


Proceedings of the 37th WIC Symposium on Information Theory in the Benelux

and

The 6th Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux

Université catholique de Louvain
Louvain-la-Neuve, Belgium
May 19–20, 2016

IEEE

UCL
Université
catholique
de Louvain

icteam

CORE 1966 - 2016 BRIDGING GAPS

Previous symposia

1. 1980 Zoetermeer, The Netherlands, Delft University of Technology
2. 1981 Zoetermeer, The Netherlands, Delft University of Technology
3. 1982 Zoetermeer, The Netherlands, Delft University of Technology
4. 1983 Haasrode, Belgium ISBN 90-334-0690-X
5. 1984 Aalten, The Netherlands ISBN 90-71048-01-2
6. 1985 Mierlo, The Netherlands ISBN 90-71048-02-0
7. 1986 Noordwijkerhout, The Netherlands ISBN 90-6275-272-1
8. 1987 Deventer, The Netherlands ISBN 90-71048-03-9
9. 1988 Mierlo, The Netherlands ISBN 90-71048-04-7
10. 1989 Houthalen, Belgium ISBN 90-71048-05-5
11. 1990 Noordwijkerhout, The Netherlands ISBN 90-71048-06-3
12. 1991 Veldhoven, The Netherlands ISBN 90-71048-07-1
13. 1992 Enschede, The Netherlands ISBN 90-71048-08-X
14. 1993 Veldhoven, The Netherlands ISBN 90-71048-09-8
15. 1994 Louvain-la-Neuve, Belgium ISBN 90-71048-10-1
16. 1995 Nieuwekerk a/d IJssel, The Netherlands ISBN 90-71048-11-X
17. 1996 Enschede, The Netherlands ISBN 90-365-0812-6
18. 1997 Veldhoven, The Netherlands ISBN 90-71048-12-8
19. 1998 Veldhoven, The Netherlands ISBN 90-71048-13-6
20. 1999 Haasrode, Belgium ISBN 90-71048-14-4
21. 2000 Wassenaar, The Netherlands ISBN 90-71048-15-2
22. 2001 Enschede, The Netherlands ISBN 90-365-1598-X
23. 2002 Louvain-la-Neuve, Belgium ISBN 90-71048-16-0
24. 2003 Veldhoven, The Netherlands ISBN 90-71048-18-7
25. 2004 Kerkrade, The Netherlands ISBN 90-71048-20-9
26. 2005 Brussels, Belgium ISBN 90-71048-21-7
27. 2006 Noordwijk, The Netherlands ISBN 90-71048-22-7
28. 2007 Enschede, The Netherlands ISBN 978-90-365-2509-1
29. 2008 Leuven, Belgium ISBN 978-90-9023135-8
30. 2009 Eindhoven, The Netherlands ISBN 978-90-386-1852-4
31. 2010 Rotterdam, The Netherlands ISBN 978-90-710-4823-4
32. 2011 Brussels, Belgium ISBN 978-90-817-2190-5
33. 2012 Enschede, The Netherlands ISBN 978-90-365-3383-6
34. 2013 Leuven, Belgium ISBN 978-90-365-0000-5
35. 2014 Eindhoven, The Netherlands ISBN 978-90-386-3646-7
35. 2015 Brussels, Belgium ISBN 978-2-8052-0277-3

Proceedings

Proceedings of the 37th Symposium on Information Theory in the Benelux and the 6th Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux. Edited by François Glineur and Jérôme Louveaux.
ISBN: (to be assigned)

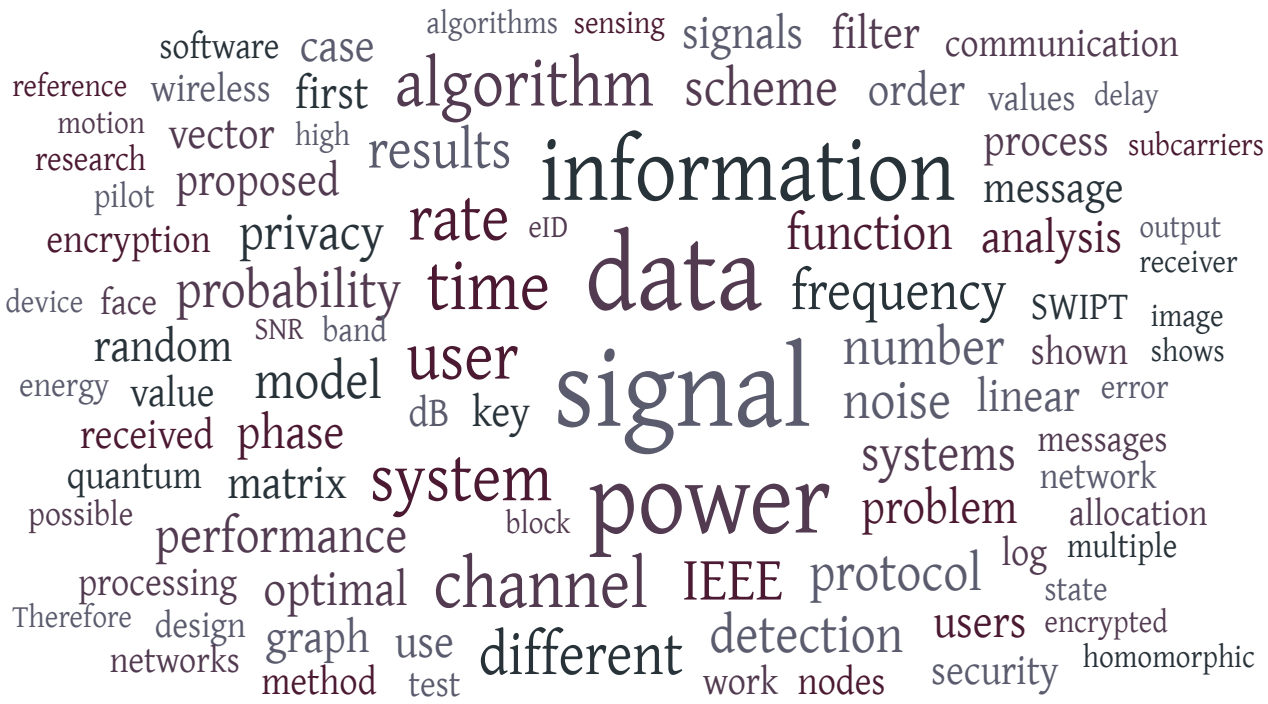
The 37th Symposium on Information Theory in the Benelux and the 6th Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux have been organized by

Université catholique de Louvain, Louvain-la-Neuve, Belgium

http://sites.uclouvain.be/sitb2016/

on behalf of the

Werkgemeenschap voor Informatie- en Communicatietheorie, the IEEE Benelux Information Theory Chapter and the IEEE Benelux Signal Processing Chapter.



Financial support from the Gauss foundation, the IEEE Benelux Information Theory Chapter, the IEEE Benelux Signal Processing Chapter and the Werkgemeenschap voor Informatie- en Communicatietheorie is gratefully acknowledged

Organizing committee
François Glineur (UCL)
Jérôme Louveaux (UCL)

Table of contents

Delay Performance Enhancement for DSL Networks through Cross-Layer Scheduling

Jeremy Van den Eynde[1] Jeroen Verdyck[2] Chris Blondia[1] Marc Moonen[2]
[1]University of Antwerp, Department of Mathematics-Computer Sciences
MOSAIC Modeling of Systems And Internet Communication
jeremy.vandeneynde@uantwerpen.be chris.blondia@uantwerpen.be
[2]KU Leuven, Department of Electrical Engineering (ESAT)
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics
jeroen.verdyck@esat.kuleuven.be marc.moonen@esat.kuleuven.be

Abstract

The quality of experience of many modern network services depends on the delay performance of the underlying communications network. In DSL networks, cross talk introduces competition for bandwidth among users. In such a competitive environment, delay performance is largely determined by the manner in which the cross-layer scheduler assigns bandwidth to the different users. Existing cross-layer schedulers optimize a simple metric, and do not consider important information that is contained within individual packets. In this paper, we present a new cross-layer scheduler, referred to as the minimal delay violation (MDV) scheduler, which optimizes a more elaborate metric that closely resembles the quality of experience of the users. Complementary to the MDV scheduler, a fast physical layer resource allocation algorithm has been developed that is based on network utility maximization. Through simulations, it is shown that the new scheduler outperforms the state of the art in cross-layer scheduling algorithms.

1 Introduction

In communications, maintaining a low delay is important for many applications such as video conferencing, VoIP, gaming, and live streaming. If many delay violations occur, quality of experience (QoE) suffers considerably for these applications. In multi-user communication systems, competition for bandwidth among users motivates the need for a scheduler that assigns bandwidth to the users. This scheduler then has a significant influence on the achieved delay performance of all applications in the network. In DSL networks, competition for bandwidth arises from physical layer resource allocation techniques that combat crosstalk, i.e. interference that results from electromagnetic coupling between different wires in a single cable binder. In the design of a scheduler for DSL systems, these physical layer mechanisms can be taken into account through the framework of cross-layer optimization.

Part of this research work was carried out at UAntwerpen, in the frame of Research Project FWO nr. G.0912.13 'Cross-layer optimization with real-time adaptive dynamic spectrum management for fourth generation broadband access networks'. Part of this research work was carried out carried out at the ESAT Laboratory of KU Leuven, in the frame of 1) KU Leuven Research Council CoE PFV/10/002 (OPTEC), 2) the Interuniversity Attractive Poles Programme initiated by the Belgian Science Policy Office: IUAP P7/23 'Belgian network on stochastic modeling analysis design and optimization of communication systems' (BESTCOM) 2012-2017, 3) Research Project FWO nr. G.0912.13 'Cross-layer optimization with real-time adaptive dynamic spectrum management for fourth generation broadband access networks', 4) IWT O&O Project nr. 140116 'Copper Next-Generation Access'. The scientific responsibility is assumed by its authors.

A cross-layer scheduler makes its scheduling decisions based on the solution to a network utility maximization (NUM) problem. Existing cross-layer schedulers optimize a simple metric, such as queue length, head-of-line delay, or average waiting time, and do not consider important information that is contained within the individual packets. In this paper, we introduce the new minimal delay violation (MDV) scheduler, which optimizes a function of the delay percentile, a measure that closely resembles the true quality of service requirements of delay sensitive traffic. Complementary to the new MDV scheduler, a fast physical layer resource allocation algorithm is developed that solves the corresponding NUM problem. The resource allocation algorithm, referred to as the NUM-DSB algorithm, is inspired by the distributed spectrum balancing (DSB) algorithm for spectrum coordination in DSL networks. The NUM-DSB algorithm decides on the appropriate power allocation for the physical layer, and can be shown to converge to a local optimum of the original NUM problem. Convergence is fast, which enables verification of the MDV scheduling algorithm through simulations.

Simulation results are obtained using the OMNeT++ framework and Matlab. The performance of the MDV scheduler is evaluated in a downstream DSL system, and is compared to the performance of both the max-weight (MW) and the max-delay utility (MDU) scheduler. Simulation results show that the MDV scheduler outperforms the MDU and MW scheduler. The MDV scheduler sometimes also demonstrates better performance with respect to throughput. Overall, when the MDV scheduler is used, it is seen that significantly fewer delay violations occur.

2 DSL system model

2.1 Physical layer

We consider an N user DSL system. DSL employs discrete multitone (DMT) modulation in order to establish K orthogonal sub channels or tones. As signal coordination is assumed not to be available, each of these tones k can be modeled as an interference channel.

$$\boldsymbol{y}_k = H_k \boldsymbol{x}_k + \boldsymbol{z}_k \tag{1}$$

In (1), $\boldsymbol{x}_k = \left[x_k^1, \ldots, x_k^N\right]^T$ is a vector containing the transmitted signal of all N users on tone k. Also, let $\boldsymbol{x}^n = \left[x_1^n, \ldots, x_K^n\right]^T$ and let $\boldsymbol{x} = \left[\boldsymbol{x}^{1^T}, \ldots, \boldsymbol{x}^{N^T}\right]^T$. Similar vector notation will be used for other signals, as well as for variables introduced later such as the bit loading, total power consumption, and data rate. Furthermore, $\boldsymbol{y}_k$ and $\boldsymbol{z}_k$ contain the received signal and noise for all N users on tone k. The average power of x_k^n is given as $s_k^n = \Delta_f \mathcal{E}\left\{|x_k^n|^2\right\}$, with $\mathcal{E}\{\cdot\}$ the expected value operator and Δ_f the tone spacing. Also, $\sigma_k^n = \Delta_f \mathcal{E}\left\{|z_k^n|^2\right\}$ is the average noise power received by user n on tone k. Finally, H_k is the $N \times N$ channel matrix, where $[H_k]_{n,m} = h_k^{n,m}$ is the transfer function between the transmitter of user m and the receiver of user n, evaluated on tone k.

The maximum achievable bit loading for user n on tone k, given transmit powers $\boldsymbol{s}_k$, is calculated as

$$b_k^n(\boldsymbol{s}_k) = \log_2\left(1 + \frac{1}{\Gamma}\frac{|h_k^{n,n}|^2 s_k^n}{\sum_{n \neq m}|h_k^{n,m}|^2 s_k^m + \sigma_k^n}\right), \tag{2}$$

with Γ the SNR gap to capacity, which incorporates the gap between ideal Gaussian signaling and the actual constellation in use. The SNR gap also accounts for the coding gain and noise margin. The data rate of user n, and the total transmit power

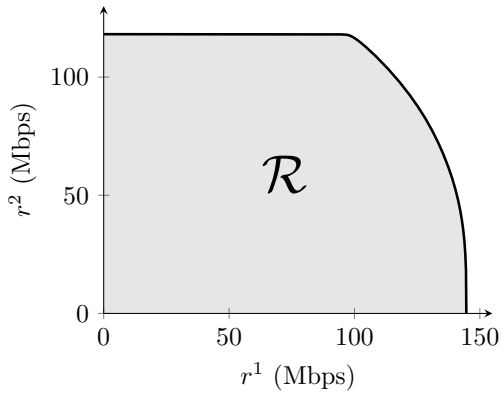


Figure 1: Rate region of a 2-user G.Fast system.

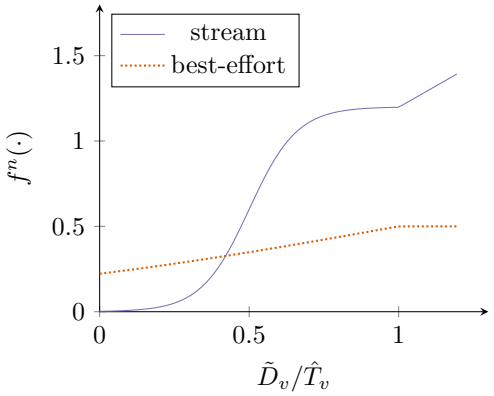


Figure 2: Weight functions for best-effort and streaming applications.

consumption of user n, are given as

$$R^n(\boldsymbol{b}^n) = f_s \sum_{k=1}^{K} b_k^n \qquad P^n(\boldsymbol{s}^n) = \sum_{k=1}^{K} s_k^n, \tag{3}$$

where f_s is the symbol rate.

The total transmit power of each user is limited to P^{tot}. The set of all possible power loadings of user n can thus be described as

$$\mathcal{S}^n = \left\{ \boldsymbol{s}^n \in \mathbb{R}_+^K \mid \; P^n(\boldsymbol{s}^n) \leq P^{\text{tot}} \right\}. \tag{4}$$

The set of all possible power loadings of the whole multi-user system is $\mathcal{S} = \mathcal{S}^1 \times \ldots \times \mathcal{S}^N$. The resulting set of achievable bit loadings is

$$\mathcal{B} = \boldsymbol{b}(\mathcal{S}) \tag{5}$$

Finally, we define the rate region as

$$\mathcal{R} = \left\{ \boldsymbol{r} \in \mathbb{R}_+^N \mid \exists\, \boldsymbol{r}' \in \boldsymbol{R}(\mathcal{B}) : \boldsymbol{r} \leq \boldsymbol{r}' \right\}. \tag{6}$$

For DSL networks with tone spacing small relative to the coherence bandwidth of the power transfer function, the rate region is a convex set [1].

As an example, the rate region of a 2-user G.Fast system that employs spectrum coordination is depicted in Figure 1. Generally, there is no power allocation that simultaneously maximizes the data rate of all users, as observed in the rate region of Figure 1. Instead, there are a number of Pareto optimal power allocation settings that achieve a data rate on the edge of the rate region. This implies the need for scheduling, i.e. choosing one of these Pareto optimal power allocation settings as the point of operation.

2.2 Upper layer & scheduling

The scheduling occurs in the upper layer, since it has the information that can help deciding the optimal point of operation. We assume that each of the N users has one traffic stream with delay upper bound $\hat{T}^n$ and allowed violation probability ϵ^n, or equivalently, conformance probability $\eta^n = 1 - \epsilon^n$. Time is divided in slots of length

τ. At slot $t \in \mathbb{N}$ the upper layer requests the physical layer for new rates, based on all available info up to time t, such as queue lengths and arrival rates. At the start of slot $t+1$, rates $\boldsymbol{r}(t+1)$ are applied in the interval $[t+1, t+2[$. There is thus a delay τ between the request and application of rates.

Traffic arrives in an infinite buffer. We denote by $a_l^n(t)$ and $Q_l^n(t)$ respectively the arrival time and length in bit of user n's l-th queued packet at the beginning of time slot t, and $Q^n(t) = \sum_{l=0}^{N^n(t)-1} Q_l^n(t)$ where $N^n(t)$ is the number of packets in user n's queue.

At the start of every slot the scheduler has to find a feasible scheduling policy that maximizes the system performance with respect to the QoS requirements. Such a policy will pick a rate $\boldsymbol{r}$ within the rate region $\mathcal{R}$. The requirements are expressed using utility functions. Such a function $u^n(r^n)$ quantifies the usefulness to user n of receiving a service r^n. Data rates $\boldsymbol{r} \in \mathcal{R}$ are then selected such that they maximize the sum of the utilities.

$$\arg\max_{\boldsymbol{r}\in\mathcal{R}} \sum_{n=1}^{N} u^n(r^n) \tag{7}$$

Ideally, $u^n(\cdot)$ is monotonically increasing, concave, and differentiable for all n.

A large family of scheduling algorithms is linear in $\boldsymbol{r}$, i.e.

$$u^n(r^n) = \omega^n r^n. \tag{8}$$

For example, the Max-Weight scheduler (MW) [2] has $\omega^n(t) = Q^n(t)$. For the Max-Delay Utility (MDU) scheduler [3], the authors give $\omega^n(t) = \frac{|u'^n(\bar{W}^n)|}{\bar{\lambda}^n}$, where u'^n is the derivative of the utility function, $\bar{W}^n$ the average waiting time, and $\bar{\lambda}^n$ the average arrival rate. It is important to note that for these linear scheduling algorithms, efficient DSL physical layer resource allocation algorithms exist [4].

The QoS requirements are expressed as a delay upper bound T^n with delay violation probability ϵ^n: $P\{D^n > \hat{T}^n\} \leq \epsilon^n$, where D^n is the packet's delay. If this delay exceeds the upper bound T^n, the packet is useless to the application. The considered performance metrics are delay violations and throughput.

3 Minimal Delay Violation Scheduler

In general, schedulers that take QoS into account aim to minimize the average delay. However, this metric offers a skewed view. Imagine twenty packets, alternating between a delay of 5ms and 55ms. This gives an average delay of 30ms per packet. If the delay requirements were 40ms, then 50% of the packets could be considered useless.

The Minimal Delay Violation (MDV) scheduler aims to minimize the delay violations, rather than the average delay. First it estimates the η^n-percentile delay $\tilde{D}^n(t)$ for the coming slots, based on the queue and observed past delays. Then, depending on the proximity of $\tilde{D}^n(t)$ to $\hat{T}^n$, a weight is defined for the user to reflect its importance. For example, if for a video v the normalized delay $\frac{\tilde{D}^v(t)}{\hat{T}^v}$ is small, then v is not important, as its delay requirements will probably not be violated, and hence it can have a lower rate assigned. If, on the other hand, $\frac{\tilde{D}^v(t)}{\hat{T}^v}$ approaches 1, then its weight should be much larger, to express it is approaching its delay upper bound.

This updated delay is then finally converted into a bit length c^n which, when divided by $r^n(t+1)$, gives an approximation to the η^n-percentile of user n's delay. It is this $\boldsymbol{c}$ that is passed on to the physical layer to find the optimal rates $\boldsymbol{r}$.

The Minimal Delay Violation (MDV) scheduler uses the utility function $u^n(r) = -\frac{c^n}{r^n}$, which is increasing, concave and differentiable on $]0, +\infty[$. At the start of every

slot, it minimizes the average of all users' η^n-percentile of the delay:

$$\arg\max_{\boldsymbol{r}\in\mathcal{R}} \sum_{n=1}^{N} -\frac{c^n(t)}{r^n} = \arg\min_{\boldsymbol{r}\in\mathcal{R}} \sum_{n=1}^{N} \frac{c^n(t)}{r^n} \tag{9}$$

We now look how c^n is constructed. Let's call $\tilde{D}^n(t) = \alpha^n \frac{\bar{q}^n(t)}{\tilde{\lambda}^n(t)} + (1-\alpha^n)\bar{d}^n(t)$, the weighted average of predicted and observed delays. Here $\alpha^n \in [0,1]$ indicates the importance of the queue. A small value means that mainly past behavior, i.e. $\bar{d}^n(t)$ which is the η^n-percentile of past delays, will influence the weight. This is useful for users that prefer a long-term average data rate, such as background jobs. A large α^n on the other hand will place more importance on the predicted delay $\frac{\bar{q}^n(t)}{\tilde{\lambda}^n(t)}$. Here $\bar{q}^n(t)$ is a measure for the queue and further explained below, and $\tilde{\lambda}^n(t) = \frac{1}{4}(\bar{\lambda}^n(t) + \sum_{s=t-2}^{t} r^n(s))$ is an estimate of the future $r^n(t+1)$, with $\bar{\lambda}^n(t)$ an average of the arrival rate. Streaming traffic benefits from this, as it can fluctuate heavily.

$\bar{q}^n(t)$ is the η^n-percentile of the user's cumulative queue size $\check{Q}^n_l(t)$, $l \in [0, N^n - 1]$:

$$\check{Q}^n_l(t) = a^n_0 r^n + \sum_{m=0}^{l'-1} Q^n_m \frac{\tilde{\lambda}^n}{r^n} + \sum_{m=l'}^{l} Q^n_m$$

The first term accounts for the head-of-line delay. The second for the packets that will be sent in the interval $[t, t+1]$, for which we already know the rates. l' is the number of packets that are transmitted in $[t, t+1[$. The final term accounts for the packets that depart in the slots $[t+1, \dots[$ at a yet unknown rate. The delay of queued packet l at a rate r^n can now simply be calculated using $\check{Q}^n_l / r^n$.

The parameter $\boldsymbol{c}$ can be expressed by $c^n = \left[\tilde{\lambda}^n \hat{T}^n f^n(\frac{\tilde{D}^n}{\hat{T}^n})\right]_1$.

The weight function $f^n(\cdot)$ transforms its argument, the proximity to $\hat{T}^n$, into a weight that reflects its importance with respect to the QoS requirements. The following functions have been defined

$f_{stream}(d) = s(\gamma = 1.2, \mu = 0.5, \sigma = 0.08, \rho = 1, x = d)$

$f_{be}(d) = s(\gamma = 1.0, \mu = 1.0, \sigma = 0.80, \rho = 0, x = d)$

with

$$s(\gamma, \mu, \sigma, \rho, x) = \begin{cases} S(x) & \text{if } x \leq 1 \\ S(1) + (x-1)\rho & \text{if } x > 1 \end{cases}$$

and the sigmoid

$$S(x) = \frac{\gamma}{1 + e^{-\frac{x-\mu}{\sigma}}}$$

They are depicted in Figure 2. These functions are tuned such that video and best-effort cooperate: if a video's delay is low then it will spare best-effort channel capacity. However if the video's delay is close to or over its delay upper bound, its weight will increase more quickly than best-effort's, which causes video's rate to increase at the cost of best-effort receiving less capacity.

4 Distributed Spectrum Balancing for Network Utility Maximization

Here, the NUM-DSB algorithm is delineated, which solves an instance of (7) for every slot t. NUM-DSB yields the optimal data rate $\boldsymbol{r}^*$, as well as the corresponding power

allocation $\boldsymbol{s}^*$. The NUM problem is non-convex on account of the bit loading being a non-convex function of the power allocation (2). Inspired by the DSB algorithm for spectrum coordination [4], our solution strategy is to construct successive per-user approximations of the rate region by defining an approximation for the bit loading that is a convex function of the power allocation. By iteratively constructing new approximations at the solution of the previous iteration, a local solution, i.e. a stationary point, of the original problem can be found.

In each iteration ℓ of the NUM-DSB algorithm, a user n will construct its own convex inner approximation of the original rate region $\mathcal{R}$. The approximation of $\mathcal{R}$ depends on the current power allocation $\boldsymbol{s}^{(\ell)}$, and is denoted as $\tilde{\mathcal{R}}(\boldsymbol{s}^{(\ell)})$. Let it be clear that, although this is not reflected in notation, the approximation $\tilde{\mathcal{R}}(\boldsymbol{s}^{(\ell)})$ is specific to user n. In order to construct $\tilde{\mathcal{R}}(\boldsymbol{s}^{(\ell)})$, it is assumed that all other users do not change their power allocation, i.e. $\boldsymbol{s}^m = \boldsymbol{s}^{m(\ell)}, \forall m \neq n$. Furthermore, the bit loading of all other users m is approximated with a lower bound hyperplane, i.e.

$$\tilde{\boldsymbol{b}}^n(\boldsymbol{s}^n; \boldsymbol{s}^{(\ell)}) = \boldsymbol{b}^n(\boldsymbol{s}) \tag{10}$$

$$\tilde{\boldsymbol{b}}^m(\boldsymbol{s}^n; \boldsymbol{s}^{(\ell)}) = \boldsymbol{b}^m(\boldsymbol{s}^{(\ell)}) + \boldsymbol{\beta}^m(\boldsymbol{s}^{(\ell)}) \circ \left(\boldsymbol{s}^n - \boldsymbol{s}^{n(\ell)}\right), \tag{11}$$

where $A \circ B$ denotes the Hadamard product of matrices A and B, and with $\beta_k^m(\boldsymbol{s}_k{}^{(\ell)})$ the directional derivative of $b_k^m(\cdot)$ at $\boldsymbol{s}_k{}^{(\ell)}$ along the n^{th} vector in the standard basis of $\mathbb{R}^n$. We want to guarantee that the value of the approximate bit loading $\tilde{b}_k^m$ remains non-negative. This can be ensured by adding a constraint on $\boldsymbol{s}^n$. Keeping in mind that $\beta_k^m(\boldsymbol{s}_k{}^{(\ell)}) < 0$, the appropriate constraint is

$$s_k^n \leq \hat{s}_k = s_k^{n(\ell)} - \max_{m \neq n} \frac{b_k^m(\boldsymbol{s}_k{}^{(\ell)})}{\beta_k^m(\boldsymbol{s}_k{}^{(\ell)})}. \tag{12}$$

The corresponding sets of all possible power loadings and resulting achievable approximate bit loadings are

$$\tilde{\mathcal{S}}^n(\boldsymbol{s}^{(\ell)}) = \{\boldsymbol{s}^n \in \mathcal{S}^n \mid \boldsymbol{s}^n \leq \hat{\boldsymbol{s}}\} \qquad \tilde{\mathcal{B}}(\boldsymbol{s}^{(\ell)}) = \tilde{\boldsymbol{b}}\big(\tilde{\mathcal{S}}^n(\boldsymbol{s}^{(\ell)}); \boldsymbol{s}^{(\ell)}\big). \tag{13}$$

Finally, the approximate rate region is defined as

$$\tilde{\mathcal{R}}(\boldsymbol{s}^{(\ell)}) = \left\{\boldsymbol{r} \in \mathbb{R}_+^N \mid \exists\, \boldsymbol{r}' \in \boldsymbol{R}\big(\tilde{\mathcal{B}}(\boldsymbol{s}^{(\ell)})\big) : \boldsymbol{r} \leq \boldsymbol{r}'\right\}. \tag{14}$$

User n thus solves the following problem, and extracts the power allocation $\boldsymbol{s}^n$ that achieves the optimal $\boldsymbol{r}$.

$$\arg\max_{\boldsymbol{r} \in \tilde{\mathcal{R}}(\tilde{\boldsymbol{s}})} \sum_{n=1}^{N} u^n(r^n) \tag{15}$$

The algorithm of choice to solve (15) is the Frank-Wolfe algorithm, which exhibits linear convergence [5] and requires no parameter tuning. This algorithm can be used as the utilities $u^n(\cdot)$ are concave and continuously differentiable by assumption, and as the rate region $\tilde{\mathcal{R}}(\boldsymbol{s}^{(\ell)})$ can be shown to be a compact convex set. The delails of the optimization algorithm are however omitted for conciseness. Then, after problem (15) has been solved, a subsequent approximation is constructed by another user at the obtained power allocation. The solutions of these successive approximations can be shown to converge to a stationary point of (7).

5 Performance

5.1 Simulation setup

The simulation consists of two parts. The NUM-DSB algorithm which is run in Matlab. The simulation of the network and upper layer scheduling is run in the OMNeT++ framework. Every $\tau = 50$ms , OMNeT++ gathers $\boldsymbol{c}$, and sends it to Matlab using the *MATLAB Engine API for C*. In the next slot, the rates $\boldsymbol{r}$ are read from Matlab, and applied to the simulated channels.

The physical layer parameters are the following. The transfer function and noise are obtained from a 99% worst case model for the physical layer of a G.Fast system with $N = 2$ users, where the respective line lengths are 450m for $n = 1$, and 390m for $n = 2$. The twisted pair cables have a line diameter of $0,5$mm, which corresponds to 24AWG. For a G.Fast system, the available per-user total transmit power is $P^{\text{tot}} = 4$dBm, the symbol rate is $f_s = 4009$Hz, the number of tones is $K = 2047$, and the tone spacing is $\Delta_f = 51.75$kHz. The SNR gap is chosen to be $\Gamma = 12.6$dB, which corresponds to BER $= 10^{-7}$, a coding gain of 3dB, and a noise margin of 6dB. The rate region that corresponds to these physical layer parameter settings is depicted in Figure 1.

The performance of the network is evaluated for 12 different traffic scenarios. Every scenario is the equivalent of one hour simulated time. Each of the N users is assigned exactly one traffic stream, the characteristics of which depend on the traffic scenario. A mix of three different kinds of traffic has been used. For video traffic, "Starwars" and "Alice in Wonderland" [6] and a 4k video entitled "The Beauty of Taiwan"* are used. Each video's packet lengths are multiplied by a constant such that the load would be closer to 1. For the second type of traffic, arrivals are determined by a Poisson process with fixed-length packets. The final traffic type kept the user's queue backlogged at all times, saturating the line. The users send packets that are encapsulated in UDP datagrams. At arrival at the next hop, the delay statistics of unfragmented packets are tracked.

5.2 Results

The simulations have been executed for the MDV scheduler, as well as for the MDU and MW scheduler. Results are displayed in Figure 3. The left plot shows the percentage of packets that violate their delay requirements. On average, MW has 7.2% of delay violations, MDU 7.4% and MDV 5.6%. Both MDV and MDU have non-zero violations in four scenarios, while MW violates delays in nine scenarios. These violations for MDV and MDU occur for scenarios in which the 4k video was playing, a very bursty video. On three out of the four scenarios, MDV outperforms MDU. The right plot of Figure 3 shows the throughput in Mbps. The results show that on average the MDV scheduler has a higher throughput (122.8 Mbps) than both the MW and MDU scheduler (121 Mbps), with differences of up to 7 Mbps (compared to MDU).

6 Conclusion

The novel cross-layer MDV scheduler has been presented, which employs a utility function to communicate its rate requirements to the physical layer. An accompanying power allocation algorithm for the physical layer (NUM-DSB) has been developed. NUM-DSB displays exceedingly fast convergence, which in turn enables the efficient

*`http://tempestvideos.skyfire.com/Sales_Optimization_Demo/beauty_taiwan_4k_final-ed.mp4`

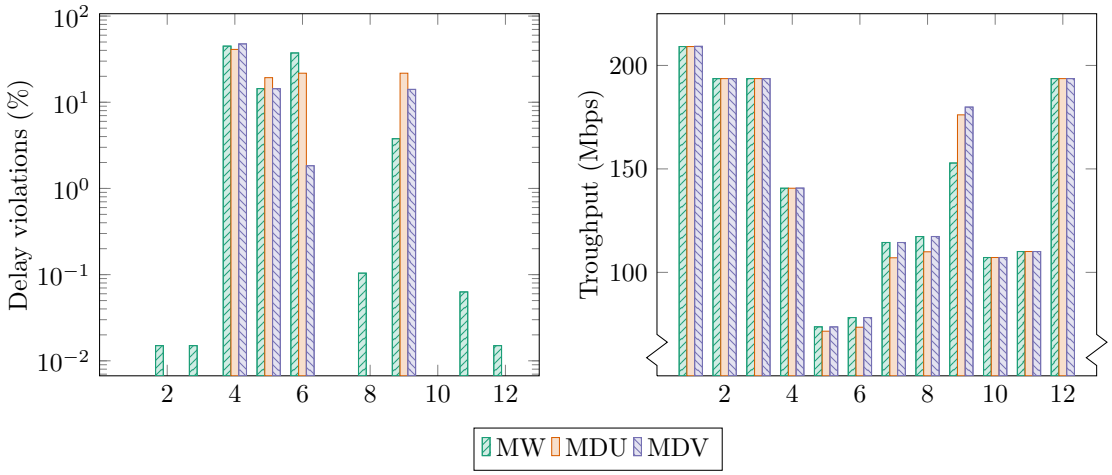


Figure 3: Delay violations (left) and throughput results (right) for the MW, MDU, and MDV schedulers. The results are displayed for 12 different traffic setups (x-axis).

execution of computer simulations that evaluate the performance of the different schedulers. These simulations have shown that, when compared to the MW and MDU scheduler, the MDV scheduler displays a significant performance improvement.

References

[1] R. Cendrillon, Wei Yu, M. Moonen, J. Verlinden, and T. Bostoen, "Optimal multiuser spectrum balancing for digital subscriber lines," *IEEE Transactions on Communications*, vol. 54, no. 5, pp. 922–933, may 2006. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1632106

[2] M. J. Neely, "Delay analysis for maximal scheduling in wireless networks with bursty traffic," in *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*. IEEE, 2008.

[3] G. Song, Y. Li, and L. J. Cimini Jr, "Joint channel-and queue-aware scheduling for multiuser diversity in wireless ofdma networks," *Communications, IEEE Transactions on*, vol. 57, no. 7, pp. 2109–2121, 2009.

[4] P. Tsiaflakis, M. Diehl, and M. Moonen, "Distributed Spectrum Management Algorithms for Multiuser DSL Networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4825–4843, oct 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4547456

[5] M. Jaggi, "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, S. Dasgupta and D. Mcallester, Eds., vol. 28. JMLR Workshop and Conference Proceedings, 2013, pp. 427–435. [Online]. Available: http://jmlr.csail.mit.edu/proceedings/papers/v28/jaggi13.pdf

[6] P. Seeling and M. Reisslein, "Video transport evaluation with H.264 video traces," *IEEE Communications Surveys and Tutorials, in print*, vol. 14, no. 4, pp. 1142–1165, 2012, Traces available at `trace.eas.asu.edu`.

Analysis of Modulation Techniques for SWIPT with Software Defined Radios

Steven Claessens Ning Pan Mohammad Rajabi Dominique Schreurs Sofie Pollin
TELEMIC Division, Department of Electrical Engineering
University of Leuven, Leuven, 3000, Belgium
steven.claessens@student.kuleuven.be
{ning.pan,mohammad.rajabi,dominique.schreurs,sofie.pollin}@kuleuven.be

Abstract

This work demonstrates the performance of software defined radios (SDRs) when transmitting multi-sine signals, which are already shown to increase RF to DC power conversion efficiency (PCE) at the receiver in a wireless power transfer system (WPT) [1]. In a system for simultaneous wireless information and power transfer (SWIPT) where the generated waveforms are modulated, however, we expect the transmitter efficiency to also be of great importance. Our goal is to evaluate the impacts on information transfer when transmitting QAM or PSK modulated power optimized waveforms (POWs) with two different SDRs. QAM and PSK modulated multi-sine waveforms are generated at fixed power levels while varying the amount of tones and modulation size and type. Error vector magnitude (EVM) is used as figure or merit for transmitter efficiency. The performed measurements show that generally, EVM decreases with increasing amount of tones. However, a high order QAM modulation of a high tone multi-sine signal at high power levels will increase EVM. Our work also shows that transmitter efficiency should not be neglected in SWIPT systems.

1 Introduction

Simultaneous wireless information and power transfer (SWIPT) is gaining interest over the last years. Increasing the radio frequency (RF) to direct current (DC) power conversion efficiency (PCE) at the receiver becomes key in SWIPT. The transmitted signal waveform can be optimized to improve PCE. The typical non-linear behavior of the diode at the rectifier results in higher PCE when excited with high PAPR signals, called power-optimized waveforms (POWs). In Figure 1, the improvement of using POWs on PCE is shown. Multi-sine waveform is a popular POW since it's PAPR can be easily controlled as explained in subsection 2.2 [2],[3].

In [4], a polynomial diode model is constructed to illustrate this phenomena. The even-order terms cause a self biasing mechanism meaning a higher time-invariant output level. It is further shown that for multi-sine POWs this DC mechanism is optimally exploited by choosing a multi-tone excitation without phase variation, resulting in a larger output DC voltage. It is later theoretically [5] and experimentally [6] shown that an upper limit on the amount of tones exists due to circuit mismatch because of increasing bandwidth and destruction of the biasing mechanism when exceeding the diode reverse breakdown voltage.

However, none of the research on SWIPT considers the impacts of different modulation schemes regarding signal distortion and received DC power. Also, WPT research generally focusses on the receiver efficiency, not taking into account the transmitter efficiency. High PAPR signals will be distorted at the transmitter because of non-linear components such as the power amplifier (PA). This paper investigates how the multi-sine signals with different modulations affect error vector magnitude (EVM) of the transmitter at various power levels. We consider different constellation sizes for PSK

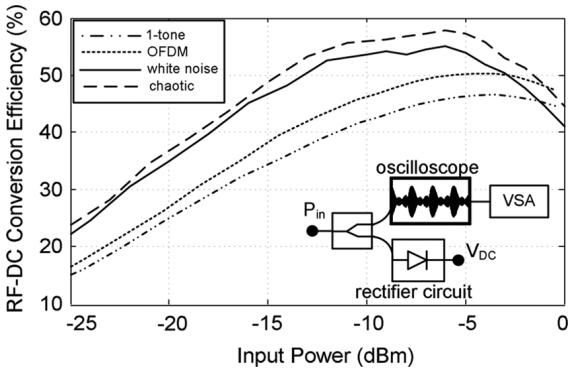


Figure 1: Rectifier RF-DC conversion efficiency versus input power for different types of POW [3].

and QAM modulations. We also demonstrate the difference in performance between a more common SDR transmitter and a high-end transmitter instrument proving the relevance of knowing transmitter efficiency in SWIPT systems.
In the next section the system set-up will be introduced. The parameters of multi-sine signals will be discussed and EVM will be explained. Next, the experimental results on EVM will be presented. Finally, conclusions will be made, we will discuss the trade-off between signal quality and optimal power transfer.

2 SWIPT System

2.1 Transmitter instruments

The multi-sine signals are generated by VST(NI PXIe-5645R) and USRP (NI-2952R). The former being a high-end instrument. We use the VST also as receiver, which is connected by cable to the transmitter instruments. The setup is presented in Figure 2 In both configurations, the EVM of the generated signal is measured by the VST to exclude common non-linear contributions due to the non-ideal receiver. In each measurement, we keep fixed average power while varying constellation size and type and the amount of tones. The carrier frequency is 2.45 GHz and the symbol rate is 50 kHz.

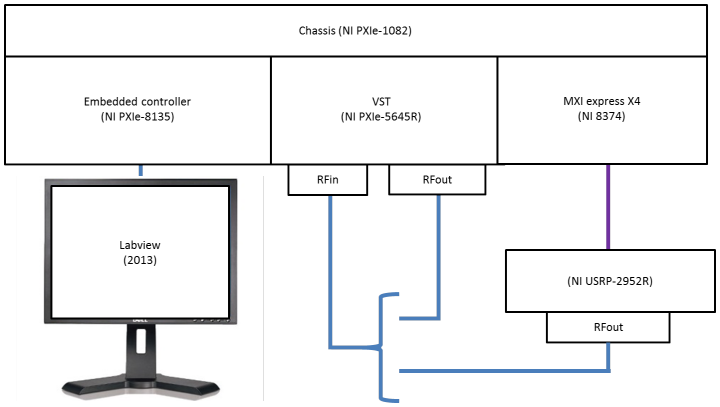


Figure 2: EVM measurements setup.

2.2 Multi-sine signals

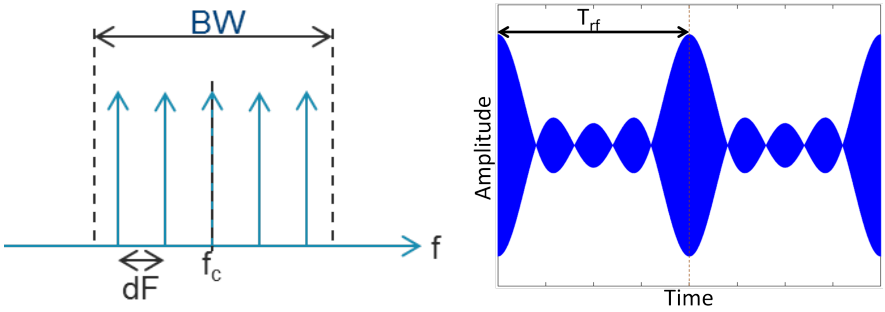
In this SWIPT system we consider multi-sine signals. The amount of tones and the frequency spacing are the most important parameters of the multi-sine waveform. As depicted in Figure 3, the considered upconverted multi-sine waveform in our work consist of N_t tones, in phase and equally spaced by Δf. The bandwidth (BW) is only dependant on the amount of tones and the frequency spacing:

$$BW_{rf} = N_t \times \Delta f. \tag{1}$$

In time domain, the period between peaks (T_{rf}) is determined by:

$$T_{rf} = \frac{1}{\Delta f}. \tag{2}$$

The PAPR increases linearly with increasing amount of tones. More energy will be located at high amplitude levels for fixed average power with increasing amount of tones. Having more energy above the forward voltage drop of a rectifier gives an intuitive explanation for the increasing PCE in WPT applications when using POWs with high PAPR. To convert the infinitely long multi-sine waveform into a symbol of limited duration we choose a segment of 1 RF period. This segment is taken symmetrically around one of the high peaks. By limiting the design to only even amount of tones, the symbol will end and start in a zero crossing. These zero crossings are useful for trailing multiple symbols, avoiding sharp transitions in the baseband symbol train. It is clear from (2) that the symbol rate is determined by the frequency spacing.



(a) Frequency (b) Time

Figure 3: Frequency and time domain representation of a 5-tone multi-sine waveform.

For fixed frequency spacing and increasing amount of tones, the minimum sampling rate also increases as the bandwidth increases. We sample each symbol 8, 16, 32 or 40 times for respectively a 2-, 4-, 8- or 10-tone multi-sine waveform, satisfying the Nyquist theorem for band-limited signals.
As shown in Figure 5, the PAPR of PSK is constant for all orders but for QAM it increases.

2.3 Error vector magnitude

EVM is our figure of merit for evaluating the quality of modulated multi-sine signals. It is defined as the RMS power value of the distance between the expected symbol and the received symbol on the constellation diagram over a collection of symbols. This is shown in Figure 4 where "O" and "X" are respectively the expected and received symbols and the decision boundaries are shown in dotted lines. EVM expressed in percentage is calculated using:

$$\begin{aligned} EVM(\%) &= \sqrt{\frac{P_{error,rms}}{P_{normalisation}}} * 100\ \%, \\ P_{error,rms} &= RMS\{|S_{received} - S_{closest\,const.point}|\}, \end{aligned} \tag{3}$$

where $P_{error,rms}$ is the average power of the difference between the received and expected symbol, S_i the complex amplitudes of the received or reference symbols and $P_{normalisation}$ is the average symbol power, which is motivated in the next section. For correct EVM measurements, the transmitted symbols should be uniformly distributed over the collection of constellation points [7]. The received symbols will be spread out

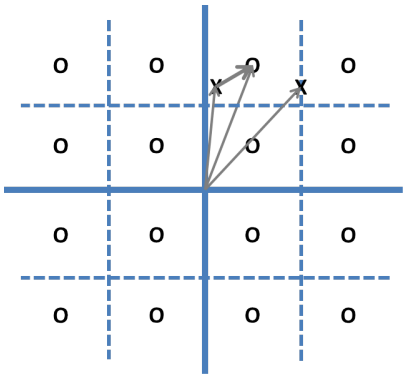


Figure 4: EVM measurement on constellation diagram.

in a cloud around the sent reference symbol point due to different types of noise such as channel noise and transmitter distortion noise. In practice, the closest constellation point to the received symbol is considered as the expected symbol. In a real system, the receiver has no prior information about the sent bitstream. Errors in EVM measurements with nondata-aided receivers occur when the received symbol crosses a decision boundary on the constellation diagram. Thus, a wrong point becomes the reference and the error distance will be underestimated. This effect has more impact for low power levels and high modulation orders [8]. Next to estimating the expected symbols it is common to normalize EVM to the symbol with maximum received power instead of the average power. $P_{normalisation}$ in (3) is then taken to be $P_{symbol,max}$. We do not use this approach as it will give unfair results when comparing waveforms with different modulations and PAPR as explained next.
To put EVM results in perspective, Figure 5 shows the theoretical EVM variation for increasing constellation size where all received symbols are located on decision boundaries closest to the reference symbols. It is clear that for increasing modulation orders, there is less noise margin for PSK symbols compared to QAM. In addition, for PSK and a given amount of tones, EVM does not depend on the normalization (i.e., to the average symbol power or maximum symbol power). Contrary to PSK, QAM symbols may have different powers and normalizing to the symbol with maxim power will give incomparable results for different modulation orders and types. When varying the amount of tones and estimating EVM by normalizing to the maxim received power

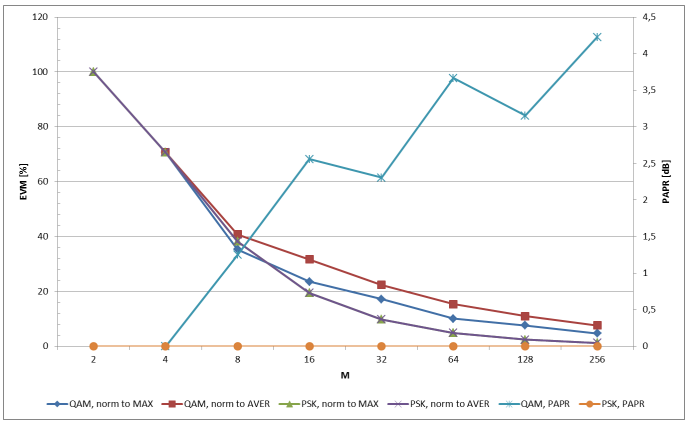


Figure 5: EVM for points at decision boundaries and PAPR for different orders of QAM and PSK.

level, distortion for waveforms with high amount of tones will also be underestimated due to increasing PAPR.

3 Experimental results

We measure EVM for different system parameters in the setup explained in subsection 2.1. A 2, 4, 8 and 10-tone multi-sine signal is modulated with different orders of QAM or PSK and transmitted by VST or USRP at various fixed power levels. For VST and USRP transmission and QAM and PSK modulation, EVM decreases with increasing amount of tones as shown in Figure 6. Only for high power levels, EVM increases for both increasing QAM modulation size and increasing amount of tones. This is because the signals high PAPR resulting in more non-linear distortion at the transmitter's PA at high power levels. For QAM, a higher modulation order will result in a higher PAPR. Contrary to QAM, all PSK symbols have the same average power thus the EVM increase at high powers is absent as shown in Figure 7.
From Figure 6 it is observed that the VST outperforms the USRP by EVM of 4% at low amount of tones and 1% at high amount of tones.
The distance between two PSK symbols is much smaller than for QAM. Hence the sensitivity for distortion is much higher comparing to PSK for high constellation size. This is proven by calculating the EVM of symbols on the decision boundary as shown in Figure 5. It is clear that for PSK, EVM has to be lower for the same error rate. This is why EVM increases for high order PSK modulation as shown in Figure 7. When looking at the EVM measurements for different orders of QAM modulation, the distortion at high power level is less for 32-QAM and 128-QAM. Because these modulation schemes are not square, PAPR is smaller as shown in Figure 5 and hence less distortion occurs than expected from the trend based on 16-QAM, 64-QAM and 256-QAM.

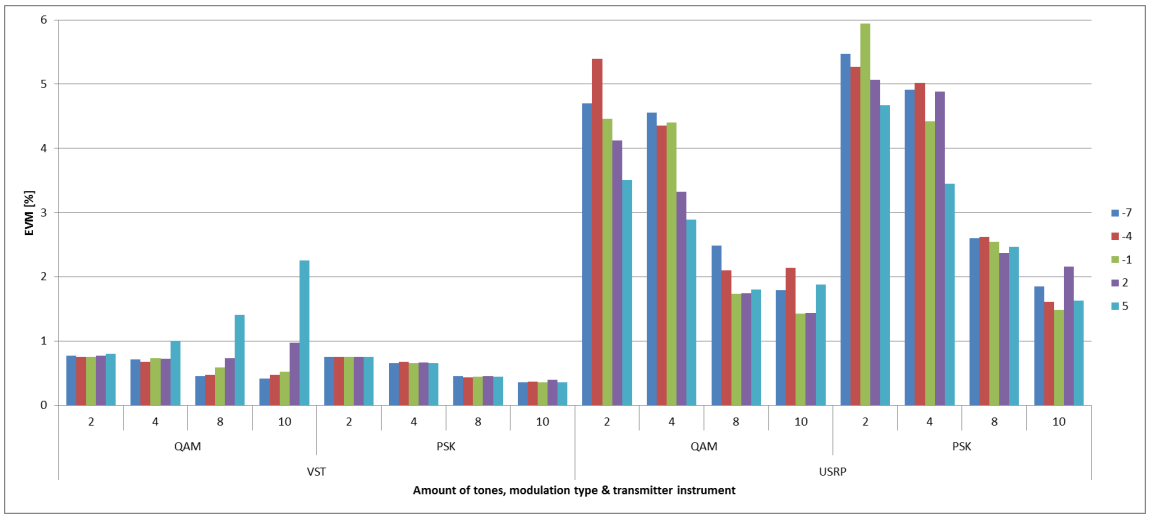


Figure 6: EVM results for 16-QAM and 16-PSK modulation of multi-sine waveform, generated by VST or USRP.

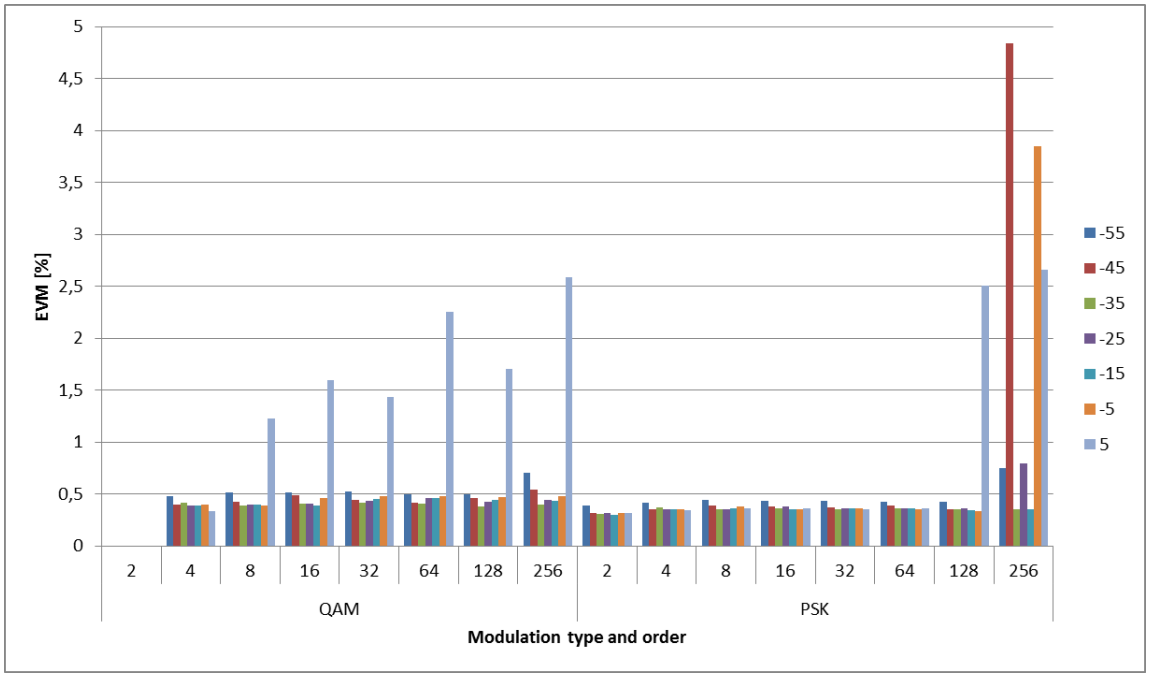


Figure 7: EVM for different orders of QAM and PSK modulation of 8-tone multi-sine waveform, transmitted with VST.

The non-linear effects at the receiver can be excluded when comparing VST and USRP transmission. Though, it would be expected that increased EVM due to the contribution of distortion at the receiver is more likely for high number of tones. Hence, the conclusions from earlier about the decrease in EVM when using higher amount of tones are still valid for each transmitter instrument. When the receiver effects on EVM are included, matched-filtering at the receiver would further increase EVM.
As mentioned, the frequency spacing in these measurements is kept constant for all configurations. This results in a much wider bandwidth for a waveform with more

tones. Figure 8 shows EVM behaviour for 16-QAM modulated 2- and 8-tone multi-sine signals, generated with VST. From this, we conclude that the EVM decrease in Figure 6 for higher amount of tones is not due to the increasing bandwidth.

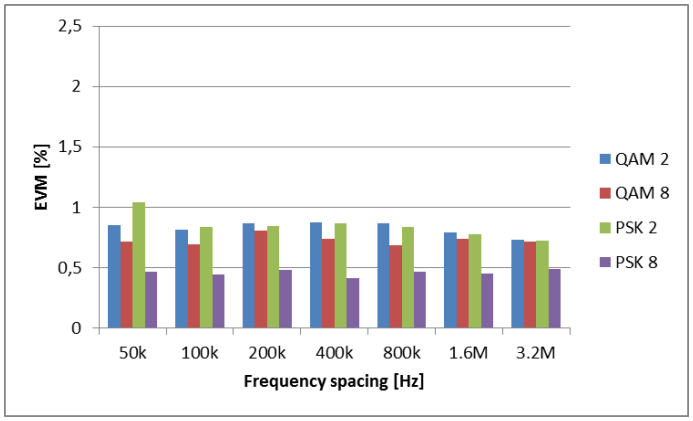


Figure 8: EVM for different frequency spacings of 16-QAM modulated 2- and 8-tone multi-sine waveform, transmitted at 0 dBm with VST.

4 Conclusion

The contribution of our work is three-fold. Firstly, we show that quality-wise, VST is better at transmitting multi-sine signals resulting in an EVM decrease of 1 % to 4 % compared to USRP transmission for respectively 10 to 2 tone signals. Secondly, we show that generally for increasing amount of tones, EVM decreases. Only for high power and QAM modulation size, EVM will increase again for increasing amount of tones due to the signals high PAPR. Thirdly, we show that at high power levels the transmitter circuits introduce more distortion when increasing the modulation order. For QAM, the signals PAPR increases with the modulation order resulting in more distortion and increasing EVM. For PSK, a higher order results in closer symbols and hence higher sensitivity to distortion, also resulting in increasing EVM.
As mentioned, for WPT the general conclusion is that multi-sine signals with high PAPR increase PCE. We have now showed that modulating these signals with QAM or PSK for information transfer has an impact on the EVM of the transmitter. High PAPR signals will improve power transfer but diminish the information link quality. For PSK all symbols have the same power level resulting in less DC output ripple. However, PSK can not be used for very high data rates due to the smaller noise margin. QAM typically has a larger noise margin but the symbols' power level vary so there will be a larger DC output ripple. The extra PAPR from QAM modulation itself will also introduce more distortion.
It is clear from our measurements that the efficiency of a common transmitter like the USRP varies significantly for different modulated multi-sine signals, proving that transmitter efficiency should be taken into account when researching SWIPT systems.

References

[1] Matthew S. Trotter, Joshua D. Griffin, and Gregory D. Durgin. Power-optimized waveforms for improving the range and reliability of RFID systems. *2009 IEEE International Conference on RFID, RFID 2009*, pages 80–87, 2009.

[2] Matthew S. Trotter and Gregory D. Durgin. Survey of range improvement of commercial RFID tags with power optimized waveforms. *RFID 2010: International IEEE Conference on RFID*, pages 195–202, 2010.

[3] a. Collado and a. Georgiadis. Optimal waveforms for efficient wireless power transmission. *IEEE Microwave and Wireless Components Letters*, 24(5):354–356, 2014.

[4] Alirio Soares Boaventura and Nuno Borges Carvalho. Maximizing DC Power in Energy Harvesting Circuits Using Multisine Excitation. *2011 IEEE MTT-S International Microwave Symposium*, 1(1):1–4, 2011.

[5] Christopher R. Valenta and Gregory D. Durgin. Rectenna performance under power-optimized waveform excitation. *2013 IEEE International Conference on RFID, RFID 2013*, pages 237–244, 2013.

[6] Ning Pan, Alirio Soares Boaventura, Mohammad Rajabi, Dominique Schreurs, Nuno Borges Carvalho, and Sofie Pollin. Amplitude and Frequency Analysis of Multi-sine Wireless Power Transfer. *Integrated Nonlinear Microwave and Millimetre-wave Circuits Workshop (INMMiC), 2015*, (1):1–3, 2015.

[7] Michael D Mckinley, Kate a Remley, Maciej Myslinski, J Stevenson Kenney, and Bart Nauwelaers. EVM Calculation for Broadband Modulated Signals. *64th ARFTG Conf Dig*, pages 45–52, 2004.

[8] Hisham A. Mahmoud and Hüseyin Arslan. Error vector magnitude to SNR conversion for nondata-aided receivers. *IEEE Transactions on Wireless Communications*, 8(5):2694–2704, 2009.

Joint Multi-objective Transmit Precoding and Receiver Time Switching Design for MISO SWIPT Systems

Nafiseh Janatian Ivan Stupia Luc Vandendorpe
Institute of Information and Communication Technologies, Electronics and Applied Mathematics, Université catholique de Louvain,
Place du Levant 2, B-1348 Louvain-la-Neuve, Belgium
nafiseh.janatian@uclouvain.be

Abstract

In this paper, we consider a time-switching (TS) co-located simultaneous wireless information and power transfer (SWIPT) system consisting of multiple multi-antenna access points which serve multiple single antenna users. In this scenario, we design jointly the optimal transmit precoding covariance matrix and the TS ratio for each receiver to maximize the utility vector made of the achieved data rates and the energy harvested of all users simultaneously. This is a non-convex multi-objective optimization problem which has been transformed into an equivalent non-convex semidefinite programming and solved using local optimization method of sequential convex programming. Numerical results illustrate the trade-off between energy harvested and information data rate objectives and show the effect of optimizing the precoding strategy and TS ratio on this trade-off.

1 Introduction

Simultaneous wireless information and power transfer (SWIPT) is a recently developed technique in which information carrying signals are also used for transferring the energy. SWIPT is a promising solution to increase the lifetime of wireless nodes and hence alleviate the energy bottleneck of energy constrained wireless networks. It is predicted that SWIPT will become an indispensable building block for many commercial and industry wireless systems in the future, including the upcoming internet of things (IoT) systems, wireless sensor networks and small cell networks [1]. The ideal SWIPT receiver architecture assumes that energy can be extracted from the same signal as that used for information decoding [2]. However, the current circuit designs are not yet able to implement this extraction, since the energy carried by the RF signal is lost during the information decoding process. As a result, a considerable effort has been devoted to the study of different practical SWIPT receiver architectures, namely, the parallel receiver architecture and the co-located receiver architecture [3]. A parallel receiver architecture equips the energy harvester and the information receiver with independent antennas for energy harvesting (EH) and information decoding (ID). In a co-located receiver architecture, the energy harvester and the information receiver share the same antennas. Two common methods to design such kind of receivers are time-switching (TS) and power-splitting (PS). In TS, the receiver switches in time between EH and ID, while in PS the receiver splits the received signal into two streams of different power for EH and ID.

SWIPT has to be realized by properly allocating the available resources and sharing them among both information transfer and energy transfer. Designing TS/PS SWIPT receivers in a point-to-point wireless environment to achieve various trade-offs between

wireless information transfer and energy harvesting is considered in [4]. In multiuser environments, researches on SWIPT focus on the power and subcarrier allocation among different users such that some criteria (throughput, harvested power, fairness, etc.) are met. For the multiuser downlink channel, various policies have been proposed for single input-single output (SISO) and multi input-single output (MISO) configurations. Resource allocation algorithm design aiming at the maximization of data transmission energy efficiency in a SISO PS SWIPT multi-user system is considered in [5] with an orthogonal frequency division multiple access (OFDMA). A MISO configuration offers the additional degree of freedom of beamforming vector optimization at the transmitter. In [6], a joint beamforming and PS ratio allocation scheme was designed to minimize the power cost under the constraints of throughput and harvested energy. The problem of joint power control and time switching in MISO SWIPT systems by considering the long-term power consumption and heterogeneous QoS requirements for different types of traffics is also studied in [7]. A MIMO interference channel with two transmitter-receiver pairs is studied in [8]. The SWIPT beamforming design for multiple cells with coordinated multipoint approach (CoMP) is also addressed in [9]. Literature overview in SWIPT shows that most of SWIPT works have considered single objective optimization (SOO) framework to formulate the problem of resource allocation or beamforming optimization. Popular objectives are classical performance metrics such as (weighted) sum rate/ throughput (to be maximized), or transmit power (to be minimized), or sum of energy harvested (to be maximized). In SOO one of these objectives is selected as the sole objective while the others are considered as constraints. This approach assumes that one of the objectives is of dominating importance and also it requires prior knowledge about the accepted values of the constraints related to the other objectives. Therefore, the fundamental approach used in this paper is the multi-objective optimization (MOO) which investigates the optimization of the vector of objectives, for nontrivial situations where there is a competition between objectives. This approach has been proposed lately for wireless information systems and is only considered for a parallel SWIPT system in [10] very recently. The system considered in [10] consists of a multi-antenna transmitter, a single-antenna information receiver, and multiple energy harvesting receivers equipped with multiple antennas. In this scenario, the trade-off between the maximization of the energy efficiency of information transmission and the maximization of the wireless power transfer efficiency is studied by means of resource allocation using an MOO framework.

In this paper, we consider a TS co-located SWIPT system consisting of multiple multi-antenna access points which serve multiple single antenna users with wireless information and power transfer. In the considered SWIPT system, we design the optimal transmit precoding covariance matrix and the time switching ratio of each receiver jointly to maximize the utility vector including the achieved information data rates and harvested energies of all users simultaneously. Since an MOO problem cannot be solved in a globally optimal way, the Pareto optimality of the resource allocation will be adopted as optimality criterion. Pareto optimality is a state of allocating the resources in which none of the objectives can be improved without degrading the other objectives [11].

The rest of this paper is organized as follows. Section II describes the system model and problem formulation. Joint multi-objective design of spatial precoding and receiver time switching is studied in section III. In Section IV, we present numerical results and finally the paper is concluded in Section V.

2 System Model and Problem Formulation

We consider a multiuser MISO downlink system for SWIPT over one single frequency band. The system consists of N_{AP} access points (APs) which are equipped with $N_{A_j}, j = 1, .., N_{AP}$ antennas and serve N_{UE} single antenna user equipments (UEs).

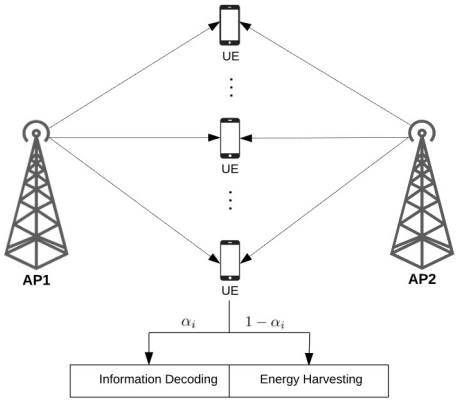


Figure 1: Time switching MISO SWIPT system

Each user is assumed to be served by multiple transmitters but the information symbols will be coded and emitted independently. Therefore, the received signal in the ith UE can be modelled as:

$$y_i = \sum_{j=1}^{N_{AP}} \boldsymbol{h}_{\boldsymbol{ij}}^{\boldsymbol{H}} \sum_{l=1}^{N_{UE}} \boldsymbol{x}_{\boldsymbol{lj}} + n_i, \tag{1}$$

where $i = 1, ..., N_{UE}, j = 1, ..., N_{AP}$, and $\boldsymbol{x}_{\boldsymbol{lj}} \in \mathbb{C}^{N_{A_j} \times 1}$ is the transmitted symbol from the jth AP to the lth UE which originates from independent Gaussian codebooks, $\boldsymbol{x}_{\boldsymbol{lj}} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{X}_{\boldsymbol{lj}})$ and $\boldsymbol{X}_{\boldsymbol{lj}} \in \mathbb{C}^{N_{A_j} \times N_{A_j}}$ denotes the transmit covariance matrix. We assume quasi-static flat fading channel for all UEs and denote by $\boldsymbol{h}_{\boldsymbol{ij}} \in \mathbb{C}^{N_{A_j} \times 1}$ the complex channel vector from the jth AP to the ith UE. Also $n_i \sim \mathcal{CN}(0, \sigma_i^2)$ is the circularly symmetric complex Gaussian receiver noise. The UEs are assumed to be capable of information decoding and also energy harvesting using time switching scheme. In particular, each reception time frame is divided into two orthogonal time slots, one for ID and the other for EH. According to (1), the achievable data rate R_i (bits/sec/Hz) and the harvested energy E_i (assuming normalized energy unit of Joule/sec), for the ith UE can be found from the following equations:

$$R_i = \log_2(1 + \frac{\sum_{j=1}^{N_{AP}} \text{trace}(\boldsymbol{H}_{\boldsymbol{ij}} \boldsymbol{X}_{\boldsymbol{ij}})}{\sigma_i^2 + \sum_{j=1}^{N_{AP}} \sum_{l=1, l \neq i}^{N_{UE}} \text{trace}(\boldsymbol{H}_{\boldsymbol{ij}} \boldsymbol{X}_{\boldsymbol{lj}})}, \tag{2}$$

$$E_i = \sum_{j=1}^{N_{AP}} \sum_{l=1}^{N_{UE}} \text{trace}(\boldsymbol{H}_{\boldsymbol{ij}} \boldsymbol{X}_{\boldsymbol{lj}}), \tag{3}$$

where $\boldsymbol{H}_{\boldsymbol{ij}} = \boldsymbol{h}_{\boldsymbol{ij}} \boldsymbol{h}_{\boldsymbol{ij}}^{\boldsymbol{H}}$. Our goal is to find the optimal transmit strategy and time switching rates to maximize the performance of all users simultaneously. Each user has its own utility vector to be optimized. Since the information data rate and harvested energy are both desirable for each user, we define the utility vector of the ith UE by $\boldsymbol{u}_i(\boldsymbol{X}_{\boldsymbol{lj}}, \alpha_i) = [\alpha_i R_i(\boldsymbol{X}_{\boldsymbol{lj}}), (1 - \alpha_i) E_i(\boldsymbol{X}_{\boldsymbol{lj}})]$ in which α_i is the fraction of timese devoted to ID in the ith receiver. As it is inferred from equations (2) and (3), these two objectives are in trade-off, because while the interference links decrease the information decoding rate, they are useful for energy harvesting. Our optimization objective is therefore to maximize the utility vector of $\boldsymbol{u}(\boldsymbol{X}_{\boldsymbol{lj}}, \alpha_i) = [\boldsymbol{u}_1(\boldsymbol{X}_{\boldsymbol{lj}}, \alpha_1), \boldsymbol{u}_2(\boldsymbol{X}_{\boldsymbol{lj}}, \alpha_2), .., \boldsymbol{u}_{N_{UE}}(\boldsymbol{X}_{\boldsymbol{lj}}, \alpha_{N_{UE}})]$ jointly via the multi-objective problem formulation. This problem can be formulated

as:

$$\begin{aligned} \underset{\boldsymbol{X_{lj}},\alpha_i}{\text{Maximize}} \quad & \boldsymbol{u}(\boldsymbol{X_{lj}},\alpha_i) \\ \text{subject to} \quad & (1)\ \sum_{l=1}^{N_{UE}} E(\boldsymbol{x_{lj}^H x_{lj}}) = \sum_{l=1}^{N_{UE}} \text{trace}(\boldsymbol{X_{lj}}) \leq P_j^{max} \\ & (2)\ \boldsymbol{X_{lj}} \succeq 0 \\ & (3)\ \alpha_i \in [0,1], \end{aligned} \tag{4}$$

where $l, i = 1, ..., N_{UE}, j = 1, ..., N_{AP}$ and the first constraint denotes the average power constraint for each AP across all transmitting antennas. In the following, we convert this problem into a SOO problem using scalarization method and then we propose an algorithm to solve this problem.

3 Joint transmit precoding and receiver time switching design

To solve the MOO problem (4), we use the weighted Chebyshev method [12] which provides complete Pareto optimal set by varying predefined preference parameters. The weighted Chebyshev goal function is:

$$f_{ch}(.) = \underset{\boldsymbol{1 \leq i \leq N_{UE}}, m=1,2}{\text{Min}} \frac{\boldsymbol{u}_i^m}{v_i^m}, \tag{5}$$

where $\boldsymbol{u}_i^m$ denotes the mth element of $\boldsymbol{u}_i(\boldsymbol{X_{lj}},\alpha_i)$ and $v_1^1, v_1^2, ..., v_{N_{UE}}^1, v_{N_{UE}}^2$ are positive weights that specify the priority of each objective. If we write the problem in epigraph form [13], this scalarization is equivalent to the following problem:

$$\begin{aligned} \underset{\boldsymbol{X_{lj}},\alpha_i,\lambda}{\text{Maximize}} \quad & \lambda \\ \text{subject to} \quad & (1)\ \alpha_i R_i(\boldsymbol{X_{lj}}) \geq \lambda v_i^1 \\ & (2)\ (1-\alpha_i) E_i(\boldsymbol{X_{lj}}) \geq \lambda v_i^2 \\ & (3)\ \sum_{l=1}^{N_{UE}} \text{trace}(\boldsymbol{X_{lj}}) \leq P_j^{max} \\ & (4)\ \boldsymbol{X_{lj}} \succeq 0 \\ & (5)\ \alpha_i \in [0,1]. \end{aligned} \tag{6}$$

This problem is a non-convex semidefinite programming (SDP) due to not only the coupled TS ratios and R_i, E_i in the first and second constraints but also the definition of R_i as presented in (2). However, using the monotonicity and concavity properties of logarithm function, and introducing the new variables $\hat{\lambda} = \log(\lambda)$, R_i, E_i, I_i and β_i, problem (6) can be represented as:

$$
\begin{aligned}
\underset{\boldsymbol{X_{lj}},\alpha_i,\beta_i,R_i,E_i,I_i,\hat{\lambda}}{\text{Maximize}} \quad & \hat{\lambda} \\
\text{subject to} \quad & \text{(C1)} \ \log(\alpha_i) + \log(R_i) \geq \hat{\lambda} + \log(v_i^1) \\
& \text{(C2)} \ \log(\beta_i) + \log(E_i) \geq \hat{\lambda} + \log(v_i^2) \\
& \text{(C3)} \ E_i = \sum_{j=1}^{N_{AP}} \sum_{l=1}^{N_{UE}} \text{trace}(\boldsymbol{H_{ij}} \boldsymbol{X_{lj}}) \\
& \text{(C4)} \ I_i = \sum_{j=1}^{N_{AP}} \sum_{l=1, l\neq i}^{N_{UE}} \text{trace}(\boldsymbol{H_{ij}} \boldsymbol{X_{lj}}) \\
& \text{(C5)} \ R_i = \log(\sigma_i^2 + E_i) - \log(\sigma_i^2 + I_i) \\
& \text{(C6)} \ \sum_{l=1}^{N_{UE}} \text{trace}(\boldsymbol{X_{lj}}) \leq P_j^{max} \\
& \text{(C7)} \ \boldsymbol{X_{lj}} \succeq 0 \\
& \text{(C8)} \ \alpha_i + \beta_i = 1 \\
& \text{(C9)} \ \alpha_i \in [0,1], \beta_i \in [0,1],
\end{aligned}
\tag{P}
$$

where (C5) is directly obtained from substituting the definition of E_i and I_i in the definition of R_i given by equation (2). This problem is a non-convex SDP because of the nonlinear equality in (C5). This problem can be solved using local optimization method of sequential convex programming (SCP). In particular, the main difficulty of problem (P) is concentrated in the nonlinear equality of (C5). This issue can be overcome by linearizing (C5) around the current iteration point and maintaining the remaining convexity of the original problem. To this end we use the first order Taylor expansion to write the linearized version of (C5) as follows:

$$
\begin{aligned}
R_i^L = & R_i(E_i^0, I_i^0) + \nabla^T R_i(E_i^0, I_i^0)[E_i - E_i^0, I_i - I_i^0] = \\
& \log(\frac{\sigma_i^2 + E_i^0}{\sigma_i^2 + I_i^0}) + \\
& \frac{1}{\sigma_i^2 + E_i^0}(E_i - E_i^0) - \frac{1}{\sigma_i^2 + I_i^0}(I_i - I_i^0),
\end{aligned}
\tag{7}
$$

where E_i^0 and I_i^0 are the points around which the equation is linearized. Now we can replace problem (P) in the kth step by the following subproblem:

$$
\begin{aligned}
\underset{\boldsymbol{X_{lj}},\alpha_i,\beta_i,R_i,E_i,I_i,\hat{\lambda}}{\text{Maximize}} \quad & \hat{\lambda} \\
\text{subject to} \quad & \text{(C1)-(C4)} \\
& \text{(C6)-(C9)} \\
& R_i \simeq \log(\frac{\sigma_i^2 + E_i^k}{\sigma_i^2 + I_i^k}) + \\
& \quad \frac{1}{\sigma_i^2 + E_i^k}(E_i - E_i^k) - \\
& \quad \frac{1}{\sigma_i^2 + I_i^k}(I_i - I_i^k).
\end{aligned}
\tag{P_k}
$$

Algorithm 1 SCP Algorithm

1: **Step 0**: Choose an initial point $\boldsymbol{w}_i^0 = [E_i^0, I_i^0]$ inside the convex set defined by (C1)-(C4), (C6)-(C9), $\gamma \in \mathbb{R}$ and a given tolerance $\epsilon > 0$. Set $k := 0$.
2: **Step 1**: For a given $\boldsymbol{w}_i^k$, solve the convex SDP of (P_k) to obtain the solution $\hat{\boldsymbol{w}}_i(\boldsymbol{w}_i^k) = [\hat{E}_i(E_i^k), \hat{I}_i(I_i^k)]$.
3: **Step 2**: If $\|\hat{\boldsymbol{w}}_i(\boldsymbol{w}_i^k) - \boldsymbol{w}_i^k\| \leqslant \epsilon$ then stop. Otherwise set $\boldsymbol{w}_i^k = \boldsymbol{w}_i^k + \gamma(\hat{\boldsymbol{w}}_i(\boldsymbol{w}_i^k) - \boldsymbol{w}_i^k)$.
4: **Step 3**: increase k by 1 and go back to step 1.

This problem is a convex SDP and it can be solved by standard optimization techniques such as Interior-point Method. In this paper, we have used CVX package to solve (P_k). The linearization point is updated with each iteration until it satisfies the termination criterion as described in Algorithm 1. It can be shown that if Algorithm 1 terminates after some iterations then $\boldsymbol{w}_i^k = [E_i^k, I_i^k], i = 1, ..., N_{UE}$ is a stationary point of problem (P). The local convergence of Algorithm 1 to a KKT point is proven in [14], under mild assumptions, and the rate of convergence is shown to be linear.

4 Numerical Results

In this section, we present numerical results to demonstrate the performance of the proposed multi-objective precoding and time switching algorithm in MISO SWIPT systems. The system is set up as follows. There are $N_{AP} = 2$ APs equipped with $N_{A_1} = N_{A_2} = 2$ antennas and $N_{UE} = 2$ single antenna user equipments. Denoting the distance of the ith user from the jth AP by d_{ij}, we assume a scenario with $d_{ij} = 10$ m $i = 1, ..., N_{UE}, j = 1, ..., N_{AP}$. At this location, channel gains are generated with Rayleigh fading and path loss effect with path loss exponent of 2. Channel bandwidth is set to 200 KHz and the spectral noise density is assumed to be $N_0 = -100$ dBm/Hz. We have set the power constraints to be $P_j^{max} = 1$ watt $j = 1, 2$. Also 100 realizations of the channel are used for averaging in simulations.

First, we investigate the trade-off between harvested energy and information data rate for the first user assuming fixed performance requirements of $E_2 = 2$ mW and $R_2 = 0.5, 2, 5$ bits/sec/Hz for the second user. Fig. (2) shows the Pareto frontier of the TS SWIPT system generated by the proposed algorithm for different directions. In other words, for each weight preference of v_1^1 and v_1^2, we find the optimal transmit covariance matrices and the optimal TS parameter, α_1, using Algorithm 1. It can be observed that the harvested energy $(1-\alpha_1)E_1$ is monotonically decreasing function with respect to the achievable data rate $\alpha_1 R_1$. This result shows that these two objectives are generally conflicting and any resource allocation algorithm that maximizes the harvested energy cannot maximize the data rate. Also as it can be seen, increasing the data rate requirement of the second user (R_2) decreases the trade-off region of the first user as a result of higher interference level. Pareto frontier of the impractical ideal SWIPT, in which EH and ID are performed simultaneously is also shown in this figure as an upper bound.

We also plot the trade-off region for two scenarios of TS1 and TS2, in which the transmit precoding matrix is optimized to maximize the harvested energy data rate and to maximize the data rate, respectively. The results in comparison with the results of the optimal MOO TS and the ideal SWIPT are depicted in Fig. (3) for $E_2 = 2$ mW and $R_2 = 0.5, 5$ bits/sec/Hz. As expected. it can be seen that the maximum harvested energy and the maximum data rate of TS1 and TS2 scenarios are equal to the two boundary points of the ideal SWIPT Pareto frontier, denoted by (R_1^1, E_1^1) and (R_1^2, E_1^2). Analytically, the Pareto frontier of TS1 and TS2 are the straight lines connecting the

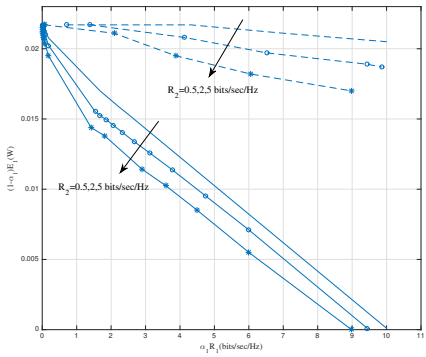


Figure 2: Pareto frontier of optimal TS and ideal SWIPT, $R_2 = 0.5, 2, 5$ bits/sec/Hz and $E_2 = 2$ mW

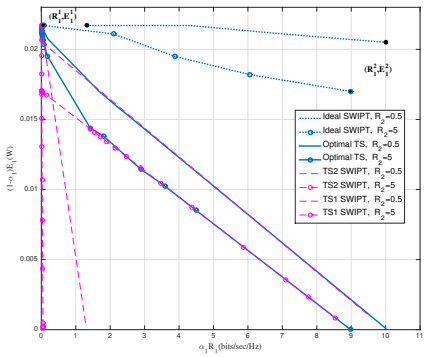


Figure 3: Comparison of optimal MOO TS SWIPT with TS1 and TS2, $R_2 = 0.5, 5$ bits/sec/Hz, $E_2 = 2$ mW

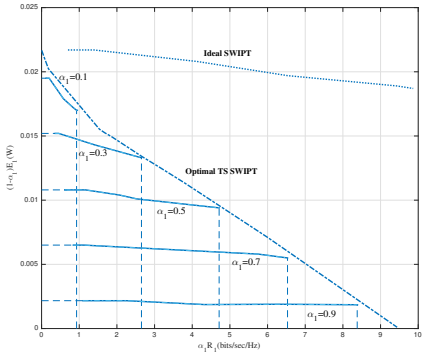


Figure 4: Effect of α_1 on the Pareto frontier, $R_2 = 2$ bits/sec/Hz, $E_2 = 2$ mW

two points $(R_1^1/R_1^2, 0)$ and $(0, E_1^1/E_1^2)$ by sweeping α_1 from 0 to 1. Simulation results in Fig. (3) verify this fact as well. As can be seen, Pareto frontier of the optimal TS reaches these two straight lines in its two boundaries. Moreover, comparing the results for $R_2 = 0.5$ and $R_2 = 5$ bits/sec/Hz, it can be inferred that in higher R_2, Pareto frontier of optimal TS approaches the Pareto frontier of TS2 in higher data rates. In other words, in a fixed data rate, more energy can be harvested using optimal TS method, since it benefits from the interference for energy harvesting.

To show the effect of TS ratio, we have also plotted the Pareto frontier of the optimal TS for $E_2 = 2$ mW, $R_2 = 5$ bits/sec/Hz and fixed values of $\alpha_1 = 0.1, 0.3, 0.5, 0.7, 0.9$ in Fig. (4). As can be seen, the lower the α_1s, the higher the E_{1max} and the lower the R_{1max}, while the optimal TS leverages the best possible E_{1max} and R_{1max} by optimizing the α_1 jointly with the precoding strategy.

5 Conclusion

In this paper, we studied the joint transmit precoding and receiver time switching design for MISO SWIPT systems. The design problem was formulated as a non-convex MOO problem with the goal of maximizing the harvested energy and information data rates for all users simultaneously. The proposed MOO problem was scalarized employing the weighted Chebyshev method. This problem is a non-convex semidefinite problem which is solved using sequence convex programming. The trade-off between energy harvested and information data rate and the effect of optimizing the precoding strategy and TS ratio on this trade-off was shown by means of numerical results.

Acknowledgment

The authors would like to thank IAP BESTCOM project funded by BELSPO, and the FNRS for the financial support.

References

[1] S. Bi, C. Ho, and R. Zhang, "Wireless powered communication: opportunities and challenges," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 117–125, 2015.

[2] L. R. Varshney, "Transporting information and energy simultaneously," in *IEEE International Symposium on Information Theory*, 2008, pp. 1612–1616.

[3] R. Zhang and C. K. Ho, "Mimo broadcasting for simultaneous wireless information and power transfer," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 1989–2001, 2013.

[4] L. Liu, R. Zhang, and K.-C. Chua, "Wireless information transfer with opportunistic energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 288–300, 2013.

[5] D. W. K. Ng, E. S. Lo, and R. Schober, "Wireless information and power transfer: Energy efficiency optimization in ofdma systems," *IEEE Transactions on Wireless Communications*, vol. 12, no. 12, pp. 6352–6370, 2013.

[6] Q. Shi, L. Liu, W. Xu, and R. Zhang, "Joint transmit beamforming and receive power splitting for miso swipt systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 6, pp. 3269–3280, 2014.

[7] Y. Dong, M. Hossain, and J. Cheng, "Joint power control and time switching for swipt systems with heterogeneous qos requirements," *IEEE Communications Letters*, vol. 20, no. 2, pp. 328 – 331, 2015.

[8] J. Park and B. Clerckx, "Joint wireless information and energy transfer in a two-user mimo interference channel," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 4210–4221, 2013.

[9] D. W. K. Ng and R. Schober, "Resource allocation for coordinated multipoint networks with wireless information and power transfer," in *IEEE Global Communications Conference*, 2014, pp. 4281–4287.

[10] S. Leng, D. W. K. Ng, N. Zlatanov, and R. Schober, "Multi-objective beamforming for energy-efficient swipt systems," *arXiv preprint arXiv:1509.05959*, 2015.

[11] E. Bjornson, E. A. Jorswieck, M. Debbah, and B. Ottersten, "Multiobjective signal processing optimization: The way to balance conflicting metrics in 5g systems," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 14–23, 2014.

[12] P. M. Pardalos, A. Migdalas, and L. Pitsoulis, *Pareto optimality, game theory and equilibria*. Springer Science & Business Media, 2008, vol. 17.

[13] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[14] Q. T. Dinh and M. Diehl, "Local convergence of sequential convex programming for nonconvex optimization," in *Recent Advances in Optimization and its Applications in Engineering*. Springer, 2010, pp. 93–102.

Target detection for DVB-T based passive radars using pilot subcarrier signal

Osama Mahfoudia[1,2], François Horlin[2] and Xavier Neyt[1]

[1]Dept. CISS,Royal Military Academy, Brussels, Belgium
[2]Dept. OPERA,Université Libre de Bruxelles, Brussels, Belgium

Abstract

Passive coherent location (PCL) radars employ non-cooperative transmitters for target detection. The cross-correlation (CC) detector, as an approximation of the optimum detector, is widely applied in PCL radars: it cross-correlates the reference signal and the surveillance signal. The CC detector is sensitive to signal-to-noise ratio (SNR) in the reference signal and thus a pre-processing of the reference signal is required. DVB-T based PCL radars can benefit from the possibility of reference signal reconstruction for SNR enhancement. The reconstruction process requires an SNR level that allows accurate signal demodulation. Hence, for low SNR values, signal reconstruction performance is limited. In this paper, we present a new approach that employs the subcarrier pilot signal for CC detection in DVB-T based PCL radars. We demonstrate the effectiveness of replacing the noisy reference signal with the a locally generated subcarrier pilot signal for CC detection.

1 Introduction

Passive coherent location (PCL) radars exploit radiations from illuminators of opportunity (IO), non-cooperative transmitter, to detect and track targets in an area of interest. The essential advantages of PCL radars are low cost, interception immunity, ease of deployment, and stealth aircraft detection capability [1, 2]. Several commercial transmitters for communication and broadcasting have been used as IO. For example, FM radio broadcast [3], satellite illumination [4], digital audio and video broadcast (DAB and DVB-T) [5] and Global System for Mobile communications (GSM) base stations [6, 7]. The architecture of PCL radars in the bistatic configuration consists of two receiving channels: a reference channel (RC) and a surveillance channel (SC). The RC captures the direct-path signal from the IO and the SC receives the target echoes.

The majority of existing PCL systems employs a cross-correlation (CC) detector. The CC detection approach is an approximation of the matched filter (MF) where a copy of the transmitted waveform is cross-correlated with the received echo to perform target detection. The exact transmitted waveform employed in MF is inaccessible in PCL radars since the IO is non-cooperative. PCL systems replace the exact waveform used in MF with the reference signal (received through RC). The reference signal is often corrupted by noise and interferences which decreases the coherent integration gain and thus degrades the CC detection performance[8].

PCL systems that exploit the DVB-T broadcasters can benefit from an enhancement of the signal-to-noise ratio (SNR) of the reference signal by demodulating and reconstructing the transmitted data. However, The reference signal reconstruction strategy is limited since it requires an SNR that allows accurate demodulation of the received signal [9, 10].

DVB-T broadcasters have attracted the interest of PCL researchers for their relatively wide bandwidth (8 MHz) allowing good range resolution. In addition, as a digital waveform its spectrum is independent of the signal content. Furthermore, the high radiated power of the DVB-T transmitters permits a considerable detection range. The DVB-T signal consists of two components: a stochastic component that results from the transmitted data randomness and a deterministic one due to the pilot subcarriers.

In this work, we consider a DVB-T based PCL radar and we introduce a new detection strategy. Precisely, we investigate the use of pilot carriers signal for detection as an alternative to the noisy reference signal. In order to do so, we adopt the statistical model developed in [8] and prove that using a locally generated pilot subcarrier signal outperforms employing the noisy reference signal for CC detection.

This paper is organized as follows. Section 2 reviews the DVB-T signal structure and introduces the average power ratio between the stochastic component and the deterministic one. Section 3 introduces the proposed detection strategy and provides the received signals model. In section 4, we derive closed-form expressions for false alarm and detection probabilities. In section 5, the simulation results validate the derived closed-form expression and show that the proposed detection strategy outperforms the use of the full noisy reference signal. Section 6 concludes the paper.

2 DVB-T signal overview

The DVB-T standard adopts the orthogonal frequency division multiplexing (OFDM) encoding method: a large number -K- of equally-spaced orthogonal subcarriers is employed to carry data ($K = 1705$ for $2K$-mode $K = 6816$ for $8K$-mode). The DVB-T signal is organized into symbols, a set of $L = 68$ symbols constructs a frame and a set of four frames composes one super-frame. The DVB-T symbols enclose three types of subcarriers: data subcarriers, transport parameter signalling (TPS) subcarriers and pilot subcarriers. There are two types of pilots: continual pilots (transmitted at known fixed frequencies) and scattered pilots (distributed following a periodic rule) [11]. Pilot subcarriers are transmitted at boosted power compared to data and TPS subcarriers. Figure 1 shows the DVB-T frame structure and emphasizes the patterns of pilot subcarriers.

The DVB-T standard employs a QAM modulation (16-QAM or 64-QAM) where the k^{th} QAM-symbol of value C_k is carried by one subcarrier of frequency f_k. The DVB-T signal s is the result of the summation over K subcarriers:

$$s(n) = \sum_{k=K_{min}}^{K_{max}} C_k e^{-j2\pi f_k n}, \tag{1}$$

where $K_{min} = 0$ and $K_{max} = 1704$ for $2K$-mode or $K_{max} = 6815$ for $8K$-mode.

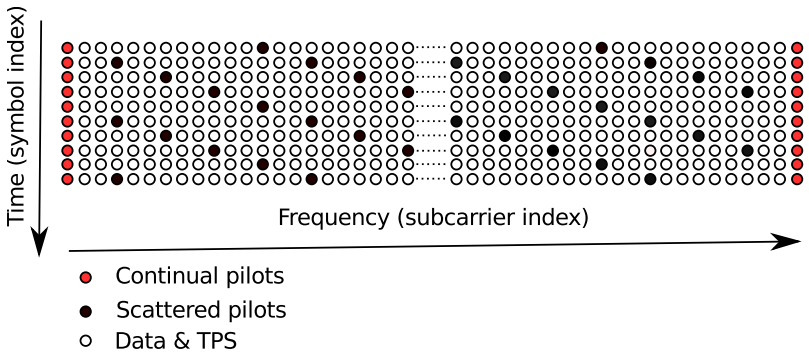


Figure 1: Pilot distribution for DVB-T signal.

The pilots are transmitted with a boosted power: an amplitude of $C_k = \pm 4/3$, thus, an average power of $E_p = 16/9$. The amplitudes of modulated data-symbols and TPS are normalized to achieve an average power of $E_d = 1$. The DVB-T signal samples $s(n)$ follow a normal distribution [12] and it can be considered as the sum of two signals $d(n)$ resulting from data subcarriers and $p(n)$ emerging from the pilot subcarriers signal:

$$s(n) = d(n) + p(n). \tag{2}$$

The data signal $d(n)$ is the sum of independent uniformly distributed QAM symbols carried by orthogonal subcarriers. Thus, the central limit theorem (CLT) leads to consider that $d(n)$ follows a normal distribution, i.e., $d(n) \sim \mathcal{CN}(0, \sigma_d^2)$. The amplitudes of pilot subcarriers ($C_k = \pm 4/3$) are generated by a Pseudo Random Binary Sequence (PRBS) generator. Hence, we consider, applying CLT, that the samples $p(n)$ follow a normal distribution, i.e., $p(n) \sim \mathcal{CN}(0, \sigma_p^2)$. Since $d(n)$ and $p(n)$ are statistically independent, we can write the variance of $s(n)$ as follows

$$\sigma_s^2 = \sigma_d^2 + \sigma_p^2. \tag{3}$$

The power ratio between the data signal and pilot signal can be calculated as follows

$$\rho = \sigma_d^2/\sigma_p^2 = (N_d E_d)/(N_p E_p), \tag{4}$$

where N_d and N_p are the number of data subcarriers and the number of pilots in one DVB-T symbol, respectively.

3 Detection strategy and signal model

In this paper, we propose a new detection approach for DVB-T based passive radars. The proposed approach takes advantage of the DVB-T signal structure that includes a deterministic part formed with pilot subcarriers. Pilot subcarrier signal can be generated knowing the broadcaster parameters such as k-mode and cyclic prefix length. Figure 2 presents the proposed approach for CC detection in noisy reference signal scenario. The received reference signal is exploited for the synchronization of the locally generated pilot subcarrier signal. The pilot signal is generated following the DVB-T standard and with blanking data and TPS subcarriers. The synchronized pilot signal replaces the noisy reference signal for CC detection.

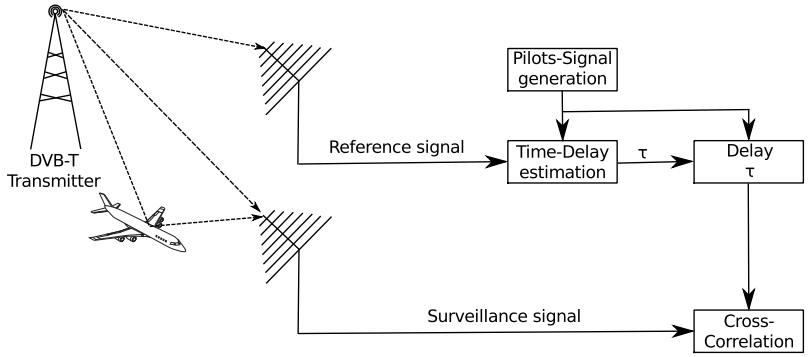


Figure 2: Passive detection employing subcarrier pilot signal.

The RC collects the reference signal formed with the direct-path signal and corrupted by noise. We note x_r the received reference signal and we consider the model in [8]:

$$x_r(n) = \beta s(n) + v(n), \tag{5}$$

where $s(n)$ is the DVB-T signal transmitted by the broadcaster, β is a scaling parameter representing the propagation losses, and $v(n)$ is i.i.d. circular complex Gaussian noise with zero mean and variance σ_v^2. We define the signal-to-noise ratio of the reference signal as

$$SNR_r = |\beta|^2 \sigma_s^2 / \sigma_v^2. \tag{6}$$

As shown in Figure 2, the surveillance signal includes direct-path signal, possible target echoes and noise contribution. We note $x_s(n)$ the received surveillance signal and we write

$$x_s(n) = \gamma s(n) + \alpha s(n-\tau) e^{j\Omega_d n} + w(n) \tag{7}$$

where γ is a scaling parameter representing the propagation losses for SC, $w(n)$ is i.i.d. circular complex Gaussian noise with zero mean and variance σ_w^2 and the target echo is characterized by τ the time-delay, Ω_d the normalized Doppler frequency and the parameter α that expresses the target reflectivity and propagation losses which is assumed to be constant during the integration time. The surveillance signal SNR is defined by

$$SNR_s = |\alpha|^2 \sigma_s^2 / \sigma_w^2. \tag{8}$$

4 Detection statistics

For the purpose of studying the performance of the proposed method, we proceeded to an analytical approach to determine the closed-form expression for the detection probability. In order to do so, the reference signal $x_r(n)$ is replaced by the time-synchronized pilot signal $p(n)$ and we formulate the binary hypothesis test as follows

$$\begin{cases} H_0 : x_s(n) = \gamma s(n) + w(n), \\ H_1 : x_s(n) = \gamma s(n) + \alpha s(n-\tau) e^{j\Omega_d n} + w(n). \end{cases} \tag{9}$$

Under both hypotheses H_0 and H_1, we calculate the statistics of the cross-correlation detector. For each range-Doppler cell, the reference signal is time-delayed and frequency shifted to match the possible target echo in the cell under test (CUT) and the detection test is given by:

$$T = \left|\bar{T}^2\right| = \left|\sum_{n=0}^{N-1} T_n\right|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \lambda, \tag{10}$$

with λ the detection threshold and T_n is the CC detector result given by

$$T_n = x_s^*(n)p(n-\tau)exp(j\Omega_d n). \tag{11}$$

Under the alternative hypothesis H_1, we calculate the mean and the variance of T_n (to retrieve the statistics for the null hypothesis, we set $\alpha = 0$). The mean value of T_n is

$$E\{T_n\} = \alpha^* \sigma_p^2, \tag{12}$$

and its variance is

$$var\left\{T_n\right\} = \left|\gamma\right|^2 (\sigma_d^2 + \sigma_p^2)\sigma_p^2 + \left|\alpha\right|^2 (\sigma_d^2 \sigma_p^2 + \phi) + \sigma_w^2 \sigma_p^2, \tag{13}$$

we introduce the pilot-data power ratio (Eq. 4)

$$var\left\{T_n\right\} = \left|\gamma\right|^2 \sigma_p^4 (1+\rho) + \left|\alpha\right|^2 (\rho\sigma_p^4 + \phi) + \sigma_w^2 \sigma_p^2, \tag{14}$$

with

$$\phi = var\{|d(n-\tau)|^2\}. \tag{15}$$

For a coherent integration time of N samples, the CC output is represented by the quantity $\bar{T}$. It follows a normal distribution with parameters (μ_0, σ_0^2) under the null hypothesis and (μ_1, σ_1^2) under the alternative one:

$$\mu_0 = 0, \tag{16}$$
$$\sigma_0^2 = N(\left|\gamma\right|^2 \sigma_p^4 (1+\rho) + \sigma_w^2 \sigma_p^2), \tag{17}$$
$$\mu_1 = N(\alpha^* \sigma_p^2), \tag{18}$$
$$\sigma_1^2 = N(\left|\gamma\right|^2 \sigma_p^4 (1+\rho) + \left|\alpha\right|^2 (\rho\sigma_p^4 + \phi) + \sigma_w^2 \sigma_p^2). \tag{19}$$

The statistic $T = |\bar{T}|^2$ under H_0 follows a central chi-squared distribution of two degrees of freedom. Thus, the false alarm probability is calculated as follows

$$P_{FA} = exp\left(-\frac{\lambda}{\sigma_0^2}\right), \tag{20}$$

the detection probability is given by the Marcum Q-function of first order:

$$P_D = Q_1\left(\sqrt{\frac{2|\mu_1|^2}{\sigma_1^2}}, \sqrt{\frac{2\sigma_0^2 ln(P_{FA}^{-1})}{\sigma_1^2}}\right). \tag{21}$$

5 Numerical results and discussion

In order to verify the validity of the detection probability expression in (21), we carried out Monte-Carlo (MC) simulations. Simulation parameters are: signal-to-noise ratio in the reference signal $SNR_r = -10dB$, coherent integration time $N = 10^5$ and false-alarm probability $P_{FA} = 10^{-2}$ (N and P_{FA} are constant for this section). Figure 3 shows the detection probability variation versus the surveillance signal SNR value for MC simulations and expression in (21). As can be seen from this figure, the analytical expression matches perfectly MC results which validates the derived closed-form expressions and the feasibility of pilot subcarrier signal detection for DVB-T based passive radars.

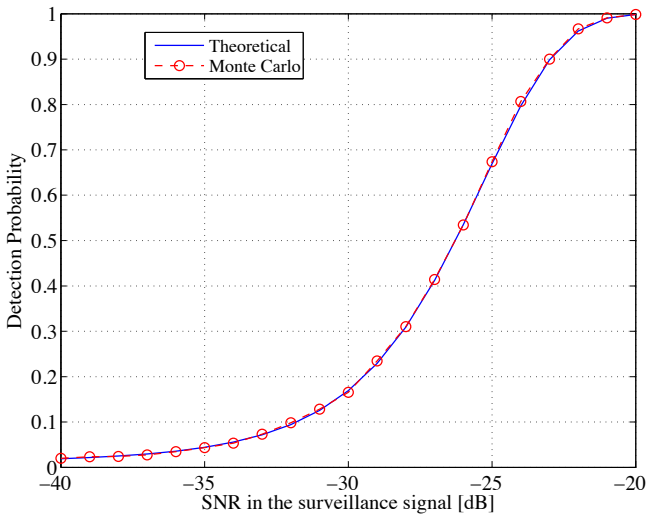


Figure 3: Validation of the detection probability expression.

After validating the feasibility of using pilot subcarrier signal as reference signal for the CC detector, the proposed method is compared to two variants of reference signal: a received reference signal with signal-to-noise ratio $SNR_r = -10dB$ and a reconstructed reference signal. Figure 4 presents the results of MC simulations for detection probability considering two variants of reference signal and the pilot subcarrier signal. For the considered scenario with $SNR_r = -10dB$, CC detection employing pilot subcarrier signal outperforms that using a noisy reference signal. Detection using reconstructed reference signal surpasses slightly that using the proposed method. This is due to the fact that the reconstructed signal contains noise-free pilot subcarriers. In addition, even if the demodulation introduces errors for $SNR_r = -10dB$, the SNR loss in the reconstructed signal is slightly higher than that due to using pilot-only signal.

To investigate the behavior of CC detection employing pilot subcarrier signal, we performed MC simulations for detection probability for a set of SNR_r values and a fixed value of $SNR_s = -30dB$. Figure 5 shows MC simulation results. As the pilot subcarrier signal is unrelated to SNR_r, the detection probability is constant for a given value of SNR_s. For the reconstructed reference signal, the detection probability converges to a non-null value for low SNR_r values. This is due to the fact that signal reconstruction for low SNR_r generates noise (caused by demodulation errors) but provides noise-free subcarrier pilots. Hence, the non-null convergence value of

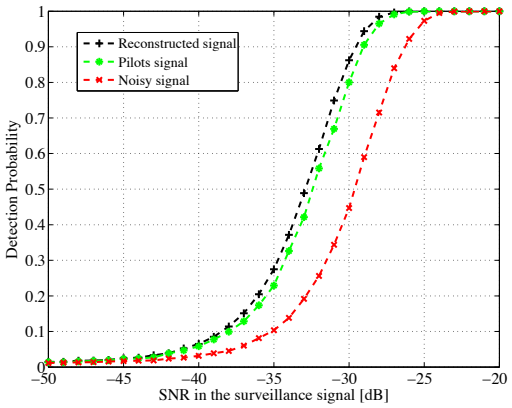


Figure 4: Detection probability for $SNR_r = -10dB$, $N = 10^5$ and $P_{FA} = 10^{-2}$.

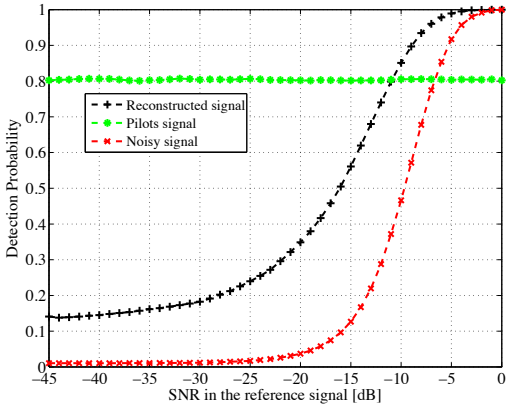


Figure 5: Detection probability for $SNR_s = -30dB$, $N = 10^5$ and $P_{FA} = 10^{-2}$.

the detection probability. For SNR_r values lower than $-11dB$, the proposed method outperforms the two other methods. Starting from $SNR_r = -11dB$, reference signal reconstruction surpasses the pilot subcarrier signal in terms of detection probability. Employing noisy reference signal outperforms the proposed method for $SNR_r = -7dB$.

The results show that the proposed method provides better performances for low SNR_r values compared to the use of noisy received reference signal. Reference signal reconstruction appears to be a solution for CC detection enhancement for noisy reference signal. However, it requires a considerable computation resources for frequency synchronization, demodulation and modulation of the received reference signal. One of the major advantages of the proposed method is that it limits the need of the received reference signal into the time-synchronization of the locally generated pilot subcarrier signal which is a large gain in computation and storage resources. Thus, we suggest employing pilot subcarrier signal for Detection in DVB-T based passive radar for low SNR_r scenario for detection probability enhancement or high SNR_r scenario for resources optimization.

6 Conclusion

In this paper, we introduced a new detection approach for DVB-T based passive radars that employs pilot subcarrier signal to replace the reference signal. We formulated closed-form expression for the detection probability and false-alarm probability and we evaluated CC detector performances analytically and using Monte-Carlo simulations for the proposed approach. We compared the proposed approach with several variants of reference signal to test its performance and limits. Based on the results it can be concluded that the proposed approach provides a solution for CC detector with noisy reference signal. In our future work, we will apply the proposed approach on real measurements.

References

[1] H. D. Griffiths and C. J. Baker, "Passive coherent location radar systems. part 1: performance prediction," *IEE Proceedings - Radar, Sonar and Navigation*, vol. 152, pp. 153–159, June 2005.

[2] C. J. Baker, H. D. Griffiths, and I. Papoutsis, "Passive coherent location radar systems. part 2: waveform properties," *IEE Proceedings - Radar, Sonar and Navigation*, vol. 152, pp. 160–168, June 2005.

[3] S. Bayat, M. Nayebi, and Y. Norouzi, "Target detection by passive coherent FM based bistatic radar," in *Radar, 2008 International Conference on*, pp. 412–415, Sept 2008.

[4] P. Marques, A. Ferreira, F. Fortes, P. Sampaio, H. Rebelo, and L. Reis, "A pedagogical passive radar using DVB-S signals," in *Synthetic Aperture Radar (APSAR), 2011 3rd International Asia-Pacific Conference on*, pp. 1–4, Sept 2011.

[5] C. Berger, B. Demissie, J. Heckenbach, P. Willett, and S. Zhou, "Signal processing for passive radar using OFDM waveforms," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, pp. 226–238, Feb 2010.

[6] X. Neyt, J. Raout, M. Kubica, V. Kubica, S. Roques, M. Acheroy, and J. Verly, "Feasibility of STAP for passive GSM-based radar," in *Radar, 2006 IEEE Conference on*, pp. 6 pp.–, April 2006.

[7] M. Kubica, V. Kubica, X. Neyt, J. Raout, S. Roques, and M. Acheroy, "Optimum target detection using illuminators of opportunity," in *Radar, 2006 IEEE Conference on*, pp. 8 pp.–, April 2006.

[8] J. Liu, H. Li, and B. Himed, "Analysis of cross-correlation detector for passive radar applications," in *Radar Conference (RadarCon), 2015 IEEE*, pp. 0772–0776, May 2015.

[9] J. Palmer, H. Harms, S. Searle, and L. Davis, "DVB-T passive radar signal processing," *Signal Processing, IEEE Transactions on*, vol. 61, pp. 2116–2126, April 2013.

[10] M. Baczyk and M. Malanowski, "Reconstruction of the reference signal in dvb-t-based passive radar," *International Journal of Electronics and Telecommunications*, vol. 57, no. 1, pp. 43–48, 2011. cited By 18.

[11] *Digital Video Broadcasting (DVB);Framing structure, channel coding and modulation for digital terrestrial television.* European Telecommunications Standards Institute, 2004.

[12] K. Polonen and V. Koivunen, "Detection of DVB-T2 control symbols in passive radar systems," in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*, pp. 309–312, June 2012.

Noise Stabilization with Simultaneous Orthogonal Matching Pursuit

Jean-François Determe*† Jérôme Louveaux† Laurent Jacques† François Horlin*

Université Libre de Bruxelles*
OPERA-WCG department
Avenue F. Roosevelt 50, CP 165/81
1050 Brussels, Belgium

jdeterme@ulb.ac.be
fhorlin@ulb.ac.be

Université Catholique de Louvain†
ICTEAM department
Avenue G. Lemaître 4, bte L4.05.01
1348 Louvain-la-Neuve, Belgium

jerome.louveaux@uclouvain.be
laurent.jacques@uclouvain.be

Abstract

This paper addresses the problem of recovering several sparse signals acquired by means of a noisy linear measurement process returning fewer observations than the dimension of the sparse signals of interest. The proposed signal model assumes that the noise is additive and Gaussian. Within the aforementioned framework, theoretical developments making use of the theory of compressive sensing show that sparse signals with similar supports can be jointly and reliably recovered by means of the greedy algorithm entitled simultaneous orthogonal matching pursuit (SOMP) provided that the linear measurements are appropriately designed. A variant of SOMP weighting each measurement vector according to its noise level is then presented. Finally, simulations confirm the benefits of weighting the measurement vectors in SOMP and show that the optimal weights predicted by the theory match the empirical ones under proper conditions.

1 Introduction to compressive sensing

In this paper, we examine the recovery of a signal $\boldsymbol{f} \in \mathbb{R}^n$ on the basis of a prescribed set of linear measurements. The matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$, which describes the linear measurement process, generates the *measurement vector* $\boldsymbol{y} \in \mathbb{R}^m$:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{f}. \tag{1}$$

Assuming that $m < n$, standard results in linear algebra show that there is no way to retrieve an arbitrary signal $\boldsymbol{f}$ on the basis of $\boldsymbol{y}$ and $\boldsymbol{\Phi}$ as $\text{rank}(\boldsymbol{\Phi}) \leq \min(m, n) = m < n$. However, many signals are known to be expressible using only a small percentage of vectors from the appropriate basis. For example, images such as those encountered in photography or medical imaging admit *sparse* representations in wavelet bases, *i.e.*, representations that only require a few wavelets to map "most" of the image. As we are about to discuss, this sparsity property can be exploited and enforced to recover $\boldsymbol{f}$ on the basis of $\boldsymbol{y}$ even though $m \ll n$. Formally, we write $\boldsymbol{f} = \boldsymbol{\Psi}\boldsymbol{x} = \sum_{j=1}^{n} x_j \boldsymbol{\psi}_j$ where $\boldsymbol{\Psi}$ represents the basis (*e.g.* a wavelet basis), $\boldsymbol{\psi}_j$ is the jth vector (or column) of $\boldsymbol{\Psi}$, and $\boldsymbol{x}$ denotes the coefficients of $\boldsymbol{f}$ in that basis. The vector $\boldsymbol{x}$ is assumed to be sparse, *i.e.*, $\|\boldsymbol{x}\|_0 := |\text{supp}(\boldsymbol{x})| := |\{j : x_j \neq 0\}|$ is low where $\text{supp}(\boldsymbol{x})$ is called the *support* of $\boldsymbol{x}$ and $|\cdot|$ denotes the cardinality. Any signal $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_0 \leq s$ is referred to as a s-sparse signal. The *compressive sensing* (CS) literature [6] has shown that if $\boldsymbol{x}$ is sufficiently sparse when compared to m and n and if $\boldsymbol{\Phi}$ satisfies some properties (see after), then it is possible to retrieve $\boldsymbol{x}$ on the basis of $\boldsymbol{\Phi}$ and the measurements $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{f} = \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{x}$. The measurement matrix $\boldsymbol{\Phi}$ is often generated randomly by using

sub-Gaussian distributions for each one of its entries [8, Chapter 9]. In this case, it can be shown that the global matrix $\boldsymbol{\Phi\Psi}$ allows the reconstruction of $\boldsymbol{x}$ with a high probability as long as $\boldsymbol{\Psi}$ represents a (deterministic) orthonormal basis [8, Theorem 9.15]. Thus, for the sake of simplicity, we will assume that $\boldsymbol{\Psi}$ is the canonical basis so that Equation (1) rewrites

$$\boldsymbol{y} = \boldsymbol{\Phi x}. \tag{2}$$

We now briefly introduce a property that, if satisfied for $\boldsymbol{\Phi}$, allows the reconstruction of any s-sparse signal $\boldsymbol{x}$ acquired by means of $\boldsymbol{\Phi}$ as in Equation (2) using algorithms of reasonable complexity. The restricted isometry property (RIP) of order s is satisfied whenever there exists a $\delta \in [0, 1[$ such that

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{\Phi x}\|_2^2 \leq (1+\delta)\|\boldsymbol{x}\|_2^2 \tag{3}$$

for all the s-sparse vectors $\boldsymbol{x}$. The smallest δ satisfying the conditions above is called the restricted isometry constant (RIC) of order s and is denoted by δ_s. For instance, it has been shown in [2] that, if $\delta_{2s} < \sqrt{2}-1$, then basis pursuit (BP) can recover any s-sparse signal $\boldsymbol{x}$ on the basis of $\boldsymbol{\Phi}$ and the measurement vector $\boldsymbol{y} = \boldsymbol{\Phi x}$. When discussing the performance of simultaneous orthogonal matching pursuit (SOMP) [10], *i.e.*, another algorithm designed for CS problems that we focus on in this paper, the results presented hereafter will also show that the probability of correct recovery improves as the RIC approaches 0. Note that choosing sub-Gaussian distributions for the entries of $\boldsymbol{\Phi}$ (as mentioned earlier) ensures that, with high probability, $\boldsymbol{\Phi}$ satisfies the RIP of order s provided that m is sufficiently large when compared to s and n [8, Theorem 9.11].

Conventions: We find useful to introduce the main notations used in this paper. For $1 \leq p < \infty$ and $\boldsymbol{x} \in \mathbb{R}^n$, we have $\|\boldsymbol{x}\|_p := (\sum_{j=1}^n |x_j|^p)^{1/p}$. Every vector is to be understood as a column vector. $\boldsymbol{x}_{\mathcal{S}}$ denotes the vector formed by the entries of $\boldsymbol{x}$ indexed within $\mathcal{S}$. Similarly, with $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$, we define $\boldsymbol{\Phi}_{\mathcal{S}}$ as the matrix formed by the columns of $\boldsymbol{\Phi}$ indexed by $\mathcal{S}$. The Moore-Penrose pseudoinverse of a matrix $\boldsymbol{\Phi}$ is written $\boldsymbol{\Phi}^+$ while its transpose is denoted by $\boldsymbol{\Phi}^{\mathrm{T}}$. The range of any matrix $\boldsymbol{\Phi}$ is denoted by $\mathcal{R}(\boldsymbol{\Phi})$. The inner product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is equal to $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}$. Finally, $\perp\!\!\!\perp$ denotes the statistical independence and $\circ$ denotes the Hadamard product.

2 Problem statement & Signal model

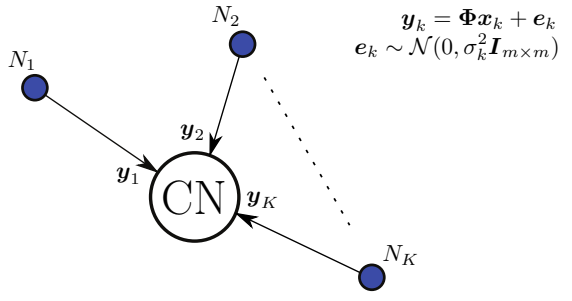


Figure 1: K nodes with different noise levels generate K measurement vectors on which the joint estimation of the support $\cup_{k=1}^K \mathrm{supp}(\boldsymbol{x}_k)$ can be performed – From [3, Figure 1]

We now turn to the particular problem addressed in this paper. While the signal model envisioned in Equation (2) involves only one sparse signals, a set of $K > 1$ sparse signals sharing a common support, *i.e.*, sharing a common structure, naturally appear in particular applications. As described in Figure 1, a set of K sensors observe a common physical phenomenon (*e.g.* a chemical reaction, an image, etc.) and return their observations to the central node (CN):

$$\boldsymbol{y}_k = \boldsymbol{\Phi x}_k + \boldsymbol{e}_k \;\; (1 \leq k \leq K) \tag{4}$$

where $\boldsymbol{e}_k \sim \mathcal{N}(0, \sigma_k^2 \boldsymbol{I}_{m\times m})$ models an additive Gaussian measurement noise

whose variance σ_k^2 is possibly different for each measurement vector $\boldsymbol{y}_k$ (or measurement channel). These discrepancies of the noise variances typically originate from the manufacturing process or from the possibly different hardware components in each sensor [3]. Typically, some local variability might be observed and this translates into different vectors $\boldsymbol{x}_k$ but the supports $\mathrm{supp}(\boldsymbol{x}_k)$ should remain nearly identical as they convey the rough structure of the phenomenon being observed. In this particular scenario, the algorithm recovering the $\boldsymbol{x}_k$ is run at the CN while the K sensors are usually cheap and their only role is limited to the acquisition of the measurements and their transfer to the CN. The curious reader will find other applications of this signal model in [1, Section 3.3]. Equation (4) can be aggregated using matrices as

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K) = \boldsymbol{\Phi}\,(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) + (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K) = \boldsymbol{\Phi}\boldsymbol{X} + \boldsymbol{E} \tag{5}$$

where $\boldsymbol{Y}, \boldsymbol{E} \in \mathbb{R}^{m \times K}$ and $\boldsymbol{X} \in \mathbb{R}^{n \times K}$. It is interesting to extend the notion of support to matrices by defining $\mathrm{supp}(\boldsymbol{X}) := \cup_{1 \leq k \leq K} \mathrm{supp}(\boldsymbol{x}_k)$. We also use the convenient definition $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)^{\mathrm{T}} \in \mathbb{R}^K$. We often refer to as multiple measurement vector (MMV) signal models whenever $K > 1$ while the single measurement vector (SMV) denomination is reserved for $K = 1$ [7] .

The main problem to be solved is thus to retrieve the support of each $\boldsymbol{x}_k$ using an algorithm of reasonable complexity. Once the estimated support $\hat{\mathcal{S}}_k$ has been obtained for each $\boldsymbol{x}_k$, it is sufficient to solve the least-squares problem obtained when assuming that only the entries of $\boldsymbol{x}_k$ indexed by $\hat{\mathcal{S}}_k$ are non-zero, *i.e.*, we compute the non-zero entries $(\hat{\boldsymbol{x}}_k)_{\hat{\mathcal{S}}_k} = \boldsymbol{\Phi}^+_{\hat{\mathcal{S}}_k} \boldsymbol{y}_k$. The next section presents a classical algorithm in the CS literature to retrieve the joint support of all the $\boldsymbol{x}_k$ and our proposed modification exploiting the differences of the noise variances σ_k^2 for each measurement channel.

3 SOMP and noise stabilization

Given that the supports of the $\boldsymbol{x}_k$ are similar, it is interesting to perform a joint support recovery [9, 3], *i.e.*, a single common support is retrieved for all the K sparse signals. A classical algorithm executing this operation is SOMP [10], which is described in Algorithm 1.

Algorithm 1: SOMP

Require: $\boldsymbol{Y} \in \mathbb{R}^{m \times K}$, $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$, $s \geq 1$

1: Initialization: $\boldsymbol{R}^{(0)} \leftarrow \boldsymbol{Y}$ and $\mathcal{S}_0 \leftarrow \emptyset$
2: $t \leftarrow 0$
3: **while** $t < s$ **do**
4: Determine the column of $\boldsymbol{\Phi}$ to be included in the support: $j_t \leftarrow \mathrm{argmax}_{1 \leq j \leq n} \sum_{k=1}^{K} |\langle \boldsymbol{\phi}_j, \boldsymbol{r}_k^{(t)} \rangle|$
5: Update the support : $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{j_t\}$
6: Projection of each measurement vector onto $\mathcal{R}(\boldsymbol{\Phi}_{\mathcal{S}_{t+1}})^{\perp}$: $\boldsymbol{R}^{(t+1)} \leftarrow (\boldsymbol{I} - \boldsymbol{\Phi}_{\mathcal{S}_{t+1}} \boldsymbol{\Phi}^+_{\mathcal{S}_{t+1}}) \boldsymbol{Y}$
7: $t \leftarrow t + 1$
8: **end while**
9: **return** $\mathcal{S}_s$ {Support at last step}

SOMP iteratively picks columns from $\boldsymbol{\Phi}$ to simultaneously approximate the K measurement vectors $\boldsymbol{y}_k$ and then compute the associated residuals $\boldsymbol{r}_k^{(t)}$ (where $\boldsymbol{r}_k^{(t)}$ is the kth column of the residual matrix $\boldsymbol{R}^{(t)}$), which are proxies of the measurement vectors that would obtained if the indexes of the columns of $\boldsymbol{\Phi}$ already picked, *i.e.*, $\mathcal{S}_t$, did not belong to $\mathcal{S} := \mathrm{supp}(\boldsymbol{X})$ in the first place. In this particular version, the sparsity level s to be enforced is supposed to be known. One column of $\boldsymbol{\Phi}$ is added to the estimated support at each iteration t (step 4). The decision rule is based upon the metric $\sum_{k=1}^{K} |\langle \boldsymbol{\phi}_j, \boldsymbol{r}_k^{(t)} \rangle|$, which accounts

for the residuals associated with all the K measurement vectors $\boldsymbol{y}_k$. The residuals at iteration t are the orthogonal projections of the original measurement vectors onto the orthogonal complement of $\mathcal{R}(\boldsymbol{\Phi}_{\mathcal{S}_{t+1}})$, *i.e.*, the subspace spanned by the columns of $\boldsymbol{\Phi}$ currently included in the estimated support $\mathcal{S}_t$ (step 6).

Algorithm 2: SOMP-NS

Require: $\boldsymbol{Y} \in \mathbb{R}^{m \times K}$, $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$, $\{q_k\}_{1 \leq k \leq K}$, $s \geq 1$

1: $\boldsymbol{R}^{(0)} \leftarrow \boldsymbol{Y}$ and $S_0 \leftarrow \emptyset$ {Initialization}
2: $t \leftarrow 0$
3: **while** $t < s$ **do**
4: Determine the column of $\boldsymbol{\Phi}$ to be included in the support:
$j_t \leftarrow \text{argmax}_{1 \leq j \leq n} \sum_{k=1}^{K} |\langle \boldsymbol{r}_k^{(t)}, \boldsymbol{\phi}_j \rangle| q_k$
5: $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{j_t\}$
6: Projection of column measurement vector onto $\mathcal{R}(\boldsymbol{\Phi}_{\mathcal{S}_{t+1}})^{\perp}$:
$\boldsymbol{R}^{(t+1)} \leftarrow (\boldsymbol{I} - \boldsymbol{\Phi}_{\mathcal{S}_{t+1}} \boldsymbol{\Phi}^{+}_{\mathcal{S}_{t+1}}) \boldsymbol{Y}$
7: $t \leftarrow t + 1$
8: **end while**
9: **return** $\mathcal{S}_s$ {Support at last step}

Interestingly, in step 4 of Algorithm 1, the K measurement channels have an identical importance when deciding which column of $\boldsymbol{\Phi}$ should be included in the support. However, the envisioned signal model suggests that some noise variances are lower than others. Thus, we propose to modify SOMP in order to weight the impact of each measurement vector on the decision metric according to their respective noise levels. We introduce simultaneous orthogonal matching pursuit with noise stabilization (SOMP-NS), which replaces the metric $\sum_{k=1}^{K} |\langle \boldsymbol{\phi}_j, \boldsymbol{r}_k^{(t)} \rangle|$ with $\sum_{k=1}^{K} |\langle \boldsymbol{\phi}_j, \boldsymbol{r}_k^{(t)} \rangle| q_k$ where the $q_k \geq 0$ $(1 \leq k \leq K)$ are the weights affected to each measurement channel. SOMP-NS is formally described in Algorithm 2. As explained in [3, Section III.B], a computationally more interesting form of SOMP-NS is obtained by first computing the multiplication $\boldsymbol{YQ}$, where the weighting matrix $\boldsymbol{Q} := \text{diag}(\boldsymbol{q}) := \text{diag}((q_1, \ldots, q_K)^{\mathrm{T}})$ intervenes, and then run SOMP on the resulting matrix.

4 On the derivation of the optimal weights

4.1 An upper bound on the probability of SOMP-NS failing

Our main objective in this paper is to present theoretical results enabling one to analytically determine the optimal weights q_k given the noise variances σ_k^2. In other words, we want to find the weights minimizing the probability of SOMP-NS failing to identify all the correct entries of the support of $\boldsymbol{X}$ for a number of iterations equal to the cardinality of the support to be recovered, *i.e.*, for the full support recovery case. For doing so, in [3] and [5], we have built a closed-form upper bound on the probability of SOMP-NS failing to make correct decisions only. The optimal weights can then be chosen so that they minimize the bound. The bound relies on several quantities of interest [5]. Among them, some only depend on the properties of $\boldsymbol{X}$, $\boldsymbol{\sigma}$, and $\boldsymbol{q}$:

- $\diamond$ The quantity $\mu_X(K) := \min_{j \in \mathcal{S}} \frac{1}{K} \sum_{k=1}^{K} |X_{j,k}| q_k$ is the lowest weighted cumulated amplitude obtained among all the rows of $\boldsymbol{X}$ indexed by the true support $\mathcal{S} = \text{supp}(\boldsymbol{X})$. Within the framework of a full support recovery, it makes sense to consider the row of $\boldsymbol{X}$ whose weighted cumulated amplitude is the lowest as this row will likely be the most difficult to identify in the presence of noise.

- $\diamond$ The quantity $\sigma(K)^2 := (1/K) \sum_{k=1}^{K} \sigma_k^2 q_k^2$ is the average noise power on the K measurement channels after applying the weights q_k.

- ◇ The quantity $\omega_\sigma(K) := (1/\sqrt{K})\|\boldsymbol{\sigma} \circ \boldsymbol{q}\|_1/\|\boldsymbol{\sigma} \circ \boldsymbol{q}\|_2 \in [1/\sqrt{K}; 1]$ quantifies to what extent $\boldsymbol{\sigma} \circ \boldsymbol{q} = (\sigma_1 q_1, \ldots, \sigma_K q_K)^{\mathrm{T}}$ is sparse. In particular, if $\boldsymbol{\sigma} \circ \boldsymbol{q}$ is 1-sparse, then $\omega_\sigma(K) = 1/\sqrt{K}$ whereas $\omega_\sigma = 1$ whenever its entries exhibit identical magnitudes. We often prefer dealing with the upper bound $\omega_\sigma := \max_{1 \leq K < \infty} \omega_\sigma(K) \in (0; 1]$.

Finally, the minimum signal-to-mean-noise ratio $\mathrm{SNR_m}$ is given by $\min_{1 \leq K < \infty} \frac{\mu_X(K)}{\boldsymbol{\sigma}(K)}$ using the definitions above. Other quantities determine how suitable to CS algorithms the measurement matrix $\boldsymbol{\Phi}$ is:

- ◇ The scalar Γ is defined as a lower bound on the ratio of SOMP-NS metric for the correct support indexes to that obtained for the incorrect ones in the noiseless case. Additional precautions should be observed when defining Γ; we refer the reader to [4, 5, 3] for further details about these matters. In the following, it is assumed that $\Gamma > 1$, which implies that, in the noiseless case, SOMP and SOMP-NS pick all the correct support indexes before choosing incorrect ones [5, 4]. In [4, Lemma 2], two valid expressions for Γ are available. For instance, $\Gamma = (1 - \delta_{|S|+1})/(\delta_{|S|+1}\sqrt{|S|})$ is correct when using SOMP(-NS) to recover a matrix $\boldsymbol{X}$ whose support cardinality is $|\mathcal{S}|$. Note that Γ increases when the RIC of order $|\mathcal{S}|+1$, *i.e.*, $\delta_{|\mathcal{S}|+1}$, decreases. This indicates, as stated in the introduction, that low values of the RIC are beneficial to the performance of SOMP(-NS).

- ◇ The scalar ρ connects the value of the noiseless SOMP-NS metric for the correct support indexes to the quantity $\mu_X(K)$. Formally, without noise, we have $\max_{j \in \mathcal{S}} \sum_{k=1}^{K} |\langle \boldsymbol{\phi}_j, \boldsymbol{r}_k^{(t)} \rangle| q_k \geq \rho \; K \mu_X(K)$. In an unpublished work, we have shown that a valid value of ρ is $\rho = (1 - \delta_{|\mathcal{S}|})(1 + \delta_{|\mathcal{S}|})/(1 + \sqrt{|\mathcal{S}| - t}\; \delta_{|\mathcal{S}|})$ for iteration t. This result indicates, once again, that SOMP-NS performance improves as the RIC $\delta_{|\mathcal{S}|}$ approaches 0.

Defining

$$\xi := \left(1 - \frac{1}{\Gamma}\right) \rho \; \mathrm{SNR_m} - \sqrt{\frac{2}{\pi}} \omega_\sigma \tag{6}$$

and assuming $\xi > 0$, we have shown [3, 5] that the probability of SOMP-NS picking at least an incorrect support index during its first $|\mathcal{S}|$ iterations is upper bounded by a quantity proportional to

$$\exp\left[-\frac{1}{8} K \xi^2\right] \tag{7}$$

where the factor of proportionality only depends on $|\mathcal{S}|$ and n. Thus, the upper bound in Equation (7) suggests that the optimal weighting strategy is that obtained whenever ξ is maximized. Since $\omega_\sigma \in (0; 1]$ and for a sufficiently high value of $(1 - 1/\Gamma)\rho$, we can neglect the variations of ξ due to ω_σ whenever the weights q_k are modified and optimize the weights on the basis of the quantity $\mathrm{SNR_m}$.

4.2 A simpler signal model

We now introduce a particular signal model that allows the computation of the weights q_k optimizing $\mathrm{SNR_m}$. We set [3, Section VII.A] $X_{j,k} = \varepsilon_{j,k} \overline{\mu_X}$ $(1 \leq k \leq K,\, j \in \mathcal{S})$ and $X_{j,k} = 0$ $(1 \leq k \leq K,\, j \notin \mathcal{S})$ where $\varepsilon_{j,k}$ denotes a Rademacher random variable, *i.e.*, a random variable returning either 1 or -1 with probability 0.5 for both values. Two different sign patterns (SP) are envisioned [3, Section VII.A]. Sign pattern 1 refers to the case where the sign pattern is identical for all the sparse vectors $\boldsymbol{x}_k$ to be recovered,

i.e., $\varepsilon_{j,k} = \varepsilon_{j,1}$ for all $1 \leq k \leq K$, $j \in \mathcal{S}$ and $\varepsilon_{j_1,1} \perp\!\!\!\perp \varepsilon_{j_2,1}$ whenever $j_1 \neq j_2$. Sign pattern 2 corresponds to the scenario for which the sign pattern is independent for each sparse signal $\boldsymbol{x}_k$ and for each measurement channel, *i.e.*, $\varepsilon_{j_1,k_1} \perp\!\!\!\perp \varepsilon_{j_2,k_2}$ whenever $j_1 \neq j_2$ and/or $k_1 \neq k_2$. Following the steps of [3, Section VII.A], we obtain the optimal weights

$$q_k = 1/\sigma_k^2 \tag{8}$$

for this particular signal model. Furthermore, we assume that $K = 2$. We also express $\boldsymbol{q}$ and $\boldsymbol{\sigma}$ using polar coordinates: $\boldsymbol{q} := (\cos\theta_q, \sin\theta_q)^{\mathrm{T}}$ and $\boldsymbol{\sigma} = (\cos\theta_\sigma, \sin\theta_\sigma)^{\mathrm{T}}$.

5 Numerical results

In this section, using the signal model described in Section 4.2, we want to

- ◇ Show that SOMP-NS outperforms SOMP whenever the noise variances are unequally distributed, *i.e.*, whenever θ_σ is not close to 45°, provided that the truly optimal weights are used. In this first part, the theoretical formula for the optimal weights $q_k = 1/\sigma_k^2$ is never used as the optimal weights are determined using the simulation results.
- ◇ Show that the optimal weights predicted by the equation $q_k = 1/\sigma_k^2$ coincide with those obtained by means of simulations. As explained later on, for the second sign pattern, the theoretical and empirical weights are identical only whenever the cardinality of the support to be recovered is sufficiently low. Such a restriction will not be observed for the first sign pattern.

To reach the two objectives above, we simply report the results already available in [3, Section VII]. For the sake of brevity, most of the technical details will be skipped. Table 1 (originally in [3, Table I]) summarizes the configurations that have been considered for the simulations. As explained in [3, Section VII], the probability of correct full support recovery is evaluated onto a discrete grid of 2-tuples $(\theta_q, \theta_\sigma)$. The empiral probability is evaluated for each 2-tuple $(\theta_q, \theta_\sigma)$ using a number of Monte-Carlo cases that is prescribed in Table 1 using the denomination "# cases".

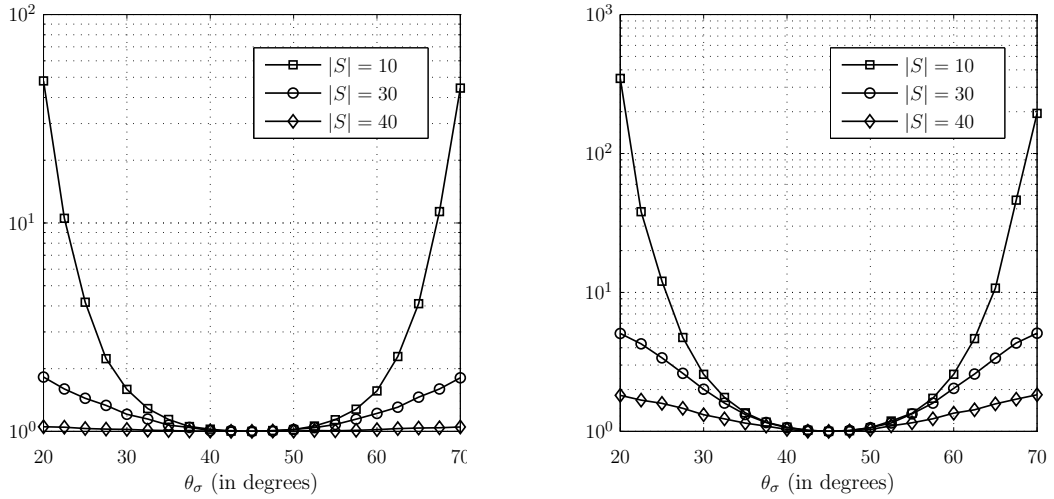
Table 1: Simulation configurations | From [3, Table I]

Configuration ID	**Sign pattern**	$\lvert\mathcal{S}\rvert$	$\overline{\mu_X}$	**# cases**
Configuration 1	1	10	2.28	$1.0\ 10^4$
Configuration 2	1	30	3.19	$1.0\ 10^4$
Configuration 3	1	40	6.94	$2.5\ 10^4$
Configuration 4	2	10	2.50	$1.0\ 10^4$
Configuration 5	2	30	3.06	$1.0\ 10^4$
Configuration 6	2	40	3.42	$1.0\ 10^4$

Figure 2 corroborates the first claim of this section, *i.e.*, SOMP-NS outperforms SOMP whenever θ_σ is different than 45°, provided that the truly optimal value of θ_q is chosen.

We now turn to our second objective, *i.e.*, providing evidence showing that the weights $q_k = 1/\sigma_k^2$ are always truly optimal for the first sign pattern while they are also correct for the sign pattern 2 provided that the cardinality $|\mathcal{S}|$ is sufficiently low.

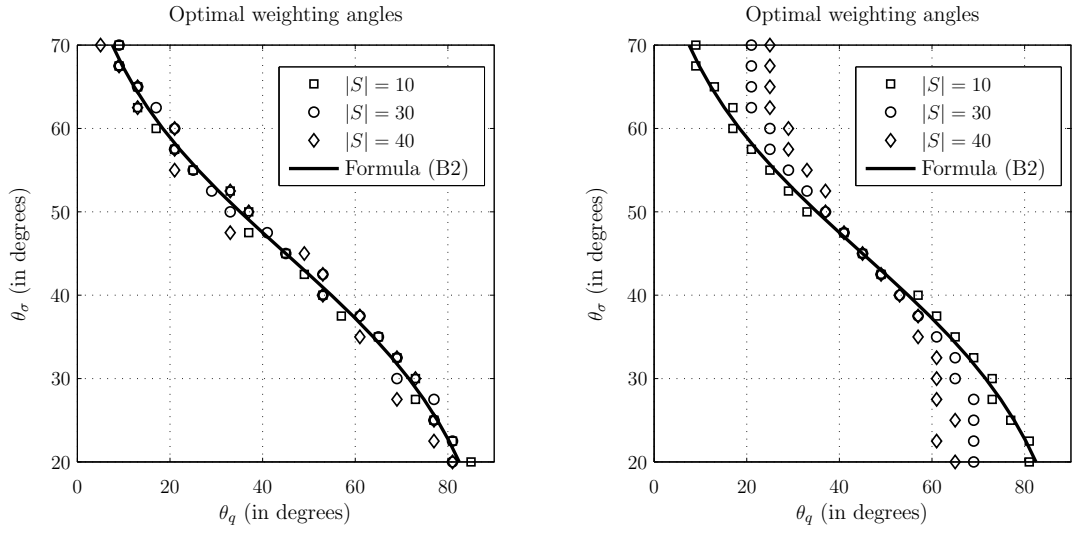
The results supporting this claim are available in Figure 3. The reasons explaining the discrepancy observed between the theoretical and empirical optimal weights in Figure 3b for the highest support cardinalities are discussed in [3, Section VII.C].



(a) Results for sign pattern 1

(b) Results for sign pattern 2

Figure 2: Simulation results ($K = 2$) – Ratio of the probability of failure of SOMP to that obtained for SOMP-NS when using the optimal weights – From [3]



(a) Results for sign pattern 1

(b) Results for sign pattern 2

Figure 3: Simulation results ($K = 2$) – Optimal weighting angles – The black curve (Formula (B2)) denotes the optimal weights predicted by $q_k = 1/\sigma_k^2$ – The markers correspond to the optimal weights derived from the simulation results – From [3]

6 Conclusion

This paper presents a novel SOMP algorithm, entitled simultaneous orthogonal matching pursuit with noise stabilization (SOMP-NS). SOMP-NS weights the impact of each measurement vector according to its noise level, which is assumed to be known. Moreover, theoretical developments suggesting that the weighting capabilities of SOMP-NS provide performance enhancements have been briefly exposed. Further, these developments have been shown to provide weighting strategies for particular signal models. Finally, numerical results have established that the weights introduced in SOMP-NS yield performance improvements provided that they are appropriately chosen. The simulations have also revealed that the weighting strategies stemming from the theoretical results are optimal in specific cases.

Disclaimer & Acknowledgements: This paper presents results [3, 4, 5] that have been published in or submitted to IEEE journals. This research has also been partially funded by the Belgian science fondation (FNRS).

References

[1] D. Baron, M. F. Duarte, M. B. Wakin, S. Sarvotham, and R. G. Baraniuk, "Distributed compressive sensing," *arXiv preprint arXiv:0901.3403*, 2009.

[2] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9, pp. 589–592, 2008.

[3] J.-F. Determe, J. Louveaux, L. Jacques, and F. Horlin, "Simultaneous orthogonal matching pursuit with noise stabilization: theoretical analysis," *arXiv preprint arXiv:1506.05324*, 2015.

[4] ——, "On the exact recovery condition of simultaneous orthogonal matching pursuit," *Signal Processing Letters, IEEE*, vol. 23, no. 1, pp. 164–168, 2016.

[5] ——, "On the noise robustness of simultaneous orthogonal matching pursuit," *arXiv preprint arXiv:1602.03400*, 2016.

[6] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[7] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *Information Theory, IEEE Transactions on*, vol. 55, no. 11, pp. 5302–5316, 2009.

[8] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Springer, 2013.

[9] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms," *Journal of Fourier analysis and Applications*, vol. 14, no. 5-6, pp. 655–687, 2008.

[10] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, 2006.

Camera Motion for Deblurring

Bart Kofoed Peter H.N. de With
Eindhoven University of Technology
Video Coding and Architectures
Eindhoven, The Netherlands
b.kofoed@tue.nl p.h.n.de.with@tue.nl

Eric Janssen
Prodrive Technologies B.V.
Son, The Netherlands
eric.janssen@prodrive-technologies.com

Abstract

Motion blur destroys high spatial frequency content, leaving deblurring as an ill-posed image restoration problem. The common practice is to prevent motion blur by reducing the exposure time, but this gives poor image quality in low-light conditions. We show that camera motion can substantially improve the image quality if 1) there are only a few distinct velocities in a scene and 2) the velocities are known prior to the exposure. We derive the camera motion trajectories that optimize the image quality of the deconvolved images. In particular, we study the case with only two distinct velocities, which occurs frequently in practice. We have built a camera that moves its field-of-view using a modified optical image stabilizer. The post-processing involves a spatially-invariant velocity-dependent deconvolution. Practical experiments confirm that our approach is more effective than existing techniques with an SNR increase of up to 4.4 dB compared to the common practice.

1 Introduction

Motion deblurring is a common and hard to solve problem in photography. Apart from issues with estimating the Point Spread Function, the inversion itself is mostly ill-posed. Deblurring algorithms are sensitive to noise and not always able to recover all details in an image. In practice it has been found that it is more effective to use an exposure time short enough to freeze all motion, although the Signal-to-Noise Ratio (SNR) becomes poor.

In this paper, we show that motion blur can be made invertible if we move the camera field-of-view during the exposure. The optimal camera motion trajectory ensures that information is captured in the best possible way, while simplifying the necessary deconvolution.

2 Prior Work

Motion blur can be made invertible by moving the camera (Motion-Invariant Photography) [1], or rapidly opening and closing the shutter during the exposure (Coded Exposure) [2]. However, Cossairt *et al.* [5] show that the image quality improvement of these techniques is negligible, due to a combination of signal-dependent noise and noise amplification during deconvolution. Although the performance gain of Motion Invariant Photography (MIP) is small, McCloskey *et al.* show that the performance gain can be improved by using more prior information about the velocity distribution [3] and implements it using Optical Image Stabilization (OIS) hardware [4],[6].

This work improves on prior work by incorporating more detailed prior information about the one-dimensional velocity distribution in a scene and verifying this with practical experiments.

3 Noise, Motion and Blur

3.1 Noise

Camera images suffer from numerous sources of degradation, among which noise is one of the most problematic in low-light conditions. Noise is either caused by electronic noise on the sensor or by the random nature of photons interacting with the detector. The latter *photon noise* is signal-dependent noise that arises from the Poisson distribution of photons reaching the detector. The variance of photon noise is approximately proportional to the number of interacting photons. In anticipation of future image sensor developments, we consider *photon-noise limited* detectors with a Quantum Efficiency of 100% during the remainder of this paper.

3.2 Motion Blur

To analyze motion blur, we adopt the space-time analysis described by Levin *et al.* [1]. This section provides a brief overview of the method. We consider one-dimensional motion blur and use one-dimensional images to facilitate the analysis.

Let $s(x)$ denote the projection of the scene, normalized such that $\mathbb{E}\left(s(x)\right) = 1$. The light level L represents the average *photon flux* per pixel, i.e. the number of photons detected per pixel per second. Hence, the photon flux at location x is $L \cdot s(x)$.

At relatively fast shutter speeds, the subject motion is approximated by a translation with constant velocity v along the x-axis. The camera moves along the x-axis as well and its position is denoted as $x_c(t)$. The image sensor integrates incoming light intensity on its surface during the exposure time T. Hence, the image captured by the sensor is written as the integral of the convolution of $s(x)$ with the *light integration curve* $h_v(t, x)$:

$$z(x) = \int_0^T h_v(t,x) * s(x) dt + n = s(x) * \int_0^T h_v(t,x) dt + n, \tag{1}$$

with n the photon noise with variance σ_n^2 and

$$h_v(t,x) = L \cdot \delta(x + vt - x_c(t)). \tag{2}$$

We analyze the properties of the convolution with $h_v(t, x)$ in the frequency domain, where convolution is equivalent to multiplication. First, set $v = 0$ and transform the light integration curve $h_0(t, x)$ to the frequency domain $H_0(\omega_t, k_x)$. This Fourier transform is rewritten as the one-dimensional Fourier transform.

$$\begin{aligned} H_0(\omega_t, k_x) &= \int_0^T \int_{-\infty}^{\infty} h_0(t,x) e^{-j2\pi(\omega_t t + k_x x)} dx dt \\ &= \int_0^T \int_{-\infty}^{\infty} L \cdot \delta(x - x_c(t)) e^{-j2\pi(\omega_t t + k_x x)} dx dt = L \int_0^T e^{-j2\pi k_x x_c(t)} e^{-j2\pi\omega_t t} dt \end{aligned} \tag{3}$$

Next, we use the fact that for a subject velocity v, time integration is equivalent to projection of $h_0(t, x)$ on the line $x = vt$ and hence $H_v(k_x) = H_0(vk_x, k_x)$, i.e. the information lies on the line $\omega_t = vk_x$. Consequently, the frequency-domain representation of the image captured by the sensor is given by:

$$Z(k_x) = H_0(vk_x, k_x) S(k_x) + N(k_x) = H_v(k_x) S(k_x) + N(k_x), \tag{4}$$

where $S(k_x)$ and $N(k_x)$ denote the frequency-domain representations of the original image and the noise, respectively.

3.3 Noise Equivalent Quanta

We aim to minimize the noise variance in the reconstructed image. A useful tool for analyzing the noise variance is the Noise Equivalent Quanta (NEQ) [7]. The NEQ measures the SNR as a function of spatial frequency k_x. For a relatively low-contrast image, photon noise is spatially white and has a variance equal to the average number of photons detected per pixel:

$$\mathbb{E}\left(|N(k_x)|^2\right) = \sigma_n^2 = H_v(0) = H(0,0) = LT. \tag{5}$$

We obtain the following expressions for the NEQ:

$$NEQ(\omega_t, k_x) = \frac{|H(\omega_t, k_x)|^2 |S(k_x)|^2}{H(0,0)}, NEQ_v(k_x) = \frac{|H_v(k_x)|^2 |S(k_x)|^2}{H_v(0)}. \tag{6}$$

An useful property is that linear operations do not change the NEQ, since the numerator and denominator are affected equally. Consequently, the NEQ is an indicator of image quality regardless of any subsequent linear filtering operations, such as deconvolution.

3.4 Energy Preservation

Levin [1] points out that the energy in the slice $\int_{-0.5}^{0.5} |H(\omega_t, k_x)|^2 d\omega_t$ for a given k_x is proportional to L^2T, which defines a fixed budget over the integration interval. Given that the noise variance is equal to $H(0,0)$ and therefore proportional to LT, we see that:

$$\int_{-0.5}^{0.5} NEQ(\omega_t, k_x) d\omega_t = \frac{|S(k_x)|^2}{H(0,0)} \int_{-0.5}^{0.5} |H(\omega_t, k_x)|^2 d\omega_t \propto |S(k_x)|^2 L. \tag{7}$$

Hence, for each slice of spatial frequency k_x we have a fixed NEQ budget proportional to $|S(k_x)|^2 L$. The distribution of the NEQ budget across ω_t is determined by both the exposure time and the camera motion path. Recall that for a subject with velocity v, the direct relation $\omega_t = vk_x$ holds. Therefore, there is a fixed NEQ budget that we can distribute across the various velocities in a scene.

3.5 Deconvolution

The image captured by the sensor is blurred in a specific way due to the relative motion between the camera and the subject. This blur is removed through deconvolution, which is a division by the blur kernel response in the frequency domain. Note that if the frequency response of the blur kernel is zero for some spatial frequency, the deconvolution becomes unstable. We therefore use the Wiener deconvolution filter, which minimizes the Mean Squared Error (MSE) of the deconvolved image:

$$W(k_x) = \frac{H_v^*(k_x)|S(k_x)|^2}{|H_v(k_x)|^2 |S(k_x)|^2 + |N(k_x)|^2}. \tag{8}$$

We then derive the expression for the MSE:

$$\begin{aligned} MSE_v(k_x) &= \mathbb{E}\left(|S_v(k_x) - \hat{S}_v(k_x)|^2\right) \\ &= \frac{H_v(0)|S(k_x)|^2}{|H_v(k_x)|^2 |S(k_x)|^2 + H_v(0)} = \frac{|S(k_x)|^2}{1 + NEQ_v(k_x)}. \end{aligned} \tag{9}$$

We see that the MSE of the deconvolved image is inversly proportional to the NEQ and recall that there is a fixed NEQ budget (Equation (7)). Hence, we should distribute the NEQ budget in an efficient way in order to minimize the reconstruction error (MSE).

4 Optimal Camera Motion

In order to minimize the total MSE, we use the following two optimality criteria which were introduced by Bando *et al.* [9]:

- ◇ **Effectiveness:** The NEQ budget must only be spent on regions in $H_0(\omega_t, k_x)$ that correspond to velocities that appear in the scene;
- ◇ **Uniformity:** The NEQ budget must be distributed uniformly among all velocities in a scene.

To facilitate the analysis, we define the function $\rho(v)$ which indicates whether a velocity occurs in a scene ($\rho(v) = 1$) or not ($\rho(v) = 0$). The function $\rho(v)$ consists of a series of M rectangle functions (denoted by Π):

$$\rho(v) = \Sigma_{m=1}^{M} \Pi(\frac{v - v_m}{2\Delta v_m}), \tag{10}$$

where v_m denotes central velocities sorted in ascending order and Δv_m denotes the width of the interval around the corresponding velocity. For example, if is known that a vehicle moves at a speed of 20 pixels/second (px/s) with an uncertainty of ± 2 px/s, then $v_m = 20$ px/s and $\Delta v_m = 2$ px/s. The blocks should not overlap, hence: $v_{m+1} - v_m \geq \Delta v_m + \Delta v_{m+1}$.

Intuitively, the camera velocity should be equal to every velocity where $\rho(v) = 1$ for an equal amount of time. Levin *et al.* [1] show that this is the case if the camera field-of-view (FOV) moves with a constant acceleration and also proves that this camera motion optimizes the information capture. If we consider each of the blocks $\Pi(\frac{v-v_m}{2\Delta v_m})$ individually, we capture the necessary information using this constant acceleration motion. Each of these M motion paths is defined by the starting velocity $v_{m,1}$, ending velocity $v_{m,2}$ and time interval T_m as follows:

$$x_m(\tau) = \frac{v_{m,2} - v_{m,1}}{2T_m}\tau^2 + v_{m,1}\tau + C, \tag{11}$$

with C being an arbitrary offset. The motion paths corresponding to the separate blocks are then concatenated over time into a continuous motion path $x_c(t)$ that captures all the necessary information. The corresponding exposure time is determined by the sum of the M time intervals: $T = \Sigma_{m=1}^{M} T_m$.

Recall that the energy in a slice for a given k_x is fixed and $\omega_t = vk_x$. Hence, energy is spread out more as the spatial frequency k_x increases. As a result, the response $H_v(k_x)$ and consequently $NEQ_v(k_x)$ are minimal at the Nyquist frequency $k_x = \pm 0.5$. We define a minimum magnitude H_{min} at $k_x = \pm 0.5$, such that

$$\forall v : \rho(v) H_{min} \leq |H_v(0.5)| \leq |H_v(k_x)|. \tag{12}$$

We find values for $v_{m,1}$, $v_{m,2}$ and T_m such that Equation (12) holds while minimizing T_m. Minimizing T_m ensures that as little as possible of the available NEQ budget is spent on velocities where $\rho(v) = 0$, fulfilling the effectiveness criterion. Moreover, we also ensure that we do not overspend our budget on velocities where $\rho(v) = 1$. Hence,

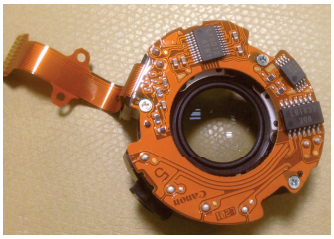


Figure 1: OIS Floating lens module. The lens can move in 2 axes.

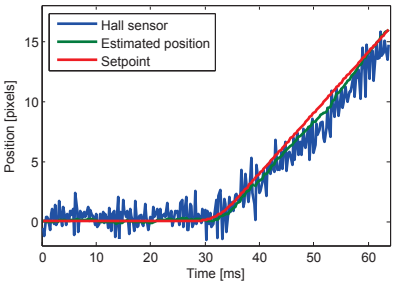


Figure 2: Lens motion trajectory for Shift Exposure with T=64ms.

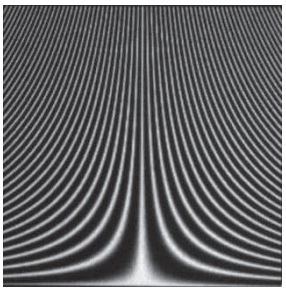


Figure 3: Test target pattern.

$|H_v(0.5)|$ will only be slightly larger than H_{min}, fulfilling the uniformity criterion as well.

We define a desired exposure time T_d, e.g. due to timing constraints, minimum video frame rate or clipping value of the detector. Next, we pick an initial value for H_{min} and do an exhaustive search for the camera motion path (i.e. triples of $v_{m,1}$, $v_{m,2}$ and T_m) that satisfy Equation (12) with the smallest possible T_m. This results in an exposure T. If the obtained $T < T_d$, we increase H_{min} and vice versa until $T = T_d$.

5 Lens Motion with Image Stabilization Hardware

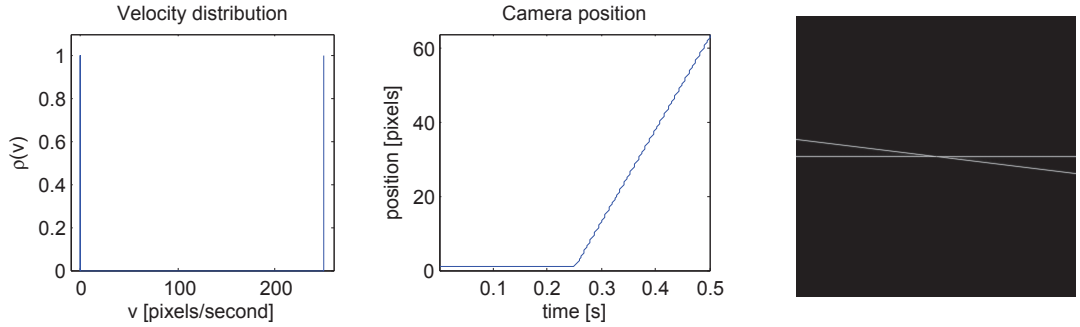
We use lens-shift OIS hardware in a Canon EF28-135 F/3.5-5.6 IS USM lens to move the field-of-view of our camera during the exposure. Lens-shift OIS systems shift a floating lens element in a plane perpendicular to the optical axis in order to compensate for unintended camera motion. The control loop measures rotation using gyroscopes in the pitch and yaw axis and shifts the lens in order to stabilize the image projected on the image sensor. We have implemented our own controller in order to control the lens position.

McCloskey *et al.* [4] implemented a similar system with a Canon EF70-200 F/4 IS USM lens. The controller design described in this section is in many ways similar to their design. We extract the floating lens module from the lens (Figure 1). The module contains voice coils for both the horizontal and vertical axis, while two magnets are mounted on the floating lens. The control logic is replaced by a standard MSP430F5438a microcontroller running a PID control loop.

Figure 2 displays an example of the lens motion. Note the noisy signal from the position sensors. A state observer on the microcontroller provides a cleaner position signal, displayed in green in Figure 2. The lens cannot accelerate instantly, therefore we ramp up the speed during 8 ms. The difference between the setpoint and the real position is within 1 pixel, which is sufficiently small for the practical test in the next section.

6 Case study

In this section we study a common practical case with a stationary background ($v_1 = 0$ px/s, $\Delta v_1 = 0$ px/s) and an object moving perpendicular to the camera with velocity $v_2 = 250$ px/s and $\Delta v_2 = 0$ px/s. The optimal camera motion path is shown in Figure 4b, alongside $\rho(v)$ and $H_0(\omega_t, k_x)$. During the remainder of this paper, we refer to this specific motion as *Shift Exposure*. An additional advantage of Shift Exposure is that the Point Spread Function is the same regardless of whether a subject moves at v_1 or



(a) Velocity distribution (b) Camera motion path (c) $H_0(\omega_t, k_x)$

v_2, removing the need for segmentation during deblurring. We compare the following camera motion paths:

1. No Motion (NM): A traditional camera.
2. Motion Invariant Photography (MIP). The camera velocity is linearly increased from v_1 to v_2 during the exposure.
3. Shift Exposure (SE). The camera velocity equals v_1 for $t < T/2$ and v_2 for $t > T/2$.

The exposure time T_d is varied between 8 ms and 128 ms. We compute the optimal camera motion path as described in Section 4 and evaluate at two different light levels with average photon flux $L_1 = 1.3 \cdot 10^3$ and $L_2 = 6.9 \cdot 10^3$ photons per second per pixel. Both are relatively low-light levels. The test target image is displayed in Figure 3. The highest spatial frequency on the test pattern corresponds to $k_x \approx 0.2$.

The camera output is simulated by computing the Point Spread Function based on the camera and subject motion paths. The test target image is blurred and noise is added afterwards. The noise model is described by Janesick [8] and includes photon noise, read noise, fixed-pattern noise and dark current. We then reconstruct using the Wiener filter that minimizes the MSE between the reconstruction and the original image.

For the practical experiment we employ the modified optical image stabilizer from Section 5. A blur-free and relatively low-noise reference image is obtained by stopping the test targets and taking a normal picture with $T = 128$ ms. This reference image is used to compute both the Wiener deconvolution filter, and the MSE of the deblurred image.

7 Discussion and Conclusion

The simulation results reveal how the image quality of a normal camera (No Motion) quickly degrades for moving subjects (Figures 5a and 5d). Whereas the MSE for still subjects decreases with increasing exposure time, the MSE of moving subjects increases for exposure time longer than 32 ms. Both MIP and SE motion paths ensure an approximately equal MSE for both still and moving subjects. MIP (Figures 5b and 5e) benefits from long exposure times, but the advantage diminishes for very long exposure times. On the other hand, the MSE of Shift Exposure (Figures 5c and 5f) keeps decreasing for longer exposure times.

At $L = L_1$, the MIP image achieves up to 2.9 dB improvement for the moving subject compared to the traditional camera, whereas Shift Exposure obtains 4.5 dB

improvement. At $L = L_2$, the improvement of MIP and SE are 2.7 dB and 5.2 dB, respectively.

The practical experiment results show similar trends. However, note that both MIP and SE benefit little of longer exposure times at the relatively bright light level $L = L_2$. The MSE of the MIP image even increases for long exposure times, which is likely due to inaccuracies in the test target and lens positions. At $L = L_1$, MIP decreases the MSE for the moving subject by up to 2.8 dB, whereas SE shows an improvement of 4.4 dB compared to the traditional camera. At $L = L_2$, the improvement of MIP and SE are 0.45 dB and 1.2 dB, respectively.

We conclude that lens motion can significantly improve the image quality of the deconvolved image. However, the advantage is limited in relatively bright circumstances. Although Motion-Invariant Photography requires relatively little prior motion information, it provides only a small improvement. Shift Exposure requires detailed prior information about the velocities in a scene, but it provides a significant performance gain of up to 4.4 dB in our practical experiments. The more accurate the prior information about the subject velocities (e.g. through motion estimation on prior frames), the better the image quality one can obtain by using the camera motion proposed in this paper.

Future work will compare the various image capture techniques using more sophisticated quality metrics than the Mean Squared Error, which does not accurately reflect the perceptual image quality. Furthermore, the concept may be extended to two-dimensional motion.

References

[1] A. Levin, P. Sand, T. S. Cho, F. Durand, and W. T. Freeman, "Motion-invariant photography," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3, p. 71, 2008.

[2] R. Raskar, A. Agrawal, and J. Tumblin, "Coded exposure photography: motion deblurring using fluttered shutter," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 795–804, 2006.

[3] S. McCloskey, K. Muldoon, and S. Venkatesha, "Motion aware motion invariance," in *Computational Photography (ICCP), 2014 IEEE International Conference on*, pp. 1–9, IEEE, 2014.

[4] S. McCloskey, K. Muldoon, and S. Venkatesha, "Motion invariance and custom blur from lens motion," in *Computational Photography (ICCP), 2011 IEEE International Conference on*, pp. 1–8, IEEE, 2011.

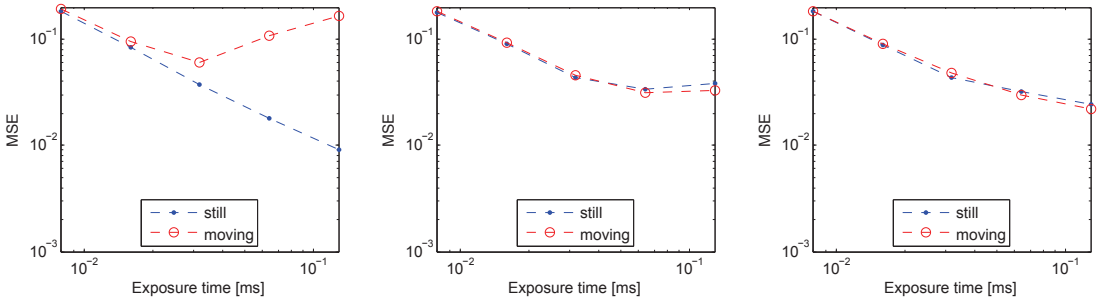
[5] O. Cossairt, M. Gupta, and S. K. Nayar, "When does computational imaging improve performance?," *Image Processing, IEEE Transactions on*, vol. 22, no. 2, pp. 447–458, 2013.

[6] S. McCloskey, "Improved motion invariant deblurring through motion estimation," in *Computer Vision–ECCV 2014*, pp. 75–89, Springer, 2014.

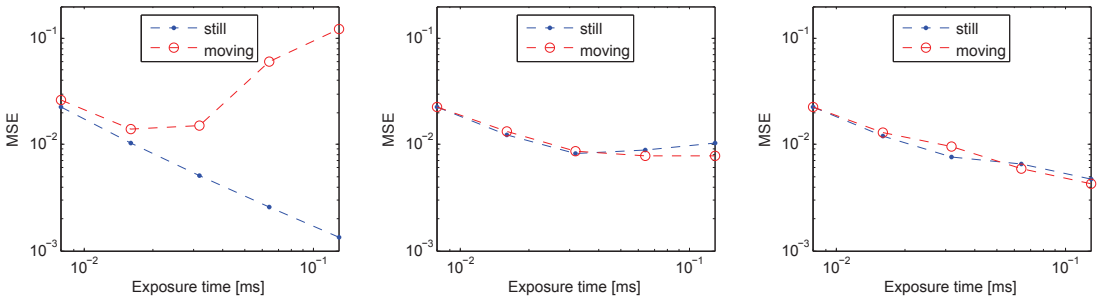
[7] B. W. Keelan, "Imaging applications of noise equivalent quanta," *Electronic Imaging*, vol. 2016, no. 13, pp. 1–7, 2016.

[8] J. R. Janesick, *Photon transfer*. SPIE press San Jose, 2007.

[9] Y. Bando, B.-Y. Chen, and T. Nishita, "Motion deblurring from a single image using circular sensor motion," in *Computer Graphics Forum*, vol. 30, pp. 1869–1878, Wiley Online Library, 2011.

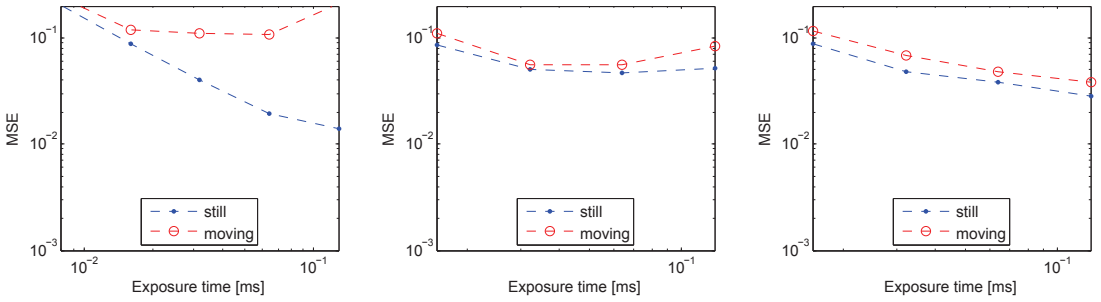


(a) $L = L_1$, No Motion (b) $L = L_1$, Motion-Invariant (c) $L = L_1$, Shift Exposure

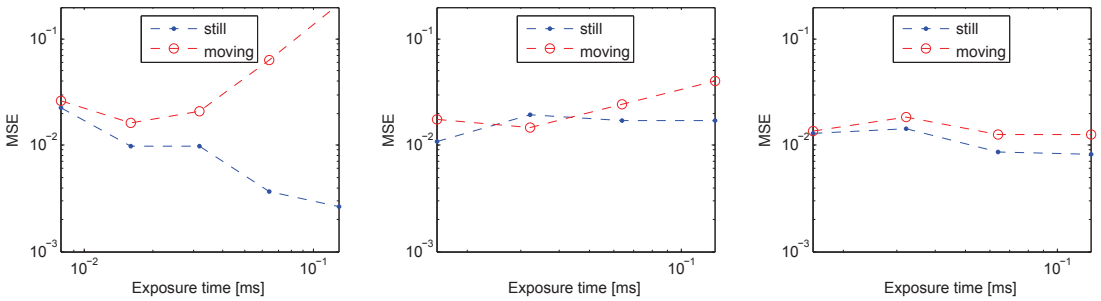


(d) $L = L_2$, No Motion (e) $L = L_2$, Motion-Invariant (f) $L = L_2$, Shift Exposure

Figure 5: Simulation results



(a) $L = L_1$, No Motion (b) $L = L_1$, Motion-Invariant (c) $L = L_1$, Shift Exposure



(d) $L = L_2$, No Motion (e) $L = L_2$, Motion-Invariant (f) $L = L_2$, Shift Exposure

Figure 6: Experimental results

Person-Independent Discomfort Detection System for Infants

C. Li[1], S. Zinger[1], W. E. Tjon a Ten[2], P. H. N. de With[1]
[1]Eindhoven University of Technology
5600MB, Eindhoven, the Netherlands
{c.li2, s.zinger, p.h.n.de.with}@tue.nl
[2]Maxima Medical Center
5500MB, Veldhoven, the Netherlands
w.tjonaten@mmc.nl

Abstract

Automatic discomfort detection for infants is important in healthcare, since infants have no ability to express their discomfort. We propose a video analysis system, based on supervised learning and classifying previously unseen infants from the testing set in a fully automated way. The first stage of our system consists of fame-based face detection, and then fit a face shape to the detected face area by using a Constrained Local Model (CLM). In the second stage, we analyze expression features by using Elongated Local Binary Patterns (ELBP), and classify expression features with an Support Vector Machine (SVM) for discomfort detection. The key contribution of our system is that the face model is infant-independent by employing a Constrained Local Model without prior knowledge about previously unseen infants. The system detects discomfort with an accuracy of 84.3%, a sensitivity of 82.4%, and specificity of 84.9% on the testing set containing videos of 11 infants. In addition, in order to increase the robustness of the system to head rotation, we introduce a face recovery method based on the symmetry of the face. With this step, the previous performance parameters increase by $3.1 - 3.8\%$ tested with videos of 2 infants containing $2,010$ frames.

1 Introduction

Discomfort and pain assessment for infants is an important and challenging task, since infants are not able to verbally communicate their discomfort. Neglecting discomfort or pain of infants can cause problems in their further development [1]. Therefore, nurses or parents normally take the responsibility of monitoring and assessing pain of infants when they are admitted to a hospital. Several behavioral and physiological changes can indicate potential discomfort of infants, such as crying, facial tension, frequent body movements, heart rate, respiratory response, blood pressure, and levels of oxygen and carbon dioxide in the blood.

To assess pain, various infant pain scoring methods have been developed, such as the PIPP (Premature Infant Pain Profile) and the COMFORT pain scale, which are based on the above-mentioned clinical parameters. All nurses responsible for infant pain assessment have to be trained to apply a pain scoring, while observing infants for a certain time period for pain estimation. However, monitoring by nurses is time consuming, expensive and, as assessments are done in intervals, changes in pain/discomfort intensity between the intervals are not detected. Therefore an automated discomfort detection system is highly attractive. Besides monitoring infants for a long period, such a system opens new possibilities for reliable disease diagnosis, such as Gastro-Esophageal Reflux Disease (GERD).

Automatic discomfort detection for infants based on video analysis is a challenging task for several reasons. At first, infants are often sleeping, so that their eyeballs are invisible. Therefore, many face detectors based on finding eyes cannot be used for this task. Secondly, most infants need pacifiers to stay calm, and this partially occludes

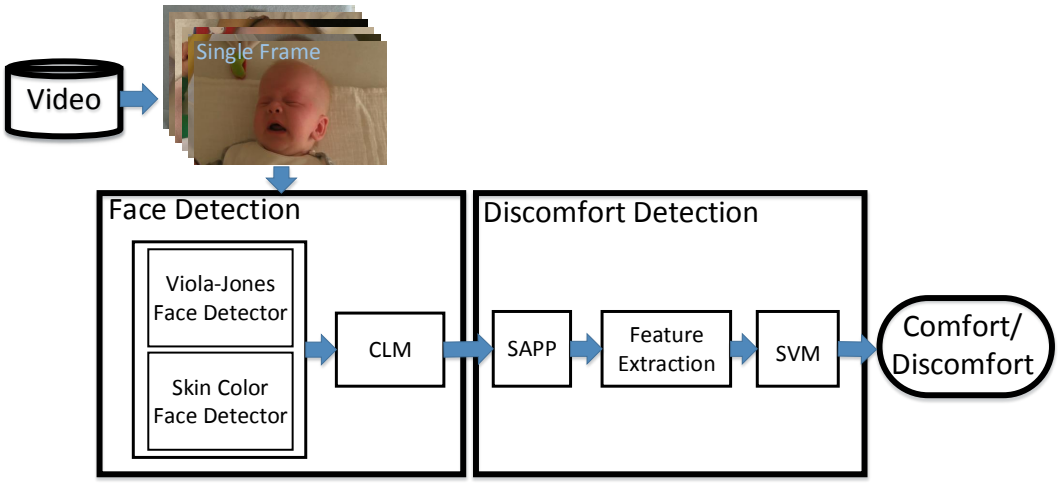


Figure 1. Block diagram of the discomfort detection system.

the features presenting facial expressions. Another complication is that discomfort can be detected in specific cases only and may be tuned to a particular child, which harms the general use of the system [2]. Since facial expression is a major feature that helps professionals to assess pain of infants, much attention has been given to analyze facial expressions. In [3], the authors propose a facial expression recognition system using an automated eye detection for face detection, together with a combination of Gabor features and Support Vector Machine (SVM) for expression classification based on the face coding system (FACS). However, infants, especially neonates, lie in their bed with their eyes closed, and an automated eye detection will fail for them. Lucey *et al.* [4] proposed a system for a pain intensity estimation based on an Active Appearance Model (AAM). They first extract shape and appearance features from AAM, and then train separate SVM classifiers for action units with these features. However, this approach requires manual labeling of the key frames of a sequence, so that the system is not fully automated. In this paper, we present our video analysis system that automatically monitors discomfort, but now in a generic way, independent of the observed infant.

We contribute to this research in two aspects. At first, our system is infant-independent and requires no prior knowledge about the previously unseen infants. Second, our system is validated on a clinical dataset. The paper is organized as follows. Section 2 explains each component of the system in detail. In Section 3, experimental results are shown for the system evaluation. Finally, conclusions are drawn in Section 4.

2 System Design

The system consists of two main components: face detection and discomfort detection, as shown in Fig. 1. The first part employs a combination of the Viola-Jones face detector and a skin-color detector for finding the face area. After that, we describe the shape of the face with a Constrained Local Model (CLM). Based on the concept of similarity-normalized appearance (SAPP), Elongated Local Binary Patterns (ELBP) are extracted as facial expression features. For classification, an Support Vector Machine (SVM) is used for distinguishing discomfort from comfort. A more detailed description of each block now follows.

2.1 Face Detection

In order to optimize the fitting of a shape model to the image, we first use a combination of a Viola-Jones face detector and a Gaussian mixture model skin color detector for locating the face area as proposed in [5]. In our system, we analyze frames in the YCbCr color space [2]. A false detection of a face directly causes a false detection of discomfort. Therefore, the system selects the detected area in the following way. If the detected area of a frame is smaller than a pre-defined threshold, then the system regards this detection as a false detection and discards this frame. Only the frames with the detected area larger than the threshold are passed on to the next stage. Then within the detected face area, we apply CLM for aligning a face shape to the frame.

A Constrained Local Model is a person-independent model compared to AAM [6]. It is a joint shape and texture model introduced by Cristinacce *et al.* in [7]. The shape is represented as a concatenated vector of X and Y coordinates, as follows:

$$\mathbf{X} = (X_1, ..., X_n, Y_1, ...Y_n)^T, \tag{1}$$

where n is the number of points in the shape of the CLM. The shape $\mathbf{X}$ can be expressed as a base shape $\bar{\mathbf{x}}$ and a linear combination of shape vectors $\mathbf{\Phi}$, denoted by

$$\mathbf{X} = T_t(\bar{\mathbf{x}} + \mathbf{\Phi}\mathbf{b}), \tag{2}$$

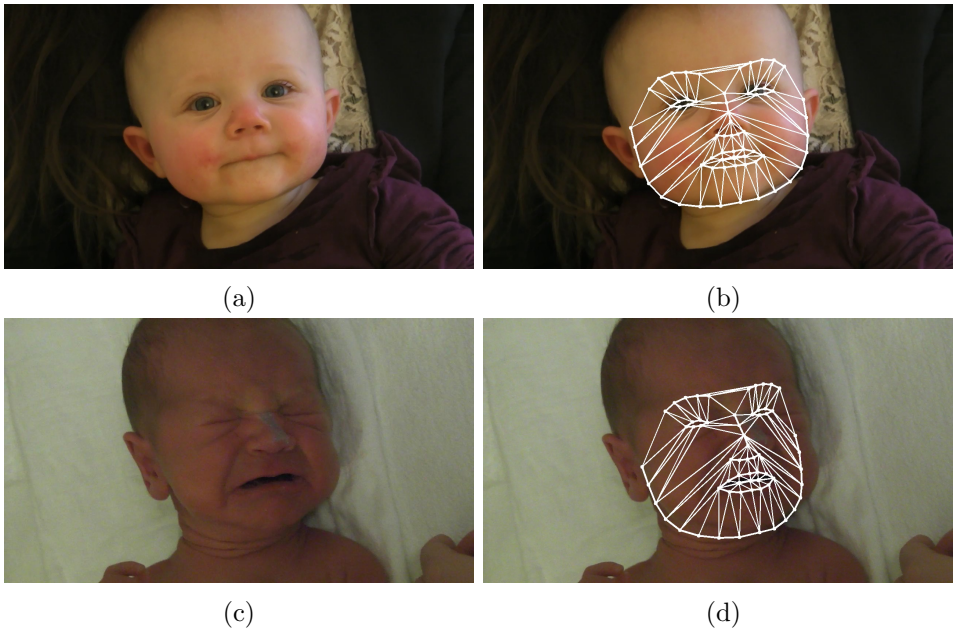
where the coefficients $\mathbf{b} = (b_1, ..., b_m)^T$ are the shape parameters. The shape model is normally computed from training data consisting of a set of images with the shape points marked manually. To obtain the base shape $\bar{\mathbf{x}}$ and the shape variation $\mathbf{b}$, the Procrustes alignment algorithm and Principle Component Analysis (PCA) are commonly used and also adopted here. To align the orientation and face scale, transformation T_t is introduced to cover these properties and it represents a similarity transform with the parameter t.

The CLM application is novel to the discomfort detection for infants with the described challenging conditions (pacifiers, no eyeballs). Hence, the fitting of the model is crucial. To fit the shape model to an image, a response map of each patch model is obtained by an exhaustive local search for each landmark around an initial shape using a feature detector. Then the shape parameters $\mathbf{b}$ are optimized by criteria that jointly maximize the detection responses over all the landmarks. In our system, we adopt the optimization strategy based on subspace-constrained mean-shifts, proposed by Saragih *et al.* [8]. An example of an original frame and a corresponding aligned face shape by CLM are shown in Fig. 2(a) and Fig. 2(b).

2.2 Discomfort Detection

Discomfort detection is performed in two steps: feature extraction and feature classification. These steps are described below.

Feature Extraction
Once the shape model is aligned to the image by CLM, facial features can be obtained based on similarity-normalized shape and SAPP defined by Lucey *et al.* [4]. SAPP refers to the face appearance after the removal of the rigid geometric variations and scale. This is achieved by warping the pixels of characteristic regions of the facial image into the similarity-normalized shape. A corresponding SAPP obtained from the face shape of Fig. 2(b) is displayed in Fig. 3(a). However, if the face is not frontal, the CLM fitting for the occluded face is not sufficiently accurate (see e.g. Fig. 2(c) and Fig. 2(d)). As a consequence, a distorted SAPP is obtained by the misalignment of the face shape (Fig. 3(b)). For the automated detection system, our contribution is to improve the robustness in applying the CLM when the faces are partly occluded. The



(a) (b)

(c) (d)

Figure 2. (a) Frame of the original video with frontal face; (b) Detected face shape of (a); (c) Frame of the original video with head rotation; (d) Detected face shape of (c);

(a) (b) (c)

Figure 3. (a): SAPP of the image shown in Fig. 2(a). (b): Distorted SAPP caused by face shape misalignment of the image shown in Fig. 2(d). (c): Restored SAPP of (b)

appearance in Fig. 3(b) is containing distortions at the wrinkles at the right eye and right side just above the mouth. In order to remove such errors, we create a frontal face orientation, and mirror the pixels from its visible part to the area where the face is occluded, thereby exploiting the symmetry of the face. This leads to a better restored SAPP, of which an example of is shown in Fig. 3(c).

ELBP is shown to have a very good performance on face recognition [9] [2]. Therefore, we extract ELBP features on SAPP for discomfort detection in our system. For ELBP calculation, the neighborhood pixels of the center pixel are defined as an ellipse with minor radius 1 in the vertical direction and major radius 2 in the horizontal direction around that pixel. The ELBP is calculated by comparing the gray value of the center pixel with the surrounding neighborhood pixels, as specified below:

$$ELBP(x) = \sum_{i=1}^{P} d(g_i - g_c)2^{i-1}. \tag{3}$$

Figure 4. Examples of frames from Database I and Database II.

Here parameter P denotes the number of the neighborhood pixels and g_c, g_i denote the gray values of the center pixel and the neighborhood pixels, respectively. The binary output function d is defined by

$$d(g_i - g_c) = \begin{cases} 1, & \text{for } (g_i - g_c) \geq 0; \\ 0, & \text{for } (g_i - g_c) < 0. \end{cases} \tag{4}$$

Feature Classification
SVM has been demonstrated to be useful in a number of facial recognition and expression recognition tasks [10]. This classifier defines the optimal hyperplane by maximizing the margin between positive and negative samples for a specified class. For detecting discomfort, we train an SVM with a linear kernel to distinguish discomfort from comfort, based on ELBP features extracted from SAPP, as described above, with the LIBSVM library [11]. The combination of the previous algorithm steps enables automated discomfort detection.

3 Experiments

In this section, we first introduce the databases used for training and testing our system. Then we describe the applied evaluation metrics. Finally, we present key performance parameters: accuracy, specificity and sensitivity for face detection and discomfort detection in our system.

3.1 Database

For training the discomfort classifier, we re-use the database from Fotiadou *et al.* [2]. This database consists of 15 videos of 8 infants, 5 displaying comfort, 1 discomfort and 9 videos containing both. Each video has a frame rate of 25 frames per second with a resolution of 1920×1080 pixels, which is denoted as Database I. For the purpose of testing, we also use a database of 106 videos from 38 infants that have been recorded at the Maxima Medical Center (MMC), Veldhoven, the Netherlands, by Slaats *et al.* [5]. Videos of the faces of infants experiencing pain resulting from various interventions, like a heel prick, placing an intravenous line, a venipuncture, a vaccination or from post-operative pain, are recorded. However, this database includes various children and situations, such as premature neonates, infants with tubes and pacifiers, and occlusions on the face. Therefore, we choose a collection of 13 videos from 11 infants without occlusion of the face, which we call Database II. From the 13 videos, 5 videos display discomfort, 7 videos display comfort and 1 video displays both. Each video has a frame rate of 30 frames per second and a spatial resolution of 1280×720 pixels. A description of both datasets is shown in Table 1. Examples of video frames from both databases are portrayed by Fig. 4.

Table 1. Database descriptions

	Infants	Videos	Total Frames
Database I	8	15	43,823
Database II	11	13	13,917

(a) (b)

Figure 5. (a) Manually labeled ground-truth image; (b) Detected face surrounded by a rectangle.

3.2 Evaluation Results

We evaluate our system in two aspects: face detection rate and discomfort detection rate. For face detection, we consider a detection to be correct when a rectangle encompassing the face mesh fitted in the image has a 70% overlap area with the rectangle in the ground-truth image. A ground-truth image has a manually annotated rectangle encompassing the eyes, nose and mouth. Fig. 5 shows an example of a manually labeled ground-truth image and a face mesh surrounded by a rectangle.

For discomfort detection, we evaluate the SVM classifier based on single frames. For each video frame in the database, the ground-truth of discomfort is provided by nurses from the Veldhoven MMC, who are experienced in giving pain scores. The performance of discomfort detection is obtained by comparing the output of the SVM classifier with the ground-truth. To measure the performance, we calculate accuracy, specificity and sensitivity of the classification results.

3.2.1 Face Detection

The Viola-Jones detector is trained with 25 infants from the database obtained by Slaats *et al.* [5], excluding the infant videos we use from Database II. Table 2 shows the performance of our CLM-based face detection system evaluated with Database I and Database II. In this table, "total frame" means all frames of each dataset, and "output frames" means frames with a face mesh detected by the face detection stage. It shows that the true detection rate over all frames is 70.9% and 81.5%. However, more importantly, the rate of true detection over the output from the CLM-based face detector is 91.6% and 96.9% for Database I and Database II, respectively.

3.2.2 Discomfort Detection

We have used infants included in Database I proposed by Fotiadou *et al.* [2] for training the discomfort detection classifier. For testing the performance of the discomfort detection, we apply infant videos from Database II. These infants are all unseen subjects to our discomfort classifier. Table 3 shows the performance of ELBP features extracted from SAPP for discomfort detection evaluated with Database II without face interpolation. It can be observed that our system achieves an accuracy of 84.3%, with a

Table 2. Performance of CLM-based face detection evaluated with Database I and Database II.

	Database I	**Database II**
Output Frames/Total Frames	77.1%	84.3%
True Detection/Total Frames	70.9%	81.5%
True Detection/Output Frames	91.9%	96.9%

Table 3. Performance of CLM-based ELBP+SAPP features evaluated with Database II.

Database II		
Accuracy	***Sensitivity***	***Specificity***
84.3%	82.4%	84.9%

Table 4. Comparison of ELBP+original SAPP and ELBP+mirrored SAPP.

	3 videos of 2 infants from Database II	
	ELBP+original SAPP	***ELBP+mirrored SAPP***
Accuracy	73.8%	77.4%
Sensitivity	85.6%	89.4%
Specificity	67.7%	70.8%

sensitivity of 82.4% and a specificity of 84.9% for unseen infants for discomfort detection. In order to evaluate the impact of face interpolation, we have manually chosen 3 videos of 2 head-rotated infants from Database II with a total of 2,010 frames, containing significant head rotations and movements. These videos form a worst-case test. This explains the lower performance scores in Table 4, which also shows that, with face interpolation, the accuracy increases by 3.6%, sensitivity by 3.8%, and specificity by 3.1%.

4 Conclusion

In this paper, we have proposed an automated and person-independent system to detect discomfort for infants. The proposed algorithm exploits CLM for infant face detection in a robust way without any prior knowledge. The robustness improvement is achieved by (1) training the CLM with a generic model that is person-independent, and (2) by solving misalignment errors in the model via face mirroring. The improved robustness brings the practical use of the system in a hospital environment where infants come and leave continuously, closer to reality. The system can detect discomfort with a score of an accuracy of 84.3%, a sensitivity of 82.4%, and a specificity of 84.9% for Database II, however, these numbers were not yet obtained for the full algorithm. False positives and false negatives for discomfort classification are mainly due to the misalignment of the fitting face shape to the image. With face interpolation, the system is more robust to head orientation. In near future, we will experiment on larger datasets, and extend the system to gastro-esophageal reflux disease patients, in order to assist the diagnosis procedure with our classification results.

References

[1] F. L. Porter, R. E. Grunau, and K. J. Anand, "Long-term effects of pain in infants," *J. Dev. Behav. Pediatr.*, vol. 20, pp. 253–261, 1999.

[2] E. Fotiadou, S. Zinger, W. E. Tjon a Ten, S. Oetomo, and P. H. N. de With, "Video-based facial discomfort analysis for infants," in *Proc. SPIE 9029, Visual Information Processing and Communication V, 90290F*, 2014, pp. 90 290F–90 290F–14.

[3] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in *Computer Vision and Pattern Recognition. CVPR. IEEE Computer Society Conference on*, vol. 2, June 2005, pp. 568–573.

[4] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. De la Torre, and J. Cohn, "AAM derived face representations for robust facial action recognition," in *Automatic Face and Gesture Recognition. FGR. 7th International Conference on*, April 2006, pp. 155–160.

[5] B. Slaats, S. Zinger, P. H. N. de With, W. Tjon a Ten, and S. Bambang Oetomo, "Video analysis for acute pain detection in infants," in *5th joint WIC/IEEE SP Symosium on Information Theory and Signal Processing*, May 2015, pp. 50–57.

[6] S. Chew, P. Lucey, S. Lucey, J. Saragih, J. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," in *Automatic Face Gesture Recognition and Workshops, IEEE International Conference on*, March 2011, pp. 915–920.

[7] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Britisch Machine Vision Conference (BMVC)*, 2006, pp. 95.1–95.10.

[8] J. M. Saragih, S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *International Conference of Computer Vision (ICCV)*, September 2009.

[9] S. Liao and A. Chung, "Face recognition by using elongated local binary patterns with average maximum distance gradient magnitude," in *Computer Vision ACCV*, ser. Lecture Notes in Computer Science, Y. Yagi, S. Kang, I. Kweon, and H. Zha, Eds. Springer Berlin Heidelberg, 2007, vol. 4844, pp. 672–679.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Wavelet-based coherence between large-scale resting-state networks: neurodynamics marker for autism?

Antoine Bernas, Evelien Barendse, Svitlana Zinger, and Albert P. Aldenkamp

Eindhoven University of Technology,
Department of Electrical Engineering,
Signal Processing Systems, VCA
P.O. Box 513, 56000 MB Eindhoven,
The Netherlands;
{a.b.bernas, s.zinger,
a.p.aldenkamp}@tue.nl

Radboud University Nijmegen,
Donders Institute for Brain, Cognition
and Behaviour,
P.O. Box 9101 6500 HB Nijmegen,
The Netherlands
e.barendse@donders.ru.nl

Abstract

Neurodynamics is poorly understood and has raised interest of neuroscientists over the past decade. When a brain pathology cannot be described through structural or functional brain analyses, neurodynamics based descriptors might be the only option to understand a pathology and maybe predict its symptomatic evolution. For example, adolescents or adults with autism have shown mixed results when their intrinsic structural and functional connectivity parameters in the brain at rest were assessed. To visualize neurodynamics parameters we use wavelet coherence maps, which show when and at which frequency two large-scale resting-state networks (RSNs) co-vary and display phase-locked behavior. Here the wavelet-based coherence coefficients are extracted from fMRI of adolescents with and without autism. More specifically, we introduce a novel metric: 'time of in-phase coherence' between pairs of resting-state networks. Results show that wavelet coherence maps can be used as neurodynamics maps, and that features such as 'time of in-phase coherence' can be calculated between pairs of resting-state networks. This wavelet-based metric shows actually weaker coherent patterns between the ventral stream and the executive control network in patient with autism.

1 Introduction

Wavelet transform and its cross-spectrum have been introduced in signal processing in the early 90's by Daubechies [1]. Its use is diverse. Usually it is applied as a filter tool, for compression, or transient detector in case of seizure pattern recognition in epilepsy for example [2]. Several trials have been done to assess brain signals in the frequency and time domain simultaneously – for observing transient or dynamic properties of a signal. However, as for the example of epileptic seizures, in neuroscience, researchers have focus on EEG signals. To our knowledge, only few studies were undertaken using fMRI-based wavelet scalograms (wavelet transform coefficient, or power, displayed in time-frequency space) [3, 4].
It has however been shown using wavelet-based metrics that resting-state fMRI BOLD signals undergo transient phases, and that the hypothesis of non-linearity of the fluctuation of the brain at rest is supported [5].

It remains unclear, though, whether in time series of larger scale of neuronal populations, such as resting-state networks, fluctuations display temporal correlations at different scale (periods), and whether these neurodynamics patterns differ from one population to another. Here, we study the correlation matrices in terms of wavelet coherence maps between pairs of RSN time-series. Wavelet coherence is an interesting measure first applied in geophysics by Torrence and Compo (1998)[6], extracted from wavelet transform and the cross-spectrum between two time series. With such maps, one can look at correlation coefficient matrices (time x frequency) showing phase-locked behavior (when adding cross-wavelet phase information). Hence, significant 'areas' in a wavelet-coherence scalogram based on RSN time series can reveal when and at which frequency two RSNs are correlated, and whether these correlations are in phase or anti-phase or even in between, i.e, in a sense of one RSN signal leading or lagging behind another. This provides us neurodynamics descriptors in a sense of causality.
We extract coherence maps from resting-state networks of adolescents with an autism spectrum disorder (ASD). As compared to their age- and IQ-matched typically developing peers (or controls), adolescents with autism displayed a less coherent covariation between the ventral stream and executive control network. The method used for extracting wavelet-based coherence descriptors and the results when applied upon the autistic brain are described in the following sections.

2 Methods

2.1 Resting-state ICA time series extraction

In order to conduct the wavelet-coherence-based analysis using resting-state time series, we first extract spatial RSN maps, using a group Independent Component Analysis (gICA) on resting-state fMRI preprocessed dataset. Then, a dual-regression is used to extract subject-specific RSN time series [7]. Finally we selected the most relevant RSNs, using an in-house built 'godness-of-fit' MatLab (The MathWorks, Inc., Natick, Massachusetts, United States) function and the resting-state networks NifTi template from Smith et al. (2009) [8].

2.2 Wavelet Coherence and time of in-phase coherence

Using the aforementioned selected RSN time series, one can build wavelet coherence maps between pairs of the time series. These maps represent localized correlation between two time series in the frequency (scale) and time space, also called scalograms. More details of how wavelet-coherence scalogram are constructed can be found in Torrence and Compo's paper (1998) [6]. Briefly, they define

$$R^2(s,t) = \frac{|\langle s^{-1} W^{XY}(s,t)\rangle|^2}{\langle s^{-1}|W^X(s,t)|^2\rangle \langle s^{-1}|W^Y(s,t)|^2\rangle} \qquad (1)$$

as a measure of coherence between two signals X and Y in the wavelet domain, i.e., after a wavelet transform $W^X(s,t)$ and $W^Y(s,t)$ was applied on the signals. As a cross-correlation, or a cross-spectrum in the Fourier space, we can extract the wavelet

cross-spectrum between the two signals, resulting in $W^{XY}(s,t) = W^{X}(s,t).W^{Y*}(s,t)$ (* denotes the conjugate). The operator ⟨.⟩ denotes the smoothing factor (in frequency and time) applied to reduce the edge effects.
Also, as we use the complex Morlet wavelet, the phase difference between X and Y is calculated with $\arg\left(R^2(s,t)\right) = \tan^1(Im(\mathrm{W}^{XY})/(Re(\mathrm{W}^{\mathrm{XY}})$. When applied on pairs of RSN time series, subject-specific wavelet-coherence scalograms are extracted (Figure 1).
By combining the wavelet-coherence definition (1) and the phase information given by $\arg\left(R^2(s,t)\right)$ we measure the average of time of in-phase coherence per scale (or periods) $c(s)$ as below.

$$c(s) = \frac{100}{N}\sum_{t=1}^{N} I\{R^2(s,t) > a_{95}\}.I\left\{-\frac{\pi}{4} < \arg\left(R^2(s,t)\right) > \frac{\pi}{4}\right\} \quad (2)$$

Here, I{.} is 1 if the condition between brackets is true, and 0 otherwise. The level of statistical significance is estimated using Monte Carlo methods with 1000 surrogate data set pairs (red noise background; more details in [9]). This metric in % (of total scan time N) is then used to assess brain dynamics between RSNs among frontotemporal networks and the default mode network of adolescents with autism.

3 Experimental results

3.1 Participants and Data acquisition

15 adolescents with ASD and 18 age- and IQ-matched controls participated in this study. Due to signal distortions caused by their braces, 2 participants with ASD and 5 control individuals were excluded from data analysis. Further in the study, an adolescent with ASD is also excluded because of a bad registration during the preprocessing of the resting-state fMRI data. In both groups participants' age ranged between 12 and 18 years old with an average of 15.3 and 14.5 years of age for ASD and controls respectively. Written informed consent was also obtained from the next of kin, caretakers, or guardians on behalf of the adolescents enrolled in this study. The study protocol was approved by the Medical Ethical Commission of the Maastricht University Medical Center (MUMC).

MRI was performed on a 3.0-Tesla unit (Philips Achieva) equipped with an 8-channel receiver-only head coil. For anatomical reference, a T1-weighted 3D fast (spoiled) gradient echo sequence was acquired with the following parameters: repetition time (TR) 8.2 ms, echo time (TE) 3.7 ms, inversion time (TI) 1022 ms, flip angle 8°, voxel size 1x1x1 mm3, field of view (FOV) 240x240 mm2, 150 transverse slices. Then, resting-state fMRI data was acquired using the whole brain single-shot multi-slice BOLD echo-planar imaging (EPI) sequence, with TR 2 s, TE 35 ms, flip angle 90°, voxel size 2x2x4 mm3, matrix 128x128, 32 contiguous transverse slices per volume, and 210 volumes per acquisition; resulting in total a resting-state acquisition of 7 minutes. For the resting-state scan, participants were instructed to lie with their eyes closed, to think of nothing but not to fall asleep. Two resting-state scans were acquired with a memory task-based fMRI in between.

3.2 Preprocessing and RSN time-series extraction

In order to extract subject-specific RSN maps, and their associated time-series, a gICA is conducted. First the following preprocessing steps are applied, using FEAT a software from FMRIB Software Library (FSL; www.fmrib.ox.ac.uk/fsl): discard of the first 3 volumes (= 6 s) allowing the magnetization to reach equilibrium; rigid-body motion correction; non-brain tissue removal; slice-timing correction; registration to the Montreal Neurological Institute (MNI) standard space (2 mm isotropic); spatial smoothing using a Gaussian kernel of 4.0 mm full-width at half-maximum (FWHM); grand-mean intensity normalization; and high-pass temporal filtering at 100 s (0.01 Hz). Then, using the FSL MELODIC tool, a temporally concatenated probabilistic ICA was applied upon all participant's fMRI scans, resulting in 34 independent component maps (our RSNs). Finally, we extracted 11 subject-specific relevant RSN maps, and their associated time-series using the method described in section 2.1. Spatial maps of these 11 relevant RSNs are available in the technical report [10]. Furthermore, based on prior results of the spatial and dynamic RSN connectivity [10,11], we only included network encompassing mostly frontal and temporal cortices, namely, the ventral stream (VENT) network, the control-executive network (EXE), the default-mode network (DMN), and the auditory system (AUDI). Figure 1 shows these three RSN activation maps.

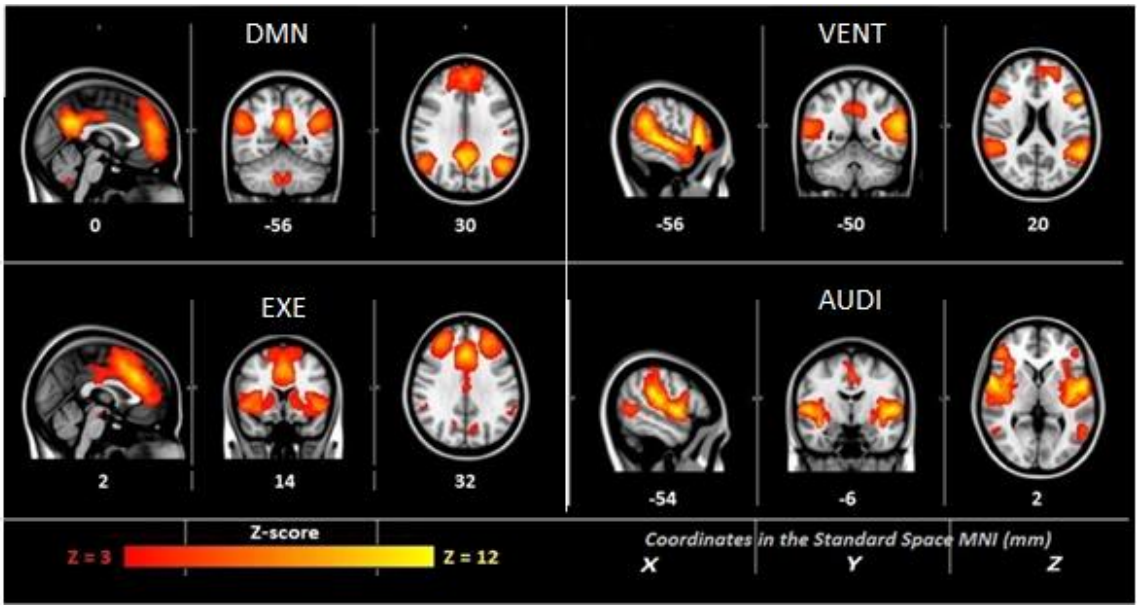


Figure 1. Four fronto-temporal relevant components. The ventral stream (VENT), executive (EXE), and default mode network (DMN) were extracted from group-IC maps overlaid in color on the MNI standard brain (2x2x2mm). Colorbar is thresholded between 3 and 15 (z-score). MNI coordinates are in mm. The left hemisphere corresponds to the right side in the images (radiological convention).

With the dual-regression method implemented in FSL [7], we extracted from the RSNs (Figure 1) a subject specific RSN time series. By using this time series one can conduct a wavelet analysis as explained in the next section.

3.3 Wavelet transforms, cross-spectrum, and coherence

Wavelet coherence maps, or scalograms, are calculated using the method and toolbox from Grindsted *et al.,* (2004) [9] (http://www.glaciology.net/wavelet-coherence). We use the complex Morlet wavelet as it has best Fourier period-wavelet scale ration of 1.03 (wavelet scale ≈ Fourier period), which helps interpreting results in frequency domain. Also, it is a complex wavelet and therefore gives us phase information, and it allows us to have directionality in the dynamic between signals (in-phase, leading, lagging, or anti-phase).

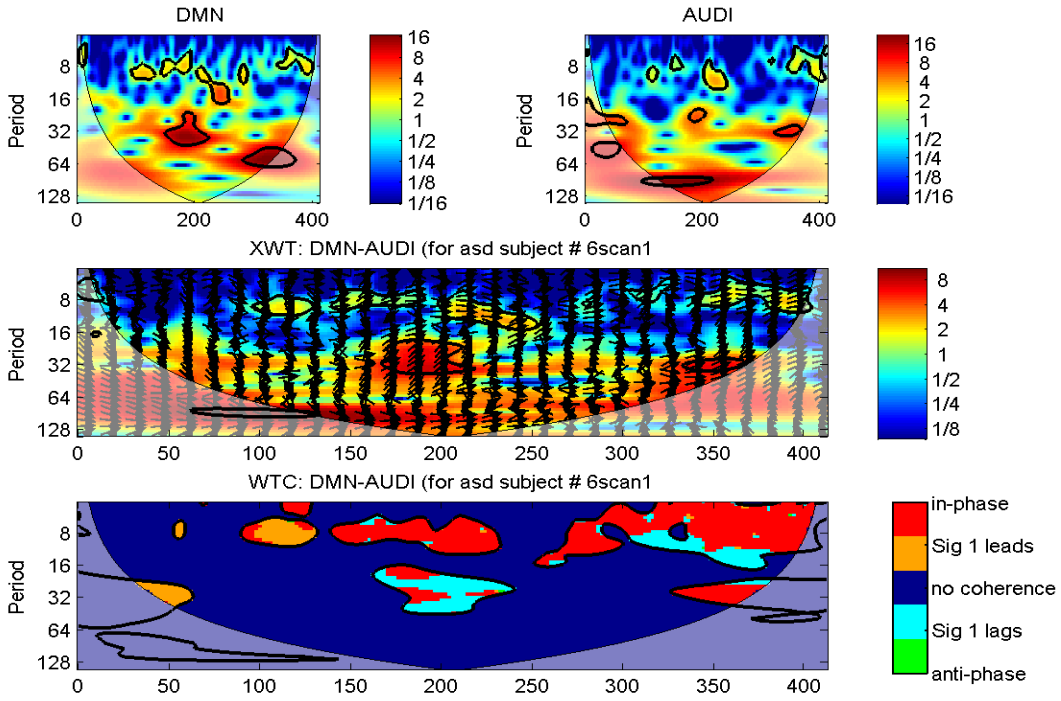


Figure 2. Example of wavelet scalograms. continuous wavelet transform on two resting-state signals: default mode network (top-left scalogram) and auditory system (top-right scalogram), and their cross wavelet transform (middle scalogram), and wavelet coherence (bottom scalogram) for the 1st scan of the 6th ASD subject. Arrows in the middle scalogram represent phase lags between the two signals. An arrow pointing to the right means that the signals are in-phase; pointing to the left: anti-phased; 90° downward: DMN leads AUDI; and 90° upward: AUDI leads DMN. This is also marked with colors in the coherent areas of the WTC, when taking the four phase difference ranges (in rad.): [0 ± π/4]; [π ± π/4] ; [-π/2 ± π/4]; [π/2 ± π/4]. For each scalogram, x-axis is time space, y-axis is scale space (in Fourier periods).

We obtained scalograms of in-phase wavelet-coherence (Figure 1, bottom scalogram, red area), for each subject and for each pair of relevant RSN time series. The ration of the red area over the full time length (7 min) of a row (a specific period) of these wavelet-coherence scalograms represents our time of in-phase coherence (in % of scan length).

Figure 3 represents for each group the group-averaged time of in-phase coherence per period, or $(1/N_{asd/con}) \sum_{i=1}^{N_{asd/con}} c(s)_{i\_asd/con}$, where $c(s)_{i\_asd/con}$ is the time of in-phase coherence (2), at the scale s, for the i[th] subject with ASD (i_asd) or control (i_con); and $N_{asd/con}$ is the sample size for ASD or controls (here, $N_{asd} = 12$; $N_{con} = 12$). These group-based measures of coherence are also repeated upon the second resting-state scan.

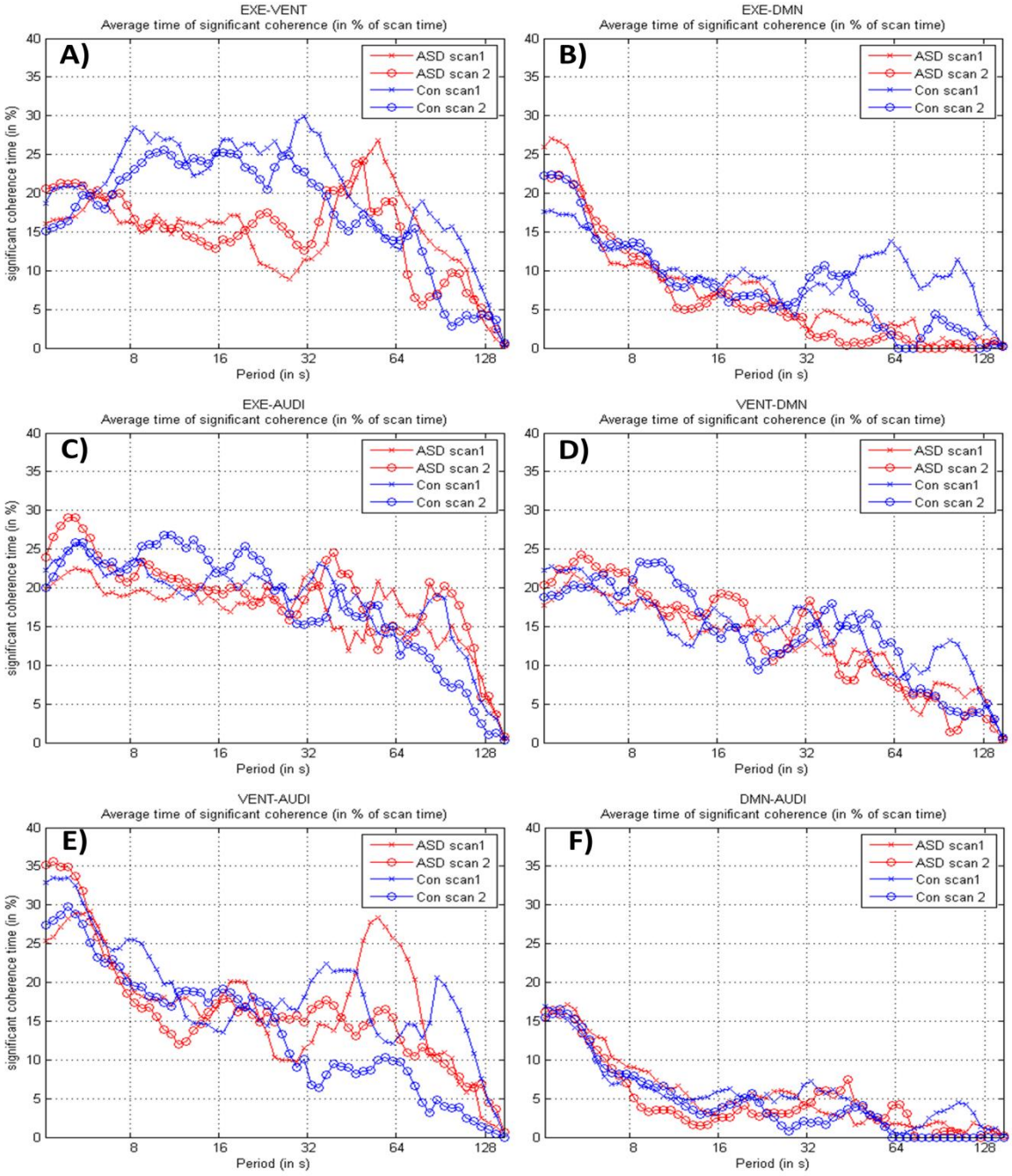


Figure 3. Average for ASD and controls, of their time of in-phase coherence (in % of scan time, Y-axis) per period (X-axis) for the pairs of networks: A) EXE-VENT, B) EXE-AUDI, C) EXE-DMN, D) VENT-AUDI, E) VENT-DMN, and F) DMN-AUDI. Blue lines are for controls and red lines for ASD participants. Crossmarks are for the 1st scan whereas circle represent the 2nd scan.

The main result in Figure 3 is that ASD adolescents show less correlation between two network time series EXE and VENT for a wide range of periods (from 8 s to 32 s, Figure 3.A). This reflects our previous findings in the report [10], where it is described that, using multivariate conditional Granger causality (measure of directed causality between multiple signals), these two networks presented a weaker effective connectivity in the direction VENT -> EXE. These two networks encompass brain areas coding for emotional processess, social interaction and executive function (cognition or behavioral outcome) – brain functions known to be a weak point in ASD.

4 Conclusion

Using the well-described wavelet coherence upon fMRI brain signals, we can visualize localized correlations in time and frequency space, between two functional networks.
We present a metric of average time of in-phase coherences and show that this could have future application in assessing neurodynamics, and gain knowledge in the transient properties of large-scale network oscillations.
Also, our features of average time of coherence can potentially describe and possibly even diagnose a chronic or a developmental brain disorder where structural and functional connectivity maps are incapable of distinguishing patients from healthy controls. .
Here we apply wavelet coherence upon functional networks extracted from an ICA decomposition using resting-state fMRI of adolescents with and without ASD. Results show that two networks coding for executive function and emotional processes display discrepancies between ASD and control adolescents. Finally other measures such as time of anti-phase coherences, or occurrences of continuous coherences per frequency can be worth investigating, especially concerning their role as brain dynamics descriptors for classifying/diagnosing a pathology.

References

[1] I. Daubechies, “The wavelet transform, time-frequency localization and signal analysis,” Inf. Theory, IEEE Trans., vol. 36, no. 5, pp. 961–1005, 1990.

[2] P. Indic and J. Narayanan, “Wavelet based algorithm for the estimation of frequency flow from electroencephalogram data during epileptic seizure,” Clin. Neurophysiol., vol. 122, no. 4, pp. 680–686, 2011.

[3] K. Müller, G. Lohmann, J. Neumann, M. Grigutsch, T. Mildner, and D. Y. von Cramon, “Investigating the wavelet coherence phase of the BOLD signal,” J. Magn. Reson. Imaging, vol. 20, no. 1, pp. 145–152, 2004.

[4] R. X. Smith, K. Jann, B. Ances, and D. J. J. Wang, “Wavelet-based regularity analysis reveals recurrent spatiotemporal behavior in resting-state fMRI,” Hum. Brain Mapp., vol. 3620, no. October 2014, pp. 3603–3620, 2015.

[5] C. Chang and G. H. Glover, “Time-frequency dynamics of resting-state brain connectivity measured with fMRI,” Neuroimage, vol. 50, no. 1, pp. 81–98, 2010.

[6] C. Torrence and G. P. Compo, “A practical guide to wavelet analysis”. Bull. Amer. Meteor. Soc., 79, 61–78, 1998

[7] C. F. Beckmann, C. E. Mackay, N. Filippini, and S. M. Smith, “Group comparison of resting-state FMRI data using multi-subject ICA and dual regression,” Hum. Brain Mapp. Conf., p. 181, 2009.

[8] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, and C. F. Beckmann, “Correspondence of the brain’s functional architecture during activation and rest.,” Proc. Natl. Acad. Sci. U. S. A., vol. 106, no. 31, pp. 13040–5, 2009.

[9] A. Grindsted, J. C. Moore, and S. Jevrejava, “Application of the cross wavelet transform and wavelet coherence to geophysical time series,” Nonlinear Process. Geophys., vol. 11, pp. 561–566, 2004.

[10] A. Bernas, “Neurodynamics in functional MRI and its clinical application”, Eindhoven University of Technology, SAI Tech. Rep., Feb. 2016. ISBN: 978-90-444-1444-8

[11] L. Q. Uddin, K. Supekar, C. J. Lynch, A. Khouzam, J. Phillips, C. Feinstein, S. Ryali, and V. Menon, “Salience network-based classification and prediction of symptom severity in children with autism.,” JAMA psychiatry, vol. 70, no. 8, pp. 869–79, 2013.

TouchSpeaker, a Multi-Sensor Context-Aware Application for Mobile Devices

Jona Beysens[1], Alessandro Chiumento[1], Sofie Pollin[1], Min Li[2]

[1]KU Leuven, Electrical Engineering, Telemic
Kasteelpark Arenberg 10 - bus 2444, Leuven

[2]NXP Semiconductors Belgium N.V.
Interleuvenlaan 80, 3001 Heverlee

jona.beysens@student.kuleuven.be, allessandro.chiumento@esat.kuleuven.be,
sofie.pollin@esat.kuleuven.be, min.li@nxp.com

Abstract

Tapping with your finger on any place on your mobile device is a promising candidate for enhanced interaction between users and their mobile device. So far the touchscreen and the accelerometer are commonly used to infer finger tap events. However, the touchscreen consumes a significant amount of power and is not always accessible (i.e., when device is used as running assistant). The accelerometer can be power efficient but can't differentiate well between a variety of contexts and positions. To address these limitations, we present *TouchSpeaker*, a novel technique for finger tap detection on mobile devices using the built-in speakers as primary sensors. We show that a combination of the speakers with other built-in sensors can distinguish between 9 different tap events with an accuracy of 98.3%, outperforming the state of the art. In addition, a robust version is implemented resulting in a false positive rate below 1%. For power constrained devices, we propose a configuration consisting of only the speakers and the accelerometer, achieving an accuracy of 95.3%.

1 Introduction

1.1 Problem statement

Nowadays, most people use the touchscreen to interact with their mobile device. However, imagine you want to interact with your mobile device when it is in the pocket or use it as a running assistant. In these cases, the touchscreen is not directly accessible or easy to use. Finally, to provide a desired action on a device from standby mode, the user usually needs to wake up the screen, unlock the device and navigate to the target application, which introduces a significant overhead. Tapping with your finger on any place on the device provides a solution to these problems. It allows users to interact with their device in a fast way without looking at it.

1.2 Earlier work

Earlier work in tap detection makes use of various mobile device's built-in sensors [1, 2]. Harisson et al. presented *TapSense*, a system that is able to identify whether the user interacts with the tip, pad, nail or knuckle [3]. The work relies on the unique acoustic signatures of these different parts of the finger when striking a touch surface. Both touchscreen and microphone should be on, resulting in significant power consumption. Next to this, McGrath et al. use the motion sensors to detect side taps and infer their location [4]. By probabilistically combining the estimate of the hand posture and tap location, an accuracy up to 97.3% is obtained. However, this method doesn't work in

situations where the mobile device can't move in the direction of the tap, because in that case there is no excitation observed in the motion sensors (i.e., tap on front when device is on table). Finally, Zang et al. combined the microphone, gyroscope and accelerometer to develop *BeyondTouch*, an application that extends the input experience by sensing one-handed, two-handed and on-table interactions [5]. Although a back tap can be detected with 97.92% accuracy in the one-handed scenario, a reduction in performance of the algorithm in real life is expected since no negative data (examples of no-taps) was included in the data set.

This paper introduces *TouchSpeaker*, a novel technique for finger tap detection using the built-in speakers as primary sensors. Speakers used as sensors (i.e., as microphones) can overcome the aforementioned limitations; these are passive devices, consuming no power for sensing. Furthermore, the mass of the diaphragm (shown in figure 1) is higher than the one in the microphone, resulting in a higher sensitivity for lower frequencies. These frequencies contain most of the information in a tap, as illustrated in figure 2. Finally, in contrast to the accelerometer, the speaker generates an excitation for every physical interaction, even when the device is on a flat surface. To increase the accuracy and the robustness to a level beyond the state of the art, the speakers are fused with other built-in sensors.

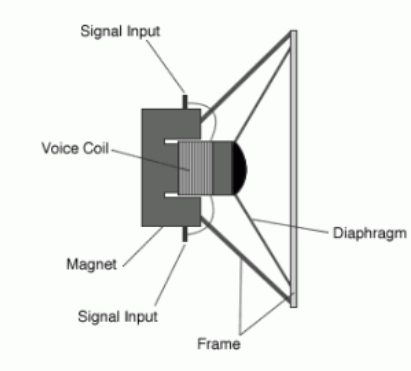


Figure 1: Physical model of a speaker [6]

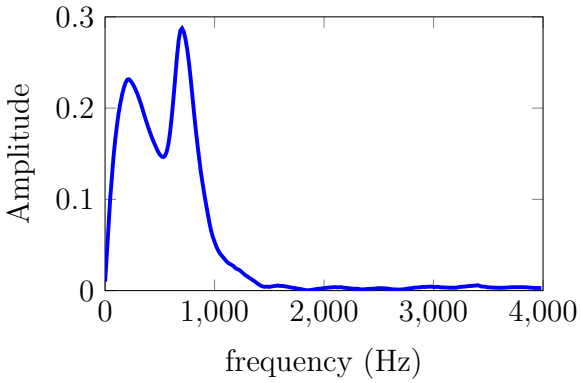


Figure 2: Frequency response of a speaker used as sensor in case of a tap event

Section 2 outlines the chosen approach to exploit the speakers as primary sensors and the combination with other built-in sensors. The results of the designed framework are presented in section 3. Section 4 describes the conclusions of this work.

2 Approach

2.1 Framework

In this section, we explain the design of the framework. The architecture is depicted in figure 3. The experimental setup contains a mobile device and a computer. Assume as an example that the user holds the device in his left hand and performs a tap on the top of the screen with the index finger of his right hand. The mobile device acquires data from the speakers and other built-in sensors in frames of 0.5 seconds and sends them to the computer. If the signal in one of the speakers exceeds the detection threshold (based on a moving average), the computer extracts characteristic features for each sensor in that frame and combines them using feature level fusion [7]. In the next step, the most relevant features are chosen using a feature selection method. The classifier

uses these to determine the context and the position of the tap, which can lead to a useful action on the device (e.g., take a selfie).

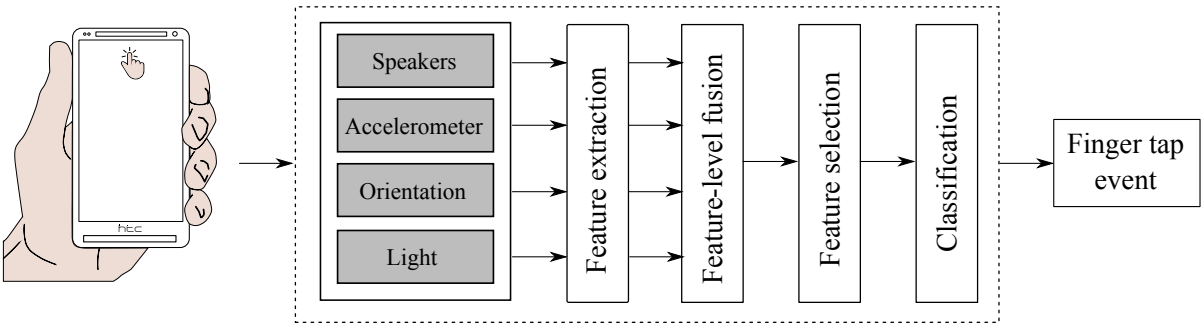


Figure 3: Structure of designed framework

2.2 Implementation

Data recording Table 1 shows the set of finger tap events that *TouchSpeaker* supports. All events (except no-tap) require physical interaction with the device. This means that taps on the surface next to the device (e.g., when on table) are rejected, because it is too difficult for the classifier to distinguish between taps on the device and taps next to the device. Furthermore, when multiple devices on a table run *TouchSpeaker*, all of them would unintentionally trigger an action when a tap is performed. Tap events can be useful in a variety of scenarios:

- ◇ take a selfie: touching a button on the screen when taking a selfie can be uncomfortable. A more convenient way to do this is by tapping on the back of the device.
- ◇ reject a call: you are in a meeting and the device in your pocket starts ringing. Taking out the device, mute it, and put it back can be cumbersome. By tapping on the pocket, you can reject the call without even looking at the phone.

An HTC One M7 is used as mobile device. Since a protective case around the device influences the acoustic signature of a tap, data is extracted with no case, a back case and a flip case. For each type of case, 10 participants were recruited. Every person performed 10 taps per finger tap event and per case, resulting in a total database of $3 \times 10 \times 10 \times 9 = 2700$ taps. Both positive (i.e., taps) and negative instances (i.e., no-taps) were included in the database. Examples of no-taps include: inserting the connector for charging, putting the device in the pocket, placing an object next to the device when it is on table. No feedback was given during the data extraction to ensure no change in the tapping behavior of the participants.

Features An overview of the features is presented in table 2. Both time and frequency domain features are considered. After extracting 36 features in total, they are combined into one large feature vector. Feature selection is performed in the software workbench WEKA [8], using a filter based approach and the information gain measure as evaluator of each candidate subset.

Classification *TouchSpeaker* is equipped with a Support Vector Machine (one versus all strategy, radial basis function kernel), Decision Tree (maximal 40 splits) and Random forest (with 100 decision trees) as supervised classification algorithms. They

Scenario	Device Orientation	Tap Position
Table	Front	Top
	Front	Bottom
	Back	Top
	Back	Bottom
Hand	Front	Top
	Front	Bottom
	Back	Top
Pocket	N/A	N/A
No-tap	N/A	N/A

Table 1: Overview of possible finger tap events

Sensor	#Feat.	Examples
Top & bottom Speakers	18	Difference amplitude, energy in tail, cumulative density function of frequency response
Accelerometer	13	Amount of peaks, sign of peaks, energy
Orientation	4	Energy in pitch
Light	1	Mean

Table 2: Overview of sensors and features

have 9 classes, corresponding to the 9 possible tap events. After training on the data set described above, they are evaluated based on 10-fold cross validation and leaving-one-user-out (LOUO) cross validation. The 2nd method measures the predictive performance on new users that are not included in the training data, leading to more realistic results compared to random cross validation.

3 Experimental Results and Discussion

This section presents the results of the designed framework. Accuracy is reported as the sum of the correctly classified instances over the total amount of instances.

3.1 General performance

Table 3 shows the accuracy and complexity of the different classifiers. All sensors are used together and no feature selection is performed. The complexity of the training and testing phase (i.e., classifying a new instance) is measured using Matlab Profiler [9]. Although the Random Forest takes most time to classify a new unseen instance based on the built model, it achieves the best results. Therefore, this classifier will be analyzed in the remainder of this section.

Classifiers	Accuracy(%)		Complexity(s)	
	10-fold CV	LOUO CV	Training	Testing
SVM	96.2	91.4	5.68	0.17
Decision Tree	92.7	88.3	1.17	0.09
Random Forest	98.3	95.8	4.92	0.66

Table 3: Accuracy and complexity of classifiers

Table 4 presents the 9×9 confusion matrix C using 10-fold cross validation for all possible events. The abbreviations in the table originate from the concatenation of

scenario, orientation and position (i.e., TFT stands for Table-Front-Top, NT for No-Tap). There is most confusion observed between taps and no-taps, shown in the last row and column of the matrix. This is due to the high similarity between taps on the device when on table and taps next to the device which are included as negative instances. Common measures used to analyze this in binary classification are the false positive rate (FPR) and false negative rate (FNR). Since we are dealing with a multi-class problem, these measures are defined here as follows ($M = 9$):

$$FPR = \frac{\sum_{j=1}^{M-1} C(M,j)}{\sum_{j=1}^{M} C(M,j)} \quad , \quad FNR = \frac{1}{M-1} \sum_{i=1}^{M-1} \frac{\sum_{j=1}^{M-1} C(i,M)}{\sum_{j=1}^{M} C(i,j)} \tag{1}$$

The FPR examines the amount of no-taps that are classified as any kind of taps whereas the FNR investigates the amount of any kind of taps that are classified as no-taps. Since incorrectly classifying no-taps as taps is perceived worse than vice versa, we will focus here on the FPR, which amounts to 6.0% in this confusion matrix.

		Predicted Tap Event								
		TFT	TFB	TBT	TBB	HFT	HFB	HBT	P	NT
Actual Tap Event	TFT	98.7	1.0	0	0	0	0	0	0	0.3
	TFB	1.3	97.7	0	0	0	0	0	0	1.0
	TBT	0	0	98.3	1.7	0	0	0	0	0
	TBB	0	0	0.7	99.3	0	0	0	0	0
	HFT	0	0	0	0	99.3	0.3	0.4	0	0
	HFB	0	0	0	0	0	100.0	0	0	0
	HBT	0	0	0	0	0.7	0	99.0	0	0.3
	P	0	0	0	0	0	0	0	100.0	0
	NT	1.0	2.0	0.6	1.0	0	0.7	0	0.7	94.0

Table 4: Confusion matrix using 10-fold cross validation

Furthermore, there is also confusion observed between taps at the top and at the bottom when the device is on the table. This can be explained by the distinct acoustic signature of the surface. Depending on the type of the surface and the exact position of the phone on this surface, a different speaker signal is observed. This is illustrated in figure 4. This was less visible for taps with a back case or flip case, because of the damping caused by the case. If the aim is to detect only taps in the hand and in the pocket, an accuracy of 99.8% and 97.5% is accomplished using 10-fold cross validation and LOUO cross validation respectively.

3.2 Improvement in robustness

Two approaches are used to lower the false positive rate and increase the robustness. Cost-sensitive classification assigns a relative higher cost to false positives compared to other errors (i.e., non-diagonal elements in the first 8 rows of the confusion matrix), making it more costly to classify a no-tap as a tap. Figure 5a shows that the FPR can

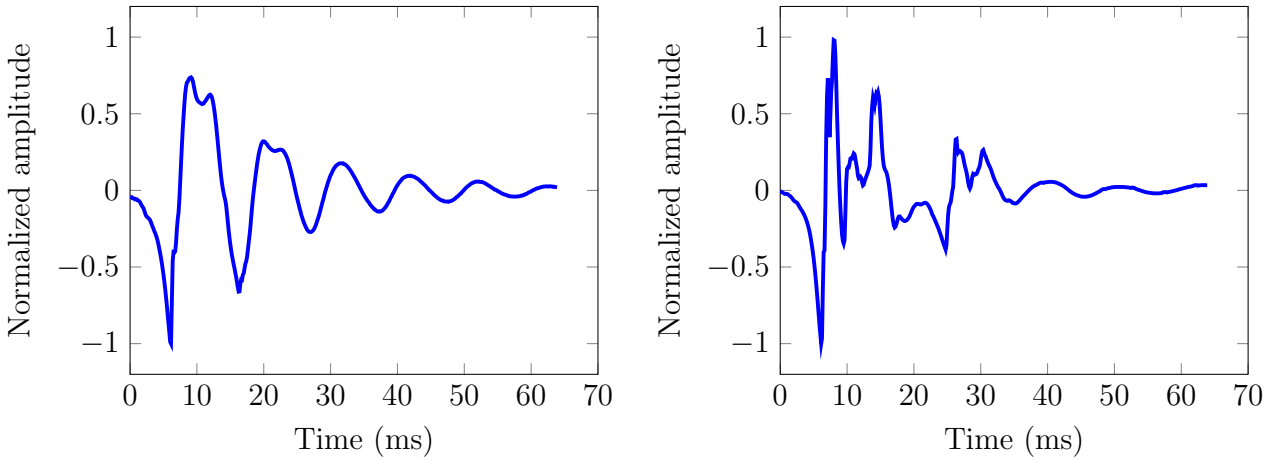
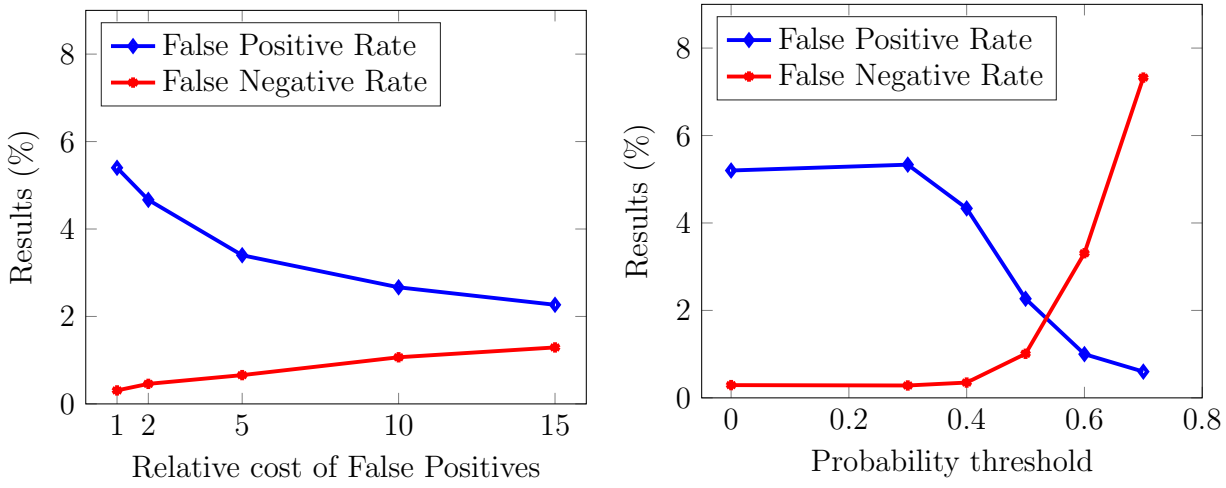


Figure 4: Top speaker signal of two different Table-Back-Top tap events when no case around the device

be reduced to 2.3% with an increased FNR of 1.3%, at a relative cost of 15. Another possibility is to use the posterior probability, which is defined as the probability of the classes, given the outcome of the classifier for a particular instance. Decisions are rejected (i.e., classify as no-taps) when the posterior probability of the class of the outcome falls below a specified threshold. Figure 5b illustrates that a FPR below 1% can be achieved at a threshold of 0.7. The downside is the removal of a significant portion of correctly classified taps, which results in a larger FNR of 7.3%. At a threshold of 0.5, the FPR equals 2.3 % with a FNR of 1.0%, which means that posterior probability rejection leads to better results than cost-sensitive classification.



(a) Cost-sensitive classification

(b) Posterior probability rejection

Figure 5: Techniques for reduction of the false positive rate

3.3 Optimal parameter selection

We can achieve 98.3% accuracy using all the sensors in this framework. The downside is the significant power consumption for sensing and processing. We found out that a

combination of the speakers and the accelerometer leads to 97.9% accuracy. This shows that the other sensors don't provide a lot of extra information. In terms of sensing, the speaker consumes no power and the accelerometer can be made very power efficient. The processing power is also reduced, since less features need to be computed.
To investigate the optimal set of features and the amount of trees in the Random Forest, the performance is tested for a variety of parameters. Figure 6 presents the results. The figure on the left shows the accuracy versus the amount of features for both evaluation methods. These features are selected using feature ranking using the information gain as evaluator. The biggest improvement is observed in the first 10 features, which are responsible for up to 91% accuracy. With 20 features, an accuracy of 95.5% is achieved.
The figure on the right illustrates the dependency of the accuracy on the amount of trees. Using a low amount of trees results in a low training error but a high generalization error (high risk of overfitting). By using more trees, the generalization error can be reduced. It is clear that using more than 30 trees is not needed, since it increases the complexity without a significant improvement in accuracy. For 10 decision trees, we get an accuracy of 97.3%. 30 decision trees can improve the accuracy to 98.0%.
From these experiments, we conclude that 20 features and 30 decision trees lead to excellent results with low complexity. For resource constrained-devices, we opt for a a configuration of 10 features and 10 decision trees to ensure acceptable results with minimal complexity. Using the combination of the speakers and the accelerometer in this configuration, an accuracy of 95.3% is achieved.

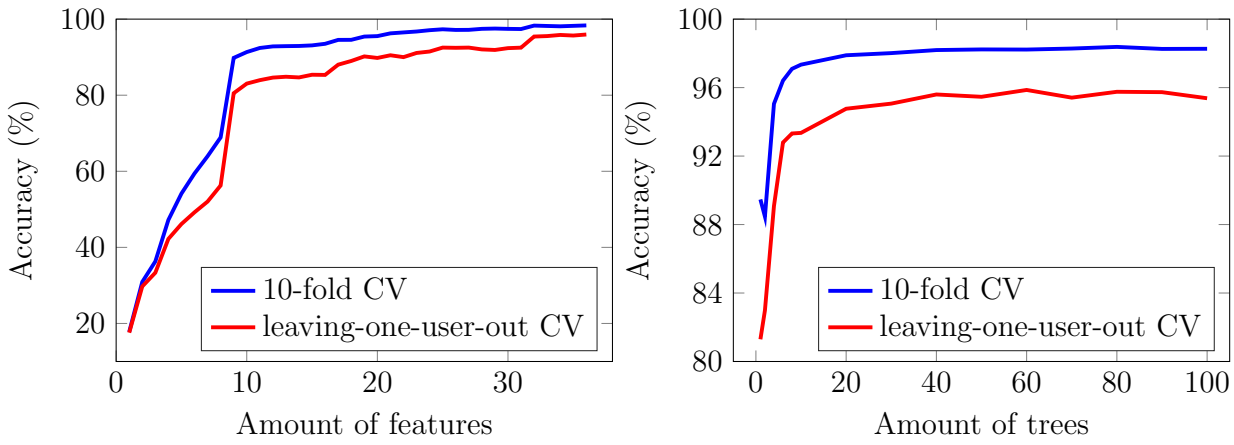


Figure 6: Accuracy versus amount of features and amount of decision trees

4 Conclusion

In this paper, a novel technique for finger tap detection is presented. *TouchSpeaker* exploits the speakers of the mobile device as primary sensors, because of low power consumption for sensing and high sensitivity for tap events. We have shown that a combination of the speakers with other built-in sensors can achieve 98.3% accuracy for 9 different finger tap events. If a restricted set of taps (hand, pocket) is envisioned, an accuracy of 99.8% can be obtained. To improve the robustness of *TouchSpeaker*, cost-sensitive classification and posterior probability rejection are proposed, leading to a false positive rate below 2.3% and 1% respectively, at the cost of a higher false negative rate. Finally, a low power solution is presented consisting of only the speakers and the accelerometer, resulting in 95.3% accuracy.

References

[1] H. Lu and Y. Li, "Gesture On: Enabling Always-On Touch Gestures for Fast Mobile Access from the Device Standby Mode," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, vol. 1, pp. 3355–3364, 2015.

[2] P. Daniel, "KnockOn app brings double-tap to lock to most Androids and wake to ones with OLED displays." [Online]. Available: http://www.phonearena.com/news/KnockOn-app-brings-double-tap-to-lock-to-most-Androids.-and-wake-to-ones-with-OLED-displays_id75033

[3] C. Harrison, J. Schwarz, and S. E. Hudson, "TapSense: enhancing finger interaction on touch surfaces," *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, pp. 627–636, 2011.

[4] W. McGrath and Y. Li, "Detecting tapping motion on the side of mobile devices by probabilistically combining hand postures," *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14*, pp. 215–219, 2014.

[5] C. Zhang, A. Guo, D. Zhang, C. Southern, R. Arriaga, and G. Abowd, "BeyondTouch: Extending the Input Language with Built-in Sensors on Commodity Smartphones," *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, pp. 67–77, 2015.

[6] D. Gouveia, "Creating a Basic Synth in XNA 4.0 Part I." [Online]. Available: http://www.david-gouveia.com/portfolio/creating-a-basic-synth-in-xna-part-i/

[7] M. Yacoub and G.-Z. Yang, *Body Sensor Networks*. Springer Science & Business Media, 2007.

[8] University of Waikato, "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[9] MathWorks Benelux, "MATLAB Profile execution time for functions." [Online]. Available: http://nl.mathworks.com/help/matlab/ref/profile.html

A Non-Convex Approach to Blind Calibration from Linear Sub-Gaussian Random Measurements

Valerio Cambareri and Laurent Jacques *

Image and Signal Processing Group (ISP Group), ICTEAM/ELEN
Université catholique de Louvain, Louvain-la-Neuve, Belgium.
E-mail: {valerio.cambareri, laurent.jacques}@uclouvain.be

Abstract

Blind calibration is a bilinear inverse problem arising in modern sensing strategies, whose solution becomes crucial when traditional calibration aided by multiple, accurately designed training signals is either infeasible or resource-consuming. We here address this problem for sensing schemes described by a linear model, in which the measured signal is projected on sub-Gaussian random sensing vectors each being affected by an unknown gain. By using multiple draws of the random sensing vectors we are able to solve this problem in its natural, non-convex form simply by projected gradient descent from a suitably chosen initialisation. Moreover, we obtain a sample complexity bound under which we are able to prove that this algorithm converges to the global optimum. Numerical evidence on the phase transition of this algorithm, as well as a practical computational sensing example support our theoretical findings.

1 Introduction

The problem of capturing an unknown input signal under sensing model uncertainties is crucial for modern sensing strategies such as Compressed Sensing (CS), in which these modelling errors inevitably occur in physical (*e.g.*, optical, analog) implementations and have direct impact on signal recovery [1]. However, if both the signal and model error remain fixed during the sensing process the use of random sensing operators in CS suggests that repeating the acquisition, *i.e.*, taking more *snapshots* under new draws of the sensing model could suffice to diversify the measurements and learn both unknown quantities. We here address the case of sensing a single unstructured vector $\boldsymbol{x} \in \mathbb{R}^n$ by collecting p independent snapshots of m independent random projections, *i.e.*,

$$\boldsymbol{y}_l = \bar{\boldsymbol{d}}\,\boldsymbol{A}_l\boldsymbol{x},\ \bar{\boldsymbol{d}} \coloneqq \mathrm{diag}(\boldsymbol{d}) \in \mathbb{R}^{m\times m},\ l \in [p] \coloneqq 1,\dots,p, \tag{1}$$

where $\boldsymbol{y}_l = (y_{1,l},\cdots,y_{m,l})^\top \in \mathbb{R}^m$ is the l-th snapshot; $\boldsymbol{d} = (d_1,\cdots,d_m)^\top \in \mathbb{R}^m_+$ is an unknown, positive and bounded gain vector that is identical throughout the p snapshots; the random matrices $\boldsymbol{A}_l \in \mathbb{R}^{m\times n}$ are independent and identically distributed (i.i.d.) and each $\boldsymbol{A}_l$ has i.i.d. rows, the i-th row $\boldsymbol{a}_{i,l}^\top \in \mathbb{R}^n$ being a centred isotropic sub-Gaussian random vector (*i.e.*, $\mathbb{E}[\boldsymbol{a}_{i,l}] = \boldsymbol{0}_n, \mathbb{E}[\boldsymbol{a}_{i,l}\boldsymbol{a}_{i,l}^\top] = \boldsymbol{I}_n$) such as i.i.d. Gaussian or Bernoulli random vectors (for a thorough introduction, see [2, Section 5.2]).

The bilinear inverse problem of jointly recovering $(\boldsymbol{x},\boldsymbol{d})$ is referred to as *blind calibration* [3, 4] and is especially motivated, but not limited to compressive imaging applications where unknown $\boldsymbol{d}$ are associated to positive gains and attenuations in a

*The authors are funded by the Belgian F.R.S.-FNRS. Part of this study is funded by the project AlterSense (MIS-FNRS).

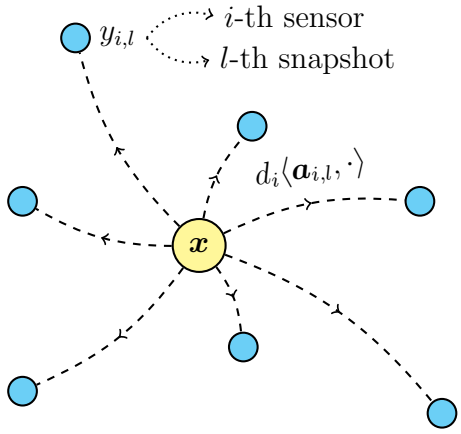


Figure 1: A general application scheme for (1).

focal plane array, while the random projections of an image $\boldsymbol{x}$ are suitably obtained by spatial light modulation injecting the l-th sensing operator $\boldsymbol{A}_l$ (*e.g.*, by random convolution with a programmable modulation medium [5–9], albeit losing sub-Gaussianity). Similar model errors concern array signal processing applications as discussed in [10,11]. More generally, the model we will describe can be adapted to any sensing scheme that fits the configuration depicted in Figure 1, *i.e.*, whenever the measurements are obtained as the projection of an unknown input $\boldsymbol{x}$ (*e.g.*, an image, time-series or channel response) on some sub-Gaussian random sensing vectors $\boldsymbol{a}_{i,l}$ known and injected during the sensing process, so that the mp outputs of this operation impinge on m sensor elements subject to some unknown gain coefficients d_i.

In such a context, when inverse problems are formulated to retrieve $\boldsymbol{x}$ the knowledge of $\boldsymbol{d}$ could critically improve the accuracy of the recovered signal; gaining its knowledge from (1) is typically achieved by the solution of convex or alternating minimisation problems [3, 4, 12] aided by the use of multiple input signals (*i.e.*, $\boldsymbol{x}_l, l \in [p]$) required to be either as independent as possible or to lie in a low-dimensional subspace, instead of randomising the sensing operator itself. More recently, *lifting* approaches [10,11] have been proposed to jointly recover $(\boldsymbol{x}, \boldsymbol{d})$ in (1), as well as more general *blind deconvolution* models [13, 14]. Their main limitation is in that a semidefinite program is solved to recover a very high-dimensional rank-one matrix $\boldsymbol{x}\boldsymbol{d}^\top$, an approach that becomes quite rapidly computationally unaffordable as m and n exceed a few hundreds.

2 Non-Convex Blind Calibration

The solution we propose is inspired by recent results on fast and provably convergent non-convex approaches to the quadratic inverse problem of phase retrieval [15–17]. We here propose to solve the blind calibration problem of recovering two unstructured vectors $(\boldsymbol{x}, \boldsymbol{d})$ in (1) by a non-convex formulation described as follows; since no *a priori* structure is given on $(\boldsymbol{x}, \boldsymbol{d})$ we must operate in an oversampling regime with respect to (w.r.t.) $mp \geq n + m$ and solve

$$(\hat{\boldsymbol{x}}, \hat{\boldsymbol{d}}) = \underset{\boldsymbol{\xi}\in\mathbb{R}^n, \boldsymbol{\gamma}\in\Pi_+^m}{\operatorname{argmin}} f(\boldsymbol{\xi}, \boldsymbol{\gamma}), \;\; f(\boldsymbol{\xi}, \boldsymbol{\gamma}) \coloneqq \tfrac{1}{2mp}\textstyle\sum_{l=1}^p \|\bar{\boldsymbol{\gamma}}\boldsymbol{A}_l\boldsymbol{\xi} - \boldsymbol{y}_l\|_2^2, \tag{2}$$

Quantity	**Finite-sample** ($p < \infty$)	**Expectation** ($\mathbb{E}_{\boldsymbol{a}_{i,l}}, p \to \infty$)
Objective: $f(\boldsymbol{\xi}, \boldsymbol{\gamma})$	$\frac{1}{2mp} \sum_{l=1}^{p} \left\| \bar{\boldsymbol{\gamma}} \boldsymbol{A}_l \boldsymbol{\xi} - \bar{\boldsymbol{d}} \boldsymbol{A}_l \boldsymbol{x} \right\|_2^2$	$\frac{1}{2m} \| \boldsymbol{\xi}\boldsymbol{\gamma}^\top - \boldsymbol{x}\boldsymbol{d}^\top \|_F^2$
Signal gradient: $\boldsymbol{\nabla}_{\boldsymbol{\xi}} f(\boldsymbol{\xi}, \boldsymbol{\gamma})$	$\frac{1}{mp} \sum_{l=1}^{p} \boldsymbol{A}_l^\top \bar{\boldsymbol{\gamma}} \left(\bar{\boldsymbol{\gamma}} \boldsymbol{A}_l \boldsymbol{\xi} - \bar{\boldsymbol{d}} \boldsymbol{A}_l \boldsymbol{x} \right)$	$\frac{1}{m} \left[\| \boldsymbol{\gamma} \|_2^2 \boldsymbol{\xi} - (\boldsymbol{\gamma}^\top \boldsymbol{d}) \boldsymbol{x} \right]$
Gain gradient: $\boldsymbol{\nabla}_{\boldsymbol{\gamma}} f(\boldsymbol{\xi}, \boldsymbol{\gamma})$	$\frac{1}{mp} \sum_{l=1}^{p} \overline{\boldsymbol{A}_l \boldsymbol{\xi}} \left(\bar{\boldsymbol{\gamma}} \boldsymbol{A}_l \boldsymbol{\xi} - \bar{\boldsymbol{d}} \boldsymbol{A}_l \boldsymbol{x} \right)$	$\frac{1}{m} \left[\| \boldsymbol{\xi} \|_2^2 \boldsymbol{\gamma} - (\boldsymbol{\xi}^\top \boldsymbol{x}) \boldsymbol{d} \right]$
Projected gain gradient: $\boldsymbol{\nabla}_{\boldsymbol{\gamma}}^{\perp} f(\boldsymbol{\xi}, \boldsymbol{\gamma})$	$\frac{1}{mp} \sum_{l=1}^{p} \overline{\boldsymbol{A}_l \boldsymbol{\xi}} \left(\bar{\boldsymbol{\varepsilon}} \boldsymbol{A}_l \boldsymbol{\xi} - \bar{\boldsymbol{\omega}} \boldsymbol{A}_l \boldsymbol{x} \right)$	$\frac{1}{m} \left[\| \boldsymbol{\xi} \|_2^2 \boldsymbol{\varepsilon} - (\boldsymbol{\xi}^\top \boldsymbol{x}) \boldsymbol{\omega} \right]$
Hessian matrix: $\mathcal{H} f(\boldsymbol{\xi}, \boldsymbol{\gamma})$	$\frac{1}{mp} \sum_{l=1}^{p} \begin{bmatrix} \boldsymbol{A}_l^\top \bar{\boldsymbol{\gamma}}^2 \boldsymbol{A}_l & \boldsymbol{A}_l^\top \overline{2\boldsymbol{\gamma} \boldsymbol{A}_l \boldsymbol{\xi} - \boldsymbol{d} \boldsymbol{A}_l \boldsymbol{x}} \\ \overline{2\boldsymbol{\gamma} \boldsymbol{A}_l \boldsymbol{\xi} - \boldsymbol{d} \boldsymbol{A}_l \boldsymbol{x}} \, \boldsymbol{A}_l & \overline{\boldsymbol{A}_l \boldsymbol{\xi}}^2 \end{bmatrix}$	$\frac{1}{m} \begin{bmatrix} \| \boldsymbol{\gamma} \|_2^2 \boldsymbol{I}_n & 2\boldsymbol{\xi}\boldsymbol{\gamma}^\top - \boldsymbol{x}\boldsymbol{d}^\top \\ 2\boldsymbol{\gamma}\boldsymbol{\xi}^\top - \boldsymbol{d}\boldsymbol{x}^\top & \| \boldsymbol{\xi} \|_2^2 \boldsymbol{I}_m \end{bmatrix}$
Initialisation: $(\boldsymbol{\xi}_0, \boldsymbol{\gamma}_0)$	$\left(\frac{1}{mp} \sum_{l=1}^{p} (\boldsymbol{A}_l)^\top \bar{\boldsymbol{d}} \boldsymbol{A}_l \boldsymbol{x}, \, \boldsymbol{1}_m \right)$	$\left(\frac{\| \boldsymbol{d} \|_1}{m} \boldsymbol{x}, \, \boldsymbol{1}_m \right)$

Table 1: Finite-sample and expected values of the objective function; its gradient components and Hessian matrix; the initialisation point $(\boldsymbol{\xi}_0, \boldsymbol{\gamma}_0)$.

given $\{\boldsymbol{y}_l\}_{l=1}^p, \{\boldsymbol{A}_l\}_{l=1}^p$ with $\Pi_+^m \coloneqq \{\boldsymbol{\gamma} \in \mathbb{R}_+^m, \boldsymbol{1}_m^\top \boldsymbol{\gamma} = m\}$ being the scaled probability simplex. The geometry of this problem can be studied by observing the objective $f(\boldsymbol{\xi}, \boldsymbol{\gamma})$ with its gradient $\boldsymbol{\nabla} f(\boldsymbol{\xi}, \boldsymbol{\gamma}) = \left[\, (\boldsymbol{\nabla}_{\boldsymbol{\xi}} f(\boldsymbol{\xi}, \boldsymbol{\gamma}))^\top \; (\boldsymbol{\nabla}_{\boldsymbol{\gamma}} f(\boldsymbol{\xi}, \boldsymbol{\gamma}))^\top \, \right]^\top$ and Hessian matrix $\mathcal{H} f(\boldsymbol{\xi}, \boldsymbol{\gamma})$, all computed in Table 1 for $p < \infty$. The finite-sample expressions therein are unbiased estimates of their expectation w.r.t. the i.i.d. sensing vectors $\boldsymbol{a}_{i,l}$. There, we evince that the set $\left\{ (\boldsymbol{\xi}, \boldsymbol{\gamma}) \in \mathbb{R}^n \times \mathbb{R}^m : \boldsymbol{\xi} = \frac{1}{\alpha} \boldsymbol{x}, \boldsymbol{\gamma} = \alpha \boldsymbol{d}, \alpha \in \mathbb{R} \setminus \{0\} \right\}$ contains the global minimisers of $f(\boldsymbol{\xi}, \boldsymbol{\gamma})$. Moreover $f(\boldsymbol{\xi}, \boldsymbol{\gamma})$ is shown to be generally non-convex[†] as there exist plenty of counterexamples $(\boldsymbol{\xi}', \boldsymbol{\gamma}')$ for which $\mathcal{H} f(\boldsymbol{\xi}', \boldsymbol{\gamma}') \not\succeq 0$.

We now elaborate on the constraint $\boldsymbol{\gamma} \in \Pi_+^m$. Firstly, we see that only one minimiser $(\boldsymbol{x}^\star, \boldsymbol{d}^\star) \coloneqq \left(\frac{\| \boldsymbol{d} \|_1}{m} \boldsymbol{x}, \frac{m}{\| \boldsymbol{d} \|_1} \boldsymbol{d} \right)$ remains for $mp \geq n+m$, which amounts to the exact solution up to a fixed $\alpha = {}^{m}/_{\| \boldsymbol{d} \|_1}$. Secondly, assuming that $\boldsymbol{d} \in \mathbb{R}_+^m$ in (1) is positive and bounded amounts to letting $\boldsymbol{d}^\star \in \mathcal{C}_\rho \subset \Pi_+^m$, $\mathcal{C}_\rho \coloneqq \boldsymbol{1}_m + \boldsymbol{1}_m^\perp \cap \rho \, \mathbb{B}_\infty^m$ for a maximum deviation $\rho > \| \boldsymbol{d}^\star - \boldsymbol{1}_m \|_\infty, \rho < 1$ (where we denoted the orthogonal complement $\boldsymbol{1}_m^\perp \coloneqq \{\boldsymbol{v} \in \mathbb{R}^m : \boldsymbol{1}_m^\top \boldsymbol{v} = 0\}$ and $\mathbb{B}_p^q$ the ℓ_p-ball in $\mathbb{R}^q$). In particular, we can specify $\boldsymbol{d}^\star = \boldsymbol{1}_m + \boldsymbol{\omega}$ for $\boldsymbol{\omega} \in \boldsymbol{1}_m^\perp \cap \rho \, \mathbb{B}_\infty^m$, as well as $\boldsymbol{\gamma} = \boldsymbol{1}_m + \boldsymbol{\varepsilon}$ for $\boldsymbol{\varepsilon} \in \boldsymbol{1}_m^\perp \cap \rho \, \mathbb{B}_\infty^m$ provided that the algorithm solving (2) is so that $\boldsymbol{\gamma}$ remains in $\mathcal{C}_\rho$. This allows us to recast (2) in terms of the variations $\boldsymbol{\omega}, \boldsymbol{\varepsilon} \in \boldsymbol{1}_m^\perp \cap \rho \mathbb{B}_\infty^m$ on the simplex Π_+^m.

While applying this constraint to the minimisation of $f(\boldsymbol{\xi}, \boldsymbol{\gamma})$ will not grant convexity in the domain of (2), we proceed by defining a *neighbourhood* of the global minimiser $(\boldsymbol{x}^\star, \boldsymbol{d}^\star)$ as follows. To begin with, we will require a notion of distance, *i.e.*,

$$\Delta(\boldsymbol{\xi}, \boldsymbol{\gamma}) \coloneqq \| \boldsymbol{\xi} - \boldsymbol{x}^\star \|_2^2 + \tfrac{\| \boldsymbol{x}^\star \|_2^2}{m} \| \boldsymbol{\gamma} - \boldsymbol{d}^\star \|_2^2.$$

Then, to account for the constraint in the gain domain we define our neighbourhood

$$\mathcal{D}_{\kappa,\rho} \coloneqq \{ (\boldsymbol{\xi}, \boldsymbol{\gamma}) \in \mathbb{R}^n \times \mathcal{C}_\rho : \Delta(\boldsymbol{\xi}, \boldsymbol{\gamma}) \leq \kappa^2 \| \boldsymbol{x}^\star \|_2^2 \}, \;\; \rho \in [0, 1).$$

Thus, $\mathcal{D}_{\kappa,\rho}$ is the intersection of an ellipsoid in $\mathbb{R}^n \times \mathbb{R}^m$, as defined by $\Delta(\boldsymbol{\xi}, \boldsymbol{\gamma}) \leq \kappa^2 \| \boldsymbol{x}^\star \|_2^2$, with $\mathbb{R}^n \times \mathcal{C}_\rho$ (and as such is a convex set). While we anticipate that local

[†]It is *biconvex* [11], *i.e.*, convex once either $\boldsymbol{\xi}$ or $\boldsymbol{\gamma}$ are fixed.

1: Initialise $\boldsymbol{\xi}_0 \coloneqq \frac{1}{mp}\sum_{l=1}^{p}(\boldsymbol{A}_l)^\top \boldsymbol{y}_l$, $\boldsymbol{\gamma}_0 \coloneqq \mathbf{1}_m$, $k \coloneqq 0$.
2: **while** stop criteria not met **do**
3: $\begin{cases}\mu_{\boldsymbol{\xi}} \coloneqq \operatorname{argmin}_{\upsilon\in\mathbb{R}} f(\boldsymbol{\xi}_k - \upsilon\boldsymbol{\nabla}_{\boldsymbol{\xi}} f(\boldsymbol{\xi}_k,\boldsymbol{\gamma}_k),\boldsymbol{\gamma}_k)\\ \mu_{\boldsymbol{\gamma}} \coloneqq \operatorname{argmin}_{\upsilon\in\mathbb{R}} f(\boldsymbol{\xi}_k,\boldsymbol{\gamma}_k - \upsilon\boldsymbol{\nabla}^{\perp}_{\boldsymbol{\gamma}} f(\boldsymbol{\xi}_k,\boldsymbol{\gamma}_k))\end{cases}$ {Line search in $\boldsymbol{\xi},\boldsymbol{\gamma}$}
4: $\boldsymbol{\xi}_{k+1} \coloneqq \boldsymbol{\xi}_k - \mu_{\boldsymbol{\xi}}\boldsymbol{\nabla}_{\boldsymbol{\xi}} f(\boldsymbol{\xi}_k,\boldsymbol{\gamma}_k)$
5: $\underline{\boldsymbol{\gamma}}_{k+1} \coloneqq \boldsymbol{\gamma}_k - \mu_{\boldsymbol{\gamma}}\,\boldsymbol{\nabla}^{\perp}_{\boldsymbol{\gamma}} f(\boldsymbol{\xi}_k,\boldsymbol{\gamma}_k)$
6: $\boldsymbol{\gamma}_{k+1} \coloneqq P_{\mathcal{C}_\rho}\underline{\boldsymbol{\gamma}}_{k+1}$
7: $k \coloneqq k+1$
8: **end while**

Algorithm 1: Non-Convex Blind Calibration by Projected Gradient Descent.

convexity can be shown on $\mathcal{D}_{\kappa,\rho}$ by testing the projected Hessian matrix against the direction $(\boldsymbol{\xi}-\boldsymbol{x}^\star,\boldsymbol{\gamma}-\boldsymbol{d}^\star)$, in a fashion similar to [16, Theorem 2.3], this directional local convexity argument is not explicitly used here (it will be reported in an upcoming journal paper [18]); a first-order analysis will actually suffice to prove our main results.

In summary, to solve (2) we will use a two-fold procedure which starts from a carefully chosen initialisation and is followed by a gradient descent algorithm. In more detail, we require an *initialisation* point $(\boldsymbol{\xi}_0,\boldsymbol{\gamma}_0)$ that lands in a small neighbourhood around $(\boldsymbol{x}^\star,\boldsymbol{d}^\star)$. Similarly to [15] we see that, at least in the signal domain, initialising $\boldsymbol{\xi}_0$ as in Table 1 grants $\mathbb{E}[\boldsymbol{\xi}_0] \equiv \boldsymbol{x}^\star$, *i.e.*, asymptotically in p this initialisation is an unbiased estimator of the exact solution we seek. As for the gain domain, for simplicity we let $\boldsymbol{\gamma}_0 \coloneqq \mathbf{1}_m$ (*i.e.*, $\boldsymbol{\varepsilon}_0 \coloneqq \mathbf{0}_m$). In Proposition 1 we will see that when mp meets a linear requirement in $n+m$, we can have $(\boldsymbol{\xi}_0,\boldsymbol{\gamma}_0)\in\mathcal{D}_{\kappa,\rho}$ for some small κ.

After this initialisation we run a projected gradient descent to solve (2), as summarised in Algorithm 1. A few simplifications related to (2) are in order: while generally we would need to ensure $\boldsymbol{\gamma}_k \in \Pi_+^m$, since the optimisation starts on $\boldsymbol{\gamma}_0 = \mathbf{1}_m$ we first project the gradient update in the gain domain as $\boldsymbol{\nabla}^{\perp}_{\boldsymbol{\gamma}} f(\boldsymbol{\xi},\boldsymbol{\gamma}) \coloneqq \boldsymbol{P}_{\mathbf{1}_m^\perp}\boldsymbol{\nabla}_{\boldsymbol{\gamma}} f(\boldsymbol{\xi},\boldsymbol{\gamma})$ (step 5:) with the projection matrix $\boldsymbol{P}_{\mathbf{1}_m^\perp} \coloneqq \boldsymbol{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top$ (see Table 1 for its expression in terms of $\boldsymbol{\varepsilon},\boldsymbol{\omega}$). Then we apply $P_{\mathcal{C}_\rho}$, *i.e.*, the projector on $\mathcal{C}_\rho$ ensuring that each $\boldsymbol{\gamma}_k \in \mathcal{C}_\rho$ (step 6:). Actually, this step is a mere theoretical requirement to prove the convergence of Algorithm 1 to $(\boldsymbol{x}^\star,\boldsymbol{d}^\star)$; we have observed that, at least numerically, the projected gradient update in step 5: alone suffices since $\boldsymbol{\gamma}_k \in \mathcal{C}_\rho$ is always verified in our experiments.

As for the line searches in step 3: of Algorithm 1 they can be solved in closed-form, and are merely introduced to improve the convergence rate w.r.t. a fixed value of $\mu_{\boldsymbol{\xi}},\mu_{\boldsymbol{\gamma}}$. Below, we obtain the conditions that ensure convergence of this descent algorithm to $(\boldsymbol{x}^\star,\boldsymbol{d}^\star)$ for fixed step sizes $\mu_{\boldsymbol{\xi}},\mu_{\boldsymbol{\gamma}}$.

3 Recovery Guarantees

The statements presented below rely on a concentration inequality proved in [18] and reported in Lemma 1, as well as some simple geometry of (2) in $\mathcal{D}_{\kappa,\rho}$ (*i.e.*, simple bounds on vector and matrix norms in this neighbourhood). Having assumed that all mp sensing vectors $\boldsymbol{a}_{i,l}$ are i.i.d. sub-Gaussian, we report the concentration inequality which allows us to control for $p<\infty$ the matrix norm of a weighted sum of $\boldsymbol{a}_{i,l}\boldsymbol{a}_{i,l}^\top$.

Lemma 1 (Weighted Concentration Inequality). *Let* $\{\boldsymbol{a}_{i,l} \in \mathbb{R}^n : i \in [m], l \in [p]\}$ *be a set of random vectors, each formed by* n *i.i.d. copies of a sub-Gaussian random variable* X *[2, Section 5.2.3] with* $\mathbb{E}\,X = 0$, $\mathbb{E}\,X^2 = 1$ *and sub-Gaussian norm* $\|X\|_{\psi_2}$. *Given* $\delta \in (0,1)$ *and* $t > 1$ *we have, with probability exceeding*

$$1 - Ce^{-c\delta^2 mp} - (mp)^{-t} \tag{3}$$

for some $C, c > 0$ *depending only on* $\|X\|_{\psi_2}$, *that*

$$\left\|\tfrac{1}{mp}\sum_{i=1}^m \sum_{l=1}^p \theta_i(\boldsymbol{a}_{i,l}\boldsymbol{a}_{i,l}^\top - \boldsymbol{I}_n)\right\|_2 \leq \delta\,\|\boldsymbol{\theta}\|_\infty,\ \forall \boldsymbol{\theta} = \{\theta_i\}_{i=1}^m \in \mathbb{R}^m \tag{4}$$

provided $n \gtrsim t\log(mp)$ *and* $mp \gtrsim \delta^{-2}(n+m)\log(\frac{n}{\delta})$.

We now establish the sample complexity required for the initialisation to lie in $\mathcal{D}_{\kappa,\rho}$ with very high probability for some κ and ρ fixed by the entity of the model error.

Proposition 1 (Initialisation Proximity). *Let* $(\boldsymbol{\xi}_0, \boldsymbol{\gamma}_0)$ *be as in Table 1. For any* $\epsilon \in (0,1)$ *we have, with probability exceeding* $1 - Ce^{-c\epsilon^2 mp} - (mp)^{-t}$ *for some* $C, c > 0$, *that* $\|\boldsymbol{\xi}_0 - \boldsymbol{x}^\star\|_2 \leq \epsilon\|\boldsymbol{x}^\star\|_2$ *provided* $n \gtrsim t\log(mp)$ *and* $mp \gtrsim \epsilon^{-2}(n+m)\log\left(n\epsilon^{-1}\right)$. *Since* $\boldsymbol{\gamma}_0 = \mathbf{1}_m$ *we also have* $\|\boldsymbol{\gamma}_0 - \boldsymbol{d}^\star\|_\infty \leq \rho < 1$. *Thus* $(\boldsymbol{\xi}_0, \boldsymbol{\gamma}_0) \in \mathcal{D}_{\kappa,\rho}$ *with the same probability and* $\kappa := \sqrt{\epsilon^2 + \rho^2} \leq \sqrt{2}$.

With analogue tools (and more general concentration properties of isotropic sub-Gaussian $\boldsymbol{a}_{i,l}$ [2, Section 5.2]) we are able to state when a single gradient descent step from any $(\boldsymbol{\xi}, \boldsymbol{\gamma}) \in \mathcal{D}_{\kappa,\rho}$ reduces, with very high probability, the distance w.r.t. the global minimum. This requires establishing a *regularity condition* on the projected gradient $\boldsymbol{\nabla}^\perp f(\boldsymbol{\xi}, \boldsymbol{\gamma}) := \left[\,(\boldsymbol{\nabla}_{\boldsymbol{\xi}} f(\boldsymbol{\xi},\boldsymbol{\gamma}))^\top\ (\boldsymbol{\nabla}_{\boldsymbol{\gamma}}^\perp f(\boldsymbol{\xi},\boldsymbol{\gamma}))^\top\,\right]^\top$ holding *uniformly*, *i.e.*, for all $(\boldsymbol{\xi}, \boldsymbol{\gamma}) \in \mathcal{D}_{\kappa,\rho}$ (similarly to [15, Condition 7.9]).

Proposition 2 (Regularity condition in $\mathcal{D}_{\kappa,\rho}$). *For any* $\delta \in (0,1)$ *there exist a constant* $\eta \in (0,1)$, $t > 1$ *and a value* $L < 3(1+\kappa)(4+\|\boldsymbol{x}^\star\|_2)$ *such that, with probability exceeding*

$$1 - C\left[me^{-c\delta^2 p} + e^{-c\delta^2 mp} + (mp)^{-t}\right]$$

for some $C, c > 0$, *we have for all* $(\boldsymbol{\xi}, \boldsymbol{\gamma}) \in \mathcal{D}_{\kappa,\rho}$, $\rho \in [0,1)$,

$$\left\langle \boldsymbol{\nabla}^\perp f(\boldsymbol{\xi}, \boldsymbol{\gamma}), \begin{bmatrix} \boldsymbol{\xi} - \boldsymbol{x}^\star \\ \boldsymbol{\gamma} - \boldsymbol{d}^\star \end{bmatrix} \right\rangle \geq \tfrac{1}{2}\eta\,\Delta(\boldsymbol{\xi}, \boldsymbol{\gamma}) \qquad \text{(Bounded curvature)}$$

$$\|\boldsymbol{\nabla}^\perp f(\boldsymbol{\xi}, \boldsymbol{\gamma})\|_2^2 \leq L^2\,\Delta(\boldsymbol{\xi}, \boldsymbol{\gamma}) \qquad \text{(Lipschitz gradient)}$$

provided $\rho < \frac{1}{14}(1 - \eta - 3\delta)$, $n \gtrsim t\log(mp)$, $p \gtrsim \delta^{-2}\log m$ *and* $mp \gtrsim \delta^{-2}(n+m)\log(\frac{n}{\delta})$.

We now use Proposition 2 to show that the neighbourhood $\mathcal{D}_{\kappa,\rho}$ is a basin of attraction to the solution $(\boldsymbol{x}^\star, \boldsymbol{d}^\star)$. If we update $\boldsymbol{\xi}_+ := \boldsymbol{\xi} - \mu_{\boldsymbol{\xi}}\boldsymbol{\nabla}_{\boldsymbol{\xi}} f(\boldsymbol{\xi}, \boldsymbol{\gamma}), \underline{\boldsymbol{\gamma}}_+ := \boldsymbol{\gamma} - \mu_{\boldsymbol{\gamma}}\boldsymbol{\nabla}_{\boldsymbol{\gamma}}^\perp f(\boldsymbol{\xi}, \boldsymbol{\gamma}), \boldsymbol{\gamma}_+ := P_{\mathcal{C}_\rho}\underline{\boldsymbol{\gamma}}_+$ from any $(\boldsymbol{\xi}, \boldsymbol{\gamma}) \in \mathcal{D}_{\kappa,\rho}$ we have that, when $\mu_{\boldsymbol{\xi}} := \mu, \mu_{\boldsymbol{\gamma}} \propto m\mu$ and μ is sufficiently small, (i) $\Delta(\boldsymbol{\xi}_+, \boldsymbol{\gamma}_+) < \Delta(\boldsymbol{\xi}, \boldsymbol{\gamma})$ and (ii) $(\boldsymbol{\xi}, \boldsymbol{\gamma}) \in \mathcal{D}_{\kappa,\rho}$. Note that the role of $P_{\mathcal{C}_\rho}$ (*i.e.*, a convex ℓ_∞-ball, that is a contraction) is technical in ensuring $(i), (ii)$ above by verifying $\|\boldsymbol{\gamma}_+ - \boldsymbol{d}^\star\|_2 \leq \|\underline{\boldsymbol{\gamma}}_+ - \boldsymbol{d}^\star\|_2$. By a recursive application of these facts (which will be expanded in [18]) we obtain the following convergence result.

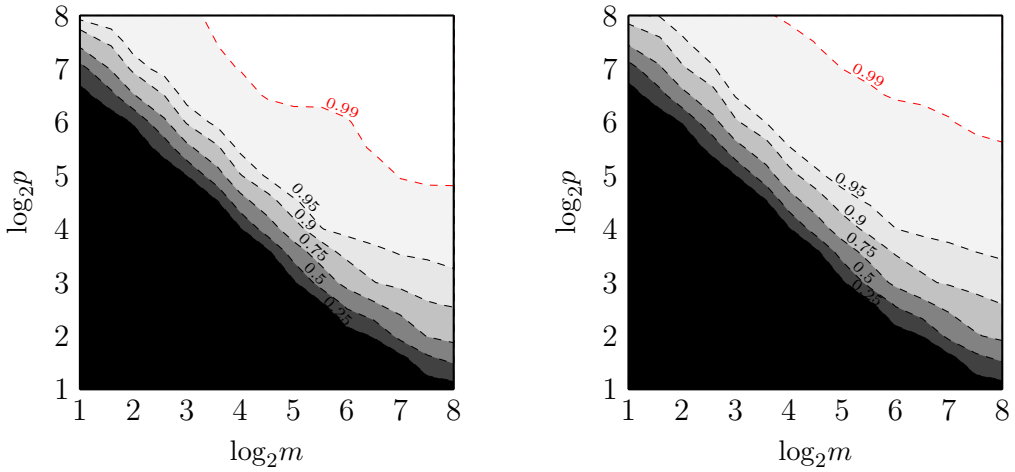


Figure 2: Empirical phase transition of (2) for $n = 2^8$, $\rho = 10^{-3}$ (left) and $\rho = 10^{-0.5}$ (right). The probability value P_ζ is reported on each contour line.

Theorem 1 (Provable Convergence to the Exact Solution)**.** *Under the conditions of Proposition 1, 2 we have that, with probability exceeding* $1 - C\big[me^{-c\delta^2 p} + e^{-c\delta^2 mp} + e^{-ce^2 mp} + (mp)^{-t}\big]$ *for some* $C, c > 0$*, Algorithm 1 with* $\mu_{\boldsymbol{\xi}} \coloneqq \mu, \mu_{\boldsymbol{\gamma}} \coloneqq \mu \frac{m}{\|\boldsymbol{x}^\star\|_2^2}$ *has error decay*

$$\Delta(\boldsymbol{\xi}_k, \boldsymbol{\gamma}_k) \leq \big(1 - \eta\mu + \tfrac{L^2}{\tau}\mu^2\big)^k \big(\epsilon^2 + \rho^2\big)\|\boldsymbol{x}^\star\|_2^2, (\boldsymbol{\xi}_k, \boldsymbol{\gamma}_k) \in \mathcal{D}_{\kappa,\rho} \tag{5}$$

at any iteration $k > 0$ *provided* $\mu \in \big(0, \tau\eta/L^2\big)$*,* $\tau \coloneqq \min\{1, \|\boldsymbol{x}^\star\|_2^2/m\}$*. Hence,*

$$\Delta(\boldsymbol{\xi}_k, \boldsymbol{\gamma}_k) \underset{k\to\infty}{\longrightarrow} 0.$$

We remark that the initialisation is critical to set the value of κ for the verification of Proposition 1,2, with its initial value appearing in (5). Let us mention finally that if additive measurement noise $\boldsymbol{N} \coloneqq (\boldsymbol{n}_1, \cdots, \boldsymbol{n}_p) \in \mathbb{R}^{m\times p}$ corrupts the p snapshots of the sensing model (1), then it can be shown that $\Delta(\boldsymbol{\xi}_k, \boldsymbol{\gamma}_k) \underset{k\to\infty}{\longrightarrow} \sigma^2$ with $\frac{1}{mp}\|\boldsymbol{N}\|_F^2 \lesssim \sigma^2$. Thus, Algorithm 1 is also robust to noise, its solution degrading gracefully with σ^2. This stability result will be proved in [18].

4 Numerical Experiments

4.1 Empirical Phase Transition

To characterise the *phase transition* of (2), that is the transition between a region in which Algorithm 1 successfully recovers $(\boldsymbol{x}^\star, \boldsymbol{d}^\star)$ with probability 1 and that in which it does with probability 0, we ran some extensive simulations by generating 256 random instances of (2) for each $n = \{2^1, \ldots, 2^8\}$ and taking the latter range for m, p. Then, we let $\rho = \{10^{-3}, 10^{-2.5}, \ldots, 1\}$, drawing $\boldsymbol{d}^\star = \mathbf{1}_m + \boldsymbol{\omega}$ with $\boldsymbol{\omega}$ drawn uniformly at random on $\mathbf{1}_m^\perp \cap \rho\,\mathbb{S}_\infty^{m-1}$. Then we evaluated $\mathrm{P}_\zeta \coloneqq \mathbb{P}\left[\max\left\{\frac{\|\hat{\boldsymbol{d}}-\boldsymbol{d}^\star\|_2}{\|\boldsymbol{d}^\star\|_2}, \frac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2}{\|\boldsymbol{x}^\star\|_2}\right\} < \zeta\right]$ on the trials with $\zeta = 10^{-3}$ chosen according to the stop criterion $f(\boldsymbol{\xi}, \boldsymbol{\gamma}) < 10^{-7}$. Of this large dataset we only report the case $n = 2^8$ in Figure 2, highlighting the contour lines of P_ζ for $\rho = \{10^{-3}, 10^{-0.5}\}$ as a function of $\log_2 m$ and $\log_2 p$. There, we do observe a phase transition at $\log_2 m + \log_2 p \simeq \log n$. Moreover, we see how an increase in ρ only slightly affects the requirements on (m, p) for a successful recovery of $(\boldsymbol{x}^\star, \boldsymbol{d}^\star)$.

(a) $\boldsymbol{x}$ *(b)* $\boldsymbol{d}, \rho = 1/2$ *(c)* $\hat{\boldsymbol{x}}$ by LS *(d)* $\hat{\boldsymbol{x}}$ by (2) *(e)* $\hat{\boldsymbol{d}}$ by (2)

Figure 3: A high-dimensional example of blind calibration for the linear sensing model (1) under unstructured gains $\boldsymbol{d}$ with $m = n, p = 4$ snapshots.

4.2 Blind Calibration of an Imaging System

To test (2) in a realistic context, we assume that $\boldsymbol{x}$ is a $n = 64 \times 64$ pixel image acquired by an imaging system following (1), in which its $m = 64 \times 64$ pixel sensor array suffers from large *pixel response non-uniformity* [19]. This is simulated by generating $\boldsymbol{d}$ as before with $\rho = 1/2$. We capture $p = 4$ snapshots with i.i.d. draws of $\boldsymbol{a}_{i,l} \sim \mathcal{N}(\boldsymbol{0}_n, \boldsymbol{I}_n)$. By running Algorithm 1, the recovered estimates $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{d}})$ attain $\max\left\{\frac{\|\hat{\boldsymbol{d}}-\boldsymbol{d}^\star\|_2}{\|\boldsymbol{d}^\star\|_2}, \frac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2}{\|\boldsymbol{x}^\star\|_2}\right\} \approx -147.38\,\text{dB}$. Instead, by fixing $\boldsymbol{\gamma} \coloneqq \boldsymbol{1}_m$ and solving (2) only in $\boldsymbol{\xi}$, *i.e.*, finding the least-squares (LS) solution $\hat{\boldsymbol{x}}$ in the presence of an unknown model error, we obtain $\frac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2}{\|\boldsymbol{x}^\star\|_2} \approx -5.50\,\text{dB}$. This confirms the practicality of our method for high-dimensional problems.

5 Conclusion

We presented and solved a non-convex formulation of the blind calibration problem for a linear model comprised of sub-Gaussian random sensing vectors. In absence of *a priori* information on the solution $(\boldsymbol{x}, \boldsymbol{d})$, the algorithm successfully recovers both unknown vectors under a sample complexity requirement that grows at a linear rate $mp \gtrsim (n+m)\log(n)$ in the signal and gain dimensions. Future developments include the achievement of a blind calibration method for CS by exploiting a low-dimensional model for $\boldsymbol{x}$ (*e.g.*, sparsity, known subspace models), which will eventually allow for a reduction of the number of samples mp *below* the dimensionality n of the input signal.

References

[1] M. A. Herman and T. Strohmer, "General deviants: An analysis of perturbations in compressed sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 342–349, 2010.

[2] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012, pp. 210–268.

[3] L. Balzano and R. Nowak, "Blind calibration of networks of sensors: Theory and algorithms," in *Networked Sensing Information and Control*. Springer, 2008, pp. 9–37.

[4] C. Bilen, G. Puy, R. Gribonval, and L. Daudet, "Convex Optimization Approaches for Blind Sensor Calibration Using Sparsity," *IEEE Transactions on Signal Processing*, vol. 62, no. 18, pp. 4847–4856, Sep. 2014.

[5] J. Romberg, "Compressive sensing by random convolution," *SIAM Journal on Imaging Sciences*, vol. 2, no. 4, pp. 1098–1128, 2009.

[6] T. Bjorklund and E. Magli, "A parallel compressive imaging architecture for one-shot acquisition," in *2013 IEEE Picture Coding Symposium (PCS)*. IEEE, 2013, pp. 65–68.

[7] K. Degraux, V. Cambareri, B. Geelen, L. Jacques, G. Lafruit, and G. Setti, "Compressive Hyperspectral Imaging by Out-of-Focus Modulations and Fabry-Pérot Spectral Filters," in *International Traveling Workshop on Interactions between Sparse models and Technology (iTWIST)*, 2014.

[8] S. Bahmani and J. Romberg, "Lifting for Blind Deconvolution in Random Mask Imaging: Identifiability and Convex Relaxation," *SIAM Journal on Imaging Sciences*, vol. 8, no. 4, pp. 2203–2238, 2015.

[9] J. P. Dumas, M. A. Lodhi, W. U. Bajwa, and M. C. Pierce, "Computational imaging with a highly parallel image-plane-coded architecture: challenges and solutions," *Opt. Express*, vol. 24, no. 6, pp. 6145–6155, Mar 2016. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-24-6-6145

[10] B. Friedlander and T. Strohmer, "Bilinear compressed sensing for array self-calibration," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov. 2014, pp. 363–367.

[11] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Problems*, vol. 31, no. 11, p. 115002, 2015.

[12] J. Lipor and L. Balzano, "Robust blind calibration via total least squares," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4244–4248.

[13] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Transactions on Information Theory*, vol. 60, no. 3, pp. 1711–1732, 2014.

[14] S. Ling and T. Strohmer, "Blind Deconvolution Meets Blind Demixing: Algorithms and Performance Bounds," *arXiv preprint arXiv:1512.07730*, 2015.

[15] E. Candès, X. Li, and M. Soltanolkotabi, "Phase Retrieval via Wirtinger Flow: Theory and Algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[16] C. D. White, S. Sanghavi, and R. Ward, "The local convexity of solving systems of quadratic equations," *arXiv:1506.07868 [math, stat]*, Jun. 2015, arXiv: 1506.07868.

[17] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *arXiv preprint arXiv:1602.06664*, 2016.

[18] V. Cambareri and L. Jacques, "Through the Haze: A Non-Convex Approach to Blind Calibration for Linear Random Sensing Models," 2016, in preparation.

[19] M. M. Hayat, S. N. Torres, E. Armstrong, S. C. Cain, and B. Yasuda, "Statistical algorithm for nonuniformity correction in focal-plane arrays," *Applied Optics*, vol. 38, no. 5, pp. 772–780, 1999.

Effect of power amplifier nonlinearity in Massive MIMO: in-band and out-of-band distortion.

Steve Blandino[1,2] Claude Desset[2] Sofie Pollin[1,2] Liesbet Van der Perre[1]

[1] KUL ESAT-TELEMIC, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[2] Imec, Kapeldreef 75, B-3001 Leuven, Belgium
Contact author: steve.blandino@imec.be

Abstract

Nonlinear power amplifiers distort the transmitted signal driving to signal degradation, in addition, out-of-band (OOB) radiation becomes a source of interference for users operating in adjacent channels. Massive MIMO relies on channel based precoding which ensures the signal is added constructively at the receiver user equipment. However, the effect of the precoding on the nonlinear distortion should be investigated since users could experience an higher bit error rate and a random user operating in adjacent bandwidth could be exposed to an increased interference. Assuming a third order polynomial model to describe the behavior of nonlinear power amplifier, we first show that in-band interference does not represent a problem in Massive MIMO due to the averaging of the signal over many antennas. Then, motivated by numerical results we find that OOB does not recombine constructively avoiding large interference. Massive MIMO allows to increase the in-band received power for the target users without increasing the interference in adjacent bands. In other words, less stringent requirements are demanded for power amplifier design, confirming that in Massive MIMO simpler hardware with respect to conventional transmission scheme is sufficient.

1 Introduction

5G, the next mobile communication standard is requiring a thousand fold system capacity increase but also reduced power consumption, high reliability and low latency, to satisfy the continuously increasing number of connected devices. Massive MIMO (MaMi) is a technology developed over the last 5 years to address these challenges. It can achieve many of the 5G design goals using existing hardware and without large new spectral resources thus it has been considered a most promising 5G radio technology.

MaMi relies on the use of a large number of independent antennas at the base station (BS) and serves simultaneously on a single resource tens of users exploiting the large multiplexing and antenna gain. These results can be achieved even with simple linear precoding and simpler hardware compared to previous communication systems [1].

MaMi must be designed with low-cost components to limit the implementation cost and power consumption of many RF chains. However, these low cost components are suffering from imperfections and non-idealities introducing distortion of the transmitted signal. Co-channel and adjacent channel interference arise in wireless transmission mostly due to nonlinear components, especially power amplifiers (PAs). The in-band distortion causes amplitude and phase deviation of the modulated symbol which degrades the received modulation increasing the bit error rate (BER). Moreover, distortion spreads the transmitted power into other bands, potentially disturbing neighboring bands. In order to limit this problems, RF power amplifiers have to be designed under constraints of limited distortion and restricted unwanted transmissions minimizing

harmful interference to devices operating in neighboring bands. The design of a new communication system should be concerned about those constraints. For example, based on LTE specs, the error vector magnitude for QPSK schemes shall be better than 17.5% and OOB radiation should not exceed $-13\,\mathrm{dBm/MHz}$ [2].

Modeling and simulation of non-ideal systems plays an important role in the evaluation of the overall communication system performance. Such analysis is essential for both design and standardization of the future communication system and should not be neglected in low cost and low power hardware solutions. In [3] the study of nonlinear PA in the context of the CDMA is proposed and the characterization of spectral regrowth at the output of nonlinear PA is derived. In MaMi scenarios, [4] quantifies the impact of PA nonlinear distortion in terms of the received signal-to-interference-plus-noise ratio (SINR) but no insight on the OOB is given. [5] studies the spatial distribution of the OOB concluding that the power received outside the band is substantially lower in MaMi than when only a single antenna is used.

In this paper we give a complete analysis of the behavior of power amplifier in MaMi in different scenario assuming a polynomial memoryless model to reproduce the behaviour of an nonlinear PA. We first verify whether in-band interference are a limiting factor in MaMi. Then we aim to verify whether coherent combination of the signal outside the band of interest occurs. In particular, we quantify the interference received by a random user operating in an adjacent band exploiting the concept of MIMO-ACLR introduced in [5]. We propose the analysis of the OOB for different system solutions and different precoding schemes.

Section 2 introduces the transmission scheme and the PA model. In Section 3 we model the transmitted PSD. Section 4 proposes numerical simulations and Section 5 concludes the paper.

2 System Model

This paper analyzes the downlink (DL) of a multiuser OFDM MaMi system. Due to the high peak-to-average power ratio (PAPR) OFDM-based systems are sensitive to nonlinear distortion, which drives the power amplifier into nonlinear region.

The base station (BS) is equipped with M antennas and serves simultaneously K single antenna users. The received DL signal $\mathbf{y} \in \mathbb{C}^{K\times 1}$ is modeled as:

$$\mathbf{y}_f = \mathbf{H}_f \mathbf{W}_f \mathbf{s}_f + \mathbf{z}_f \tag{1}$$

where the index f represents the subcarrier. The stochastic channel between the BS and the user equipments (UEs) is $\mathbf{H}_f \in \mathbb{C}^{K\times M}$ and it is modeled as $\mathcal{CN}(0, \mathbf{R})$. TDD is used, including an uplink (UL) pilot phase in order to let the BS estimate the channel. Based on this knowledge, the precoder $\mathbf{W}_f \in \mathbb{C}^{M\times K}$ is computed. $\mathbf{s} \in \mathbb{C}^{K\times 1}$ are the actual stochastic zero-mean data symbols and we further assume $E[||\mathbf{s}||^2] = \gamma$. The additive term $\mathbf{z}_f \in \mathbb{C}^{K\times 1}$ is the independent receiver noise generated with distribution $\mathcal{CN}(0, \sigma^2 I)$. The transmitter chain is simplified in Figure 1. The IFFT is performed on the precoded transmitted symbol vector $\tilde{\mathbf{x}}_f = \mathbf{W}_f \mathbf{s}_f$ and the cyclic prefix is added obtaining the time domain symbol vector $\tilde{\mathbf{x}}$ which is fed to M RF chains and identical power amplifiers (PAs) before the antennas. To model the behavior of the PAs, we assume a polynomial memoryless model [3]:

$$\mathbf{x}[n] = \sum_{\substack{p=1 \\ p\text{: odd}}}^{P} b_p \tilde{\mathbf{x}}[n] |\tilde{\mathbf{x}}[n]|^{p-1}, \tag{2}$$

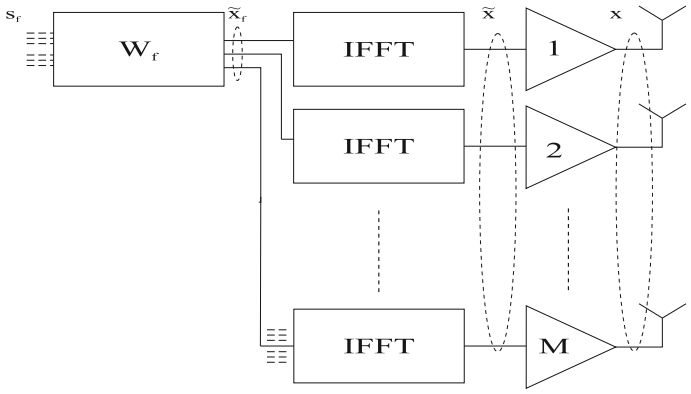


Figure 1: Simplified block diagram of the downlink of a OFDM-MaMi transmitter.

where b_p are complex coefficients representing the p-th order amplitude and phase distortions of the amplifier. Assuming $\mathbf{x}_f$ to be the corresponding frequency-domain of $\mathbf{x}$, (1) can thus be rewritten as:

$$\mathbf{y}_f = b_1 \mathbf{H}_f \mathbf{x}_f + \mathbf{H}_f \sum_{\substack{p=3 \\ p\text{: odd}}}^{P} b_p \mathbf{x}_{p_f} + \mathbf{z}_f, \tag{3}$$

which separates the extra term introduced by the PA nonlinearities from the desired signal.

3 Analysis of PA non-linearity

In this section we describe the effect of the PA nonlinearity on the transmitted PSD.

The power spectrum density of the signal at the output of the nonlinear device can be computed from the autocorrelation function of the output signal by using Wiener-Khinchin theorem:

$$\mathbf{S}_{\mathbf{xx}}(f) = \mathscr{F}(\mathbf{R}_{\mathbf{xx}}(\eta)). \tag{4}$$

We first assume $\mathbf{s}$ a circularly symmetric i.i.d. stationary process which gives the following data covariance matrix:

$$\begin{aligned} \mathbf{R}_{\mathbf{ss}}(\eta) =& \mathbb{E}\{\mathbf{s}[n]\mathbf{s}^H[n+\eta]\} = \\ & \begin{cases} \frac{\gamma}{K}\mathbf{I}_K, & \text{if } \eta = 0 \\ \mathbf{0}_K, & \text{otherwise} \end{cases}, \end{aligned} \tag{5}$$

where γ represents the total output power. The autocorrelation function of the precoded signal is:

$$\begin{aligned} \mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(\eta) =& \mathbb{E}\{\tilde{\mathbf{x}}[n]\tilde{\mathbf{x}}^H[n+\eta]\} \\ =& \mathbb{E}\{(\mathbf{W}[n] * \mathbf{s}[n])(\mathbf{W}[n+\eta] * \mathbf{s}[n+\eta])^H \\ =& \frac{\gamma}{K} \sum_n \mathbf{W}[n]\mathbf{W}^H[n+\eta]. \end{aligned} \tag{6}$$

The probability distribution of the precoded symbol $\tilde{\mathbf{x}}$ can be assumed to follow a Gaussian distribution. The Gaussian assumption holds by central limit theorem because the DL signal consists of the combination of many independent streams. The formulation of the autocorrelation function of the output of nonlinearity is greatly simplified if the input signal is assumed to have Gaussian probability distribution and in particular following [3][6][7] we can write:

$$\begin{aligned}\mathbf{R_{xx}}(\eta) =&\mathbb{E}\{\mathbf{x}[n]\mathbf{x}^H[n+\eta]\}\\ =&\mathbb{E}\{(b_1\tilde{\mathbf{x}}[n]+b_3\tilde{\mathbf{x}}[n]|\tilde{\mathbf{x}}[n]|^2)\cdot\\ &\cdot(b_1\tilde{\mathbf{x}}[n+\eta]+b_3\tilde{\mathbf{x}}[n+\eta]|\tilde{\mathbf{x}}|^2)^H\}\\ =&|b_1|^2\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}+\left[4\Re(b_1b_3^*)\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(0)+4(b_3)^2\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(0)\right]\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}+\\ &+2|b_3|^2{\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}}^3,\end{aligned} \tag{7}$$

where for simplicity $P=3$ has been considered. Now we can insert (7) in (4):

$$\begin{aligned}\mathbf{S_{xx}}(f) =&\tilde{b}_1\mathscr{F}(\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})+\tilde{b}_2\mathscr{F}(\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})+\tilde{b}_3\mathscr{F}({\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}}^3)=\\ &(\tilde{b}_1+\tilde{b}_2)\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(f)+\tilde{b}_3(\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(f)*\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(f)*\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(f)),\end{aligned} \tag{8}$$

where $\tilde{b}_1=|b_1|^2$, $\tilde{b}_2=4\Re(b_1b_3^*)\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(0)+4(b_3)^2\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(0)$, $\tilde{b}_3=2|b_3|^2$ and $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}=\mathscr{F}(\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})$. We also observe that in an OFDM system $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(f)=0,\forall|f|>B/2$, where B indicates the transmission bandwidth. The first component in (8) corresponds to the linear output term and the coefficient $\tilde{b}_1$ models the amplifier gain. The second term corresponds to the in-band nonlinear effect and depends on the input power level $\mathbf{R}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(0)$. The last term in (8) gives rise to the spectral regrowth. The convolution product indeed produces a spectral spreading over the baseband bandwidth.

4 Impact of nonlinear PAs in MaMi

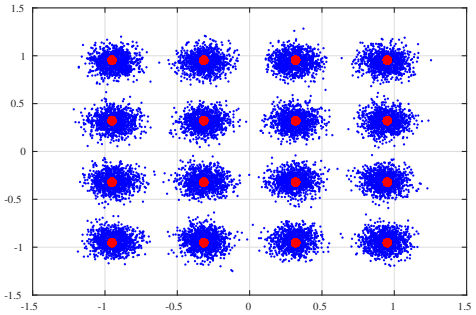
The impact of the nonlinear PAs in MaMi systems is now studied numerically. First we introduce the metrics used to analyze the system performance. Then we verify the effect of the in-band interference on the BER. Moreover the received PSD is simulated and the impact of nonlinear PA on the OOB is quantified, for different system load and precoding design.

To quantify the received constellation error we use the error vector magnitude (EVM). EVM is a measure of the difference between the ideal symbols and the measured symbols after the equalization. The difference is called the error vector. The EVM result is defined as the square root of the ratio of the mean error vector power to the mean reference power [2].

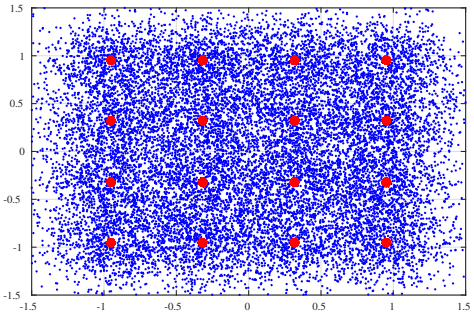
To evaluate the OOB radiation traditionally in single antenna system the Adjacent Channel Leakage power Ratio (ACLR) is used as a measure of the amount of power leaking into adjacent channels. The authors in [5] extended the ACLR for a multi-antenna system, in a random point in space for a fading environment as:

$$\mathrm{ACLR}=\frac{\max\left(\int_{F_\mathrm{l}-\frac{B}{2}}^{F_\mathrm{l}+\frac{B}{2}}\mathbb{E}[S_{y_\theta y_\theta}(f)]df,\int_{F_\mathrm{h}-\frac{B}{2}}^{F_\mathrm{h}+\frac{B}{2}}\mathbb{E}[S_{y_\theta y_\theta}(f)]df\right)}{\int_{F_\mathrm{c}-\frac{B}{2}}^{F_\mathrm{c}+\frac{B}{2}}\mathbb{E}[S_{y_\theta y_\theta}(f)]df}, \tag{9}$$

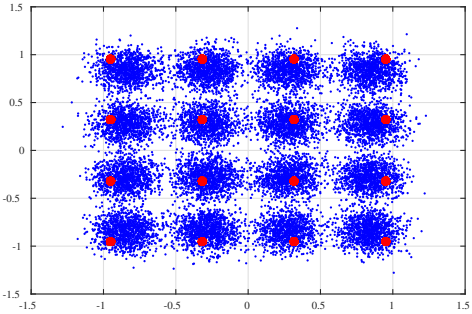
where F_c indicate the carrier frequency, F_l and F_h respectively the lower and the higher BS adjacent channel center frequency. Based on [2], assuming a channel bandwidth $B_\mathrm{ch}=20\,\mathrm{MHz}$, we set $B=18\,\mathrm{MHz}$, $F_l=F_c-B_\mathrm{ch}$ and $F_h=F_c+B_\mathrm{ch}$.



(a) Received PSD in a 100x10 MaMi system when SNR=0 and $P_{\text{IBO}} = 0\,\text{dB}$



(b) Received PSD in a 20x10 system when SNR=0 and $P_{\text{IBO}} = 0\,\text{dB}$



(c) Received PSD in a 100x10 MaMi system when SNR=0 and $P_{\text{IBO}} = -30\,\text{dB}$

Figure 2: Received 16-QAM

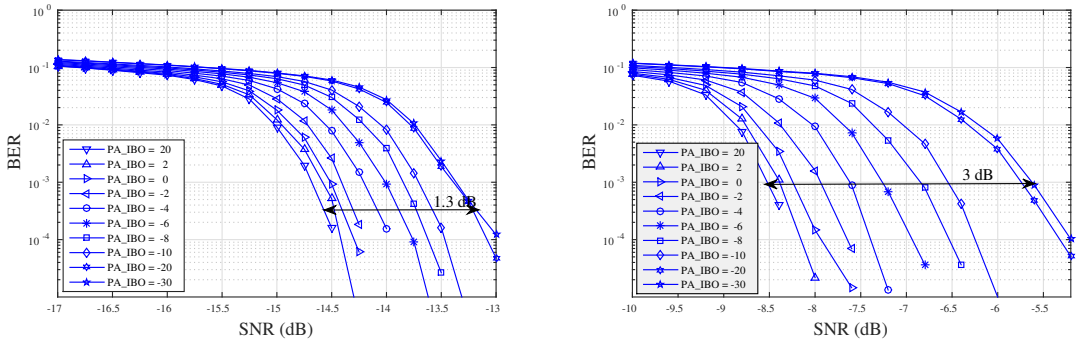
4.1 In-band distortion

Some of the inter-modulation product appear within the bandwidth causing in-band interference, as seen in equation (8), leading to an EVM degradation and worst performance in terms of BER.

We assume a 20 MHz OFDM system with 2048 subcarriers of which 1200 are actively allocated, based on LTE. A raised-cosine pulse shaping filter with a roll-off factor of 0.22 is used. The RF power amplifier follows a third order polynomial model. The PA operating point is generally characterized by the Power Input Backoff P_{IBO}, measuring the input signal margin with respect to $P_{\text{1-dB}}$, the 1-dB compression point where saturation effects become noticeable. We consider a multi-tap independent Rayleigh channel model. The average transmitted power per antenna is normalized to 0 dB. Based on $M = 100$ antennas, the total output power is hence $\gamma = 20\,\text{dB}$.

Figure 2(a) shows the received 16-QAM when $M = 100$, $K = 10$ and $P_{\text{IBO}} = 0\,\text{dB}$. The estimated EVM is -18 dB. We notice that the amplification distortion is uncorrelated with the symbol and the distortion is averaged out over the antennas. The averaging effect is obvious observing Figure 2(b) which proposes the received 16-QAM when $M = 20$, $K = 10$ and $P_{\text{IBO}} = 0\,\text{dB}$. The EVM increases to -9.9 dB. The result indicates that in-band distortion can be averaged out by increasing the number of antennas used enabling in MaMi setting the use of high efficient PA working in nonlinear region. As example Figure 2(c) shows the case in which $P_{\text{IBO}} = -30\,\text{dB}$, effectively operating in complete saturation which result to an EVM of -14.30 dB.

In OFDM system the P_{IBO} is usually a positive value of several dB. However, in MaMi it is possible to work in complete saturation as illustrated on Figures 3(a) and



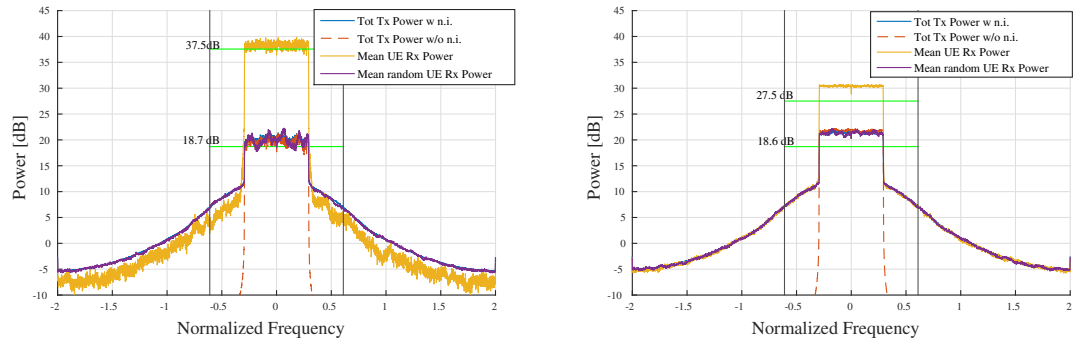
(a) QPSK (b) 16-QAM

Figure 3: P_{IBO} impact on BER, from linear operation (+20 dB) through moderate saturation (0 dB) up to complete saturation (-30 dB).

3(b). A completely saturated PA only leads 1.3 dB degradation when a QPSK is used while 3 dB degradation is observed when 16-QAM is employed.

In-band distortions do not represent a real problem in a MaMi system since the distortion can be lowered by increasing the number of antennas and negligible degradation in term of BER is observed.

4.2 Simulated in-band and out-of-band PSD

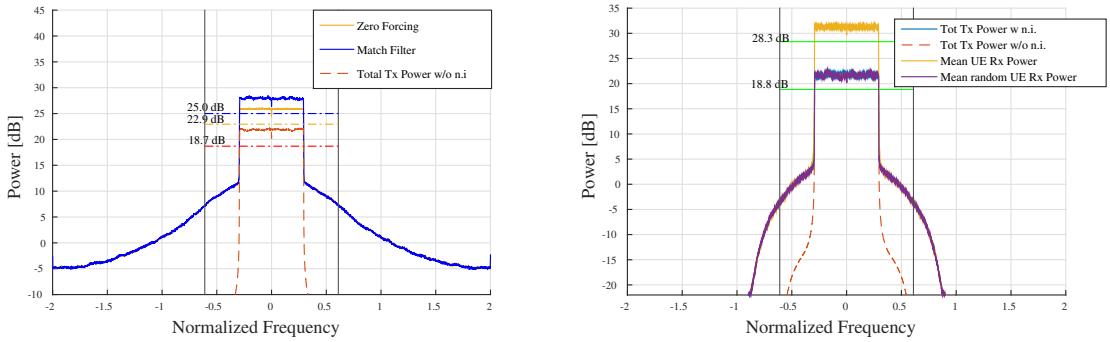


(a) $K = 1$ (b) $K = 10$

Figure 4: MaMi PSD with $M = 100$, $P_{\text{IBO}} = -30$ dB. The desired UE is compared to a random UE at a similar distance. The total BS output PSD is also provided with or without the non-ideal (n.i.) linearity behavior.

While in-band interferences do not recombine constructively, it is now not clear what the impact of precoding is on the OOB interference at the UE, or more importantly, at any possible location. A resulting concern is how the OOB will affect users operating in adjacent channels. In existing communication system OOB radiation is indeed the limiting factor, which force the use of low efficient PAs.

Figure 4(a) shows the power spectral density for a system with $M = 100$ transmitting antennas and $K = 1$ user. Observing the received PSD of the target user, we see that the effect of the array gain applies only in-band, giving a 20 dB gain, while the



(a) Comparison between ZF and MF precoding when $K = 25$, $P_{\text{IBO}} = -30\,\text{dB}$.

(b) $K = 10$ and $P_{\text{IBO}} = 0\,\text{dB}$.

Figure 5: MaMi PSD with $M = 100$.

received power in the adjacent band is equal to the transmitted power. This is interesting because, assuming a worst case scenario in which a user operating in an adjacent channel experiences exactly the channel of the target user, it does not receive any interference enhancement due to the applied precoding. Moreover, the received power of a user positioned at a random position follows the same behavior as the transmitted power both within and out of the band. The simulation highlights that coherent combination occurs only in band and that a user randomly located in space is not experiencing this combination gain. MaMi provides in this scenario an ACLR $= -39\,\text{dB}$, despite operating in complete PA saturation.

Similar observations can be extracted from Figure 4(b) where the system has been extended to $K = 10$ users. The in-band array gain is still 20 dB, but due to the presence of more users sharing the transmitted power, the relative difference between a desired user and a random user reduces to 10 dB. Here ACLR $= -28.7\,\text{dB}$ is obtained. The frequency fluctuations are also smaller with $K = 10$ than with $K = 1$ due to the averaging effects.

Let us assess the impact of different precoding. Figure 5(a) shows the performance of a MaMi with $M = 100$ and $N = 25$ using both ZF and MF. MaMi provides ACLR $= -26.3\,\text{dB}$ using MF while ACLR $= -24\,\text{dB}$ when ZF is applied. The difference between ZF and MF is approximately 2 dB. Applying the ZF precoding, the in-band gain goes to zero when increasing the number of the users in the system. This is straightforward from the definition of the ZF. The effect of canceling interference is equivalent reducing the number of antennas from M to $M - K$. Note that for smaller K the performance of ZF and MF tends to nearly the same but it is worth to remark that the load of the system and the computed precoder should be jointly designed to meet the OOB requirements.

In general, the transmitted signal power spectrum is regulated by a spectral mask. Some power input backoff is usually adopted to bound spectral regrowth within the mask limit. Figure 5(b) shows the PSD for a system with $M = 100$ and $K = 10$ when $P_{\text{IBO}} = 0\,\text{dB}$. As expected, increasing the power input backoff reduces the nonlinear distortion and ACLR $= 47.5$ dB is obtained. Increasing the backoff reduces the nonlinear distortion, but causes lower PA efficiency therefore minimizing power backoff is desirable. The impact of the P_{IBO} on the OOB and the minimum P_{IBO} which meets 3GPP mask limit will therefore have to be studied in future work.

We have verified that MaMi systems require less stringent OOB specifications thanks to the array gain observed only within the band through the large number of transmitting antennas. Even in single-user scenarios, the signal averaging over antennas and subcarriers is sufficient to prevent coherent combination of the OOB interference. MaMi enables to increase the in-band received power on selected UEs without

increasing the interference on adjacent bands.

5 Conclusion

This paper analyzes the effect of nonlinear PAs in MaMi. The analysis presented is based on the polynomial model of the PA. We show that both in-band interference and OOB interference do not recombine constructively and the array gain is experienced only inside the desired bandwidth. PAs working in complete saturation lead to negligible BER degradation and do not generate harmful interference to devices operating in neighboring bandwidth. Even in worst case scenario, when a random user shares the same channel as a target user, OOB radiations are not increased. Simulations confirm the benefit of the non-coherent addition of components coming from the different antennas. We prove that fully saturated PAs in MaMi provide better ACLR compared to traditional communication systems. Due to the high antenna gain, MaMi radiates less power while providing the same QoS as traditional systems, hence the amount of interference a user in adjacent channel is experiencing is lower with a MaMi setting.

Usually the power from the different PAs dominates the BS power consumption, due to the large output power. The use of high-efficiency nonlinear PAs would enable great gains in terms of energy efficiency. The operating region of the PAs is often determined by the OOB radiations and MaMi systems require less stringent OOB specifications thus PAs need relaxed linearity requirements with respect to traditional communication system.

In conclusion, in MaMi systems, in-band interference and OOB radiations are not limiting factors enabling the use of highly efficient saturated PAs and reducing the BS power consumption.

References

[1] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

[2] Access, Evolved Universal Terrestrial Radio, Base Station BS Radio Transmission and Reception, 3GPP TS 36.104, V10, 2011.

[3] K. G. Gard, H. M. Gutierrez, and M. B. Steer, "Characterization of spectral regrowth in microwave amplifiers based on the nonlinear transformation of a complex gaussian process," *IEEE Transactions on Microwave Theory and Techniques*, vol. 47, no. 7, Jul 1999.

[4] Y. Zou, O. Raeesi, L. Antilla, A. Hakkarainen, J. Vieira, F. Tufvesson, Q. Cui, and M. Valkama, "Impact of power amplifier nonlinearities in multi-user massive MIMO downlink," in *2015 IEEE Globecom Workshops (GC Wkshps)*, Dec 2015.

[5] C. Mollén, U. Gustavsson, T. Eriksson, and E. G. Larsson, "Out-of-band radiation measure for MIMO arrays with beamformed transmission," *arXiv preprint arXiv:1510.05513*, 2015.

[6] I. Reed, "On a moment theorem for complex gaussian processes," *IRE Transactions on Information Theory*, vol. 8, no. 3, pp. 194–195, 1962.

[7] KM Gharaibeh, KG Gard, and MB Steer, "Estimation of co-channel nonlinear distortion and sndr in wireless systems," *Microwaves, Antennas & Propagation, IET*, vol. 1, no. 5, pp. 1078–1085, 2007.

Bandwidth Impacts of a Run-Time Multi-sine Excitation Based SWIPT

Ning Pan Mohammad Rajabi Dominique Schreurs Sofie Pollin
TELEMIC Division
Department of Electrical Engineering
University of Leuven, Leuven, 3000, Belgium
{ning.pan, mohammad.rajabi, dominique.schreurs,sofie.pollin}@kuleuven.be

Abstract

Simultaneous wireless information and power transfer (SWIPT) utilizes radio frequency (RF) source as an information and energy carrier, which enables wireless networks consisting of a large amount of devices operating permanently without replacing batteries. SWIPT has attracted considerable attention since it is a promising method towards realization of future information and communication technology concepts such as internet of things (IoT). A number of studies have investigated techniques such as pre-coding and resource allocation to improve SWIPT performance. However, the transmission signal bandwidth (BW) impacts on the system have never been studied yet. Our paper originally scrutinizes how the transmission signal's BW affects received information and power. We consider a single link SWIPT system composed of a transmitter emitting RF signals and a receiver that is able to decode information and harvest energy simultaneously, which includes a power splitter after the antenna at the receiver. In addition, this SWIPT system is assumed as run-time forbidding any hardware modifications. A theoretical model describing channel capacity and received power depending on the circuit, which is a function of the input signal, is constructed specifically for SWIPT. In the simulation, the power splitting coefficient is fixed. The network throughput and harvested power is calculated based on the integrated SWIPT model. Simulation results show that a larger BW may lead to higher network throughput while the harvested power is degraded.

1 Introduction

The past decade has seen the rapid development in SWIPT techniques, since it is one of the possible approaches to realize permanently operating wireless networks without battery replacement [1]-[2]. There exist a growing number of studies investigating an efficient SWIPT system, benefiting from the ubiquitous and controllable characteristics of the radio frequency (RF) source [2]-[5].

Two performance metrics, the link throughput and harvested RF power, are considered to evaluate SWIPT system performance. Most of the previous studies are devoted to optimize the system applying techniques such as scheduling, resource allocation, and beamforming by trading-off the link-throughput and harvested power [2]-[4]. However, most of them ignore signal waveform's influence on the harvested power, by simplifying power conversion efficiency (PCE) as constant. Researchers in the hardware community have reported that the PCE depends not only on the circuit design but also the signal waveform design [6]-[16].

A group of research have proved experimentally that high peak-to-average-ratio (PAPR) waveform improve the PCE [6]-[8][13], while another work shows that CW excitations has a higher PCE compared to modulated signals. In order to have a concrete understanding for these ad-hoc measurements, theoretical model including excitation signals' characteristics and rectifier's non-linear behavior is necessary. Several studies investigating transmission signals have been carried out to model the rectifier's behavior [8][15][16]. The authors of [8] have varified that a multi-sine signal with zero-phase difference between tones will have the highest PCE, by approximating the diode's nonlinearity behavior using Taylor expansions. Their work would have been more complete if the other circuits' components influence are included. Reference [15] constructed a theoretical model calculating PCE specifically for high PAPR waveform excited rectifier based on the measured output voltage. However, this model is only valid in high input power case since it assumes perfect matching and ideal diode. The work of [16] analyzed amplitude and bandwidth impacts of the multi-sine signal based on the statistical information of output signal after the rectifier. The authors have showed that the PCE improves with a higher PAPR waveform, while the system performance degraded with increasing BW because of input network mismatch. Nevertheless, this analysis is unable to neglect the matching network's influence.

In a SWIPT system, BW has both effects on capacity and power. The BW analysis of capacity typically follows Shannon's theory [17], where a higher BW is preferred. Referring to [16], a wide BW may restrict system performance because of the input matching network. However, there is no work on the BW analysis for the multi-sine based SWIPT system to the best of our knowledge. This work analyzes BW impacts on capacity and PCE based on a combined information and power transfer model specifically for a multi-sine based run-time SWIPT, where only transmission signals can be tuned.

The rest of this paper is organized as follows. Section 2 demonstrates the single-link SWIPT model involving the related parameters and the performance metrics in SWIPT. Section 3 analyzes BW influence on the run-time system performance. In the end, section 4 concludes this work's contributions.

2 System Model

We are considering a multi-sine based run-time SWIPT system consisting of a source and a destination node transferring information and power simultaneously as can been seen in Fig. 1. The source transmits multi-sine signals to the destination via a wireless channel. After reception by the antenna with adapted matching network, the signal is divided into two parts with a fixed ratio ρ. One part goes through the information decoder to achieve the required data. In the information decoder, the received RF signal is first converted into baseband by the mixer; then the baseband signal passes through the signal processor to retrieve the source information. Additionally, the remaining signal is harvested by the rectifier. The RF signal goes through a non-linear rectifying component such as a diode after the adapted matching network. The DC output is finally obtained after an RC filter. The DC output is accumulated in the energy storage device with a feeding line connected to the information decoder for charging purposes.

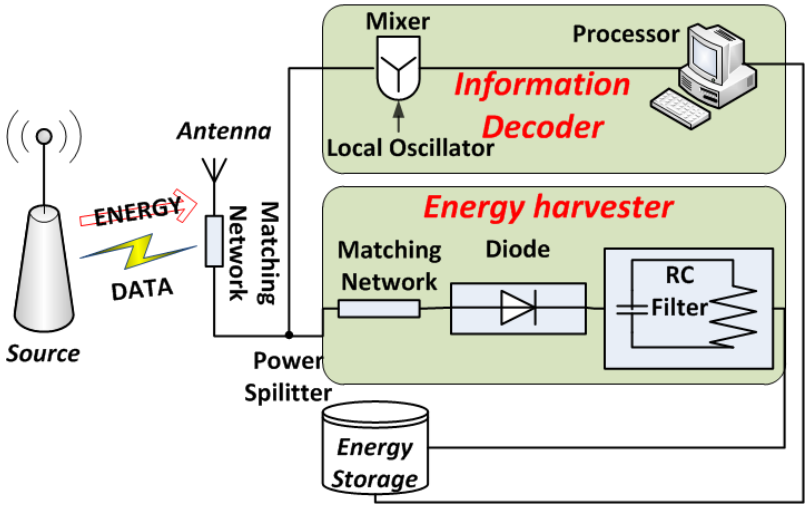


Figure 1: Single-link multi-sine based SWIPT system.

In this model, we define the time domain signal with lowercase letter with time variation t as $a(t)$, and the frequency domain signal with uppercase letter with frequency variation f as $A(f)$.

2.1 Joint Wireless Channel

In the SWIPT system, the RF signals first travel through the wireless channel before arriving at the destination. Specifically, the source node transmits an N_t-tone multi-sine wave with bandwidth BW. The time-domain transmission signal $s(t)$ is expressed as

$$s(t) = \Re\left(\sum_{n=1}^{N_t} x_n(t)\, e^{j2\pi f_n}\right). \tag{1}$$

Hereby, $x_n(t) = a_n e^{j\phi_n}$ is the n th baseband tone, where the statistical expectation of the signal equals the average power of each tone P_n as $\mathbb{E}\left[|x_n(t)|^2\right] = P_n$ with signal amplitude a_n and phase ϕ_n. In addition, the multi-sine signal is assumed to be equally distributed around the carrier frequency f_c with $f_n = f_c - \frac{B}{2} + \frac{(2n-1)B}{2N}$.

Then the RF signal passes the wireless channel described by the following model. Denote the frequency-selective channel between the source node to the user's antenna as $H(f)$, the noise as $N(f)$, and the frequency response of $s(t)$ as $S(f)$, the received signal at the antenna is expressed as

$$Y(f) = H(f)\, S(f) + N(f). \tag{2}$$

2.2 Separate Data and Power Reception

As aforementioned, the RF signal is split before the information decoder and energy harvester by a power divider. The RF signal of ρ portion goes into the information decoder assuming perfect matching and no distortion, where the channel capacity is

calculated based on the power and noise level. $(1-\rho)$ portion of the RF signal goes into the energy harvester producing a DC output. In this subsection, we model the rectifier behavior by investigating the circuit topology applying Kirchhoff's law and Shockley diode model. Clearly, the RF to DC model would have been more accurate and applicable in high frequency if parasitic effects were included. Hereby, we are considering the circuit's model in the baseband ignoring the parasitic effects, which reveals the essential non-linear rectifying behavior despite its simplicity.

Since the diode, which is the rectifier's key component, is a non-linear device and its performance strongly depends on the instantaneous voltage on the diode, the time domain signal $y_{in}(t)$ is obtained through the inverse Fourier transform of the signal response in frequency domain $Y(f)$ for further investigation. Denoting the antenna equivalent impedance as R_{ant}, the instantaneous voltage on the diode is computed based on the received signal in (2) by the following equation

$$v_{in}(t) = y_{in}(t) \cdot \sqrt{|R_{ant}|} = \mathcal{F}^{-1}\{Y(f)\} \cdot \sqrt{|R_{ant}|}. \tag{3}$$

In this work we consider a general rectifier topology, which can be extended to various topologies by including other components such as a clamper circuit. As shown

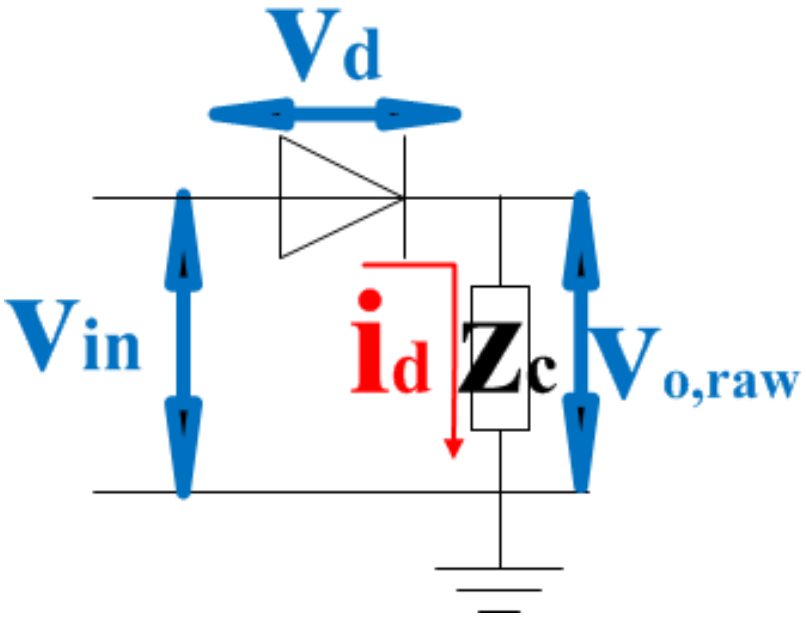


Figure 2: A general rectifier configuration.

in Fig. 2, a general rectifier consists of a diode and a load component which is an RC filter in most cases.

Obtaining the instantaneous voltage source $v_{in}(t)$ via (3), the output voltage $v_{o,raw}$ is derived by solving the equation sets below,

$$v_{in} = v_d + v_{o,raw}, \tag{4a}$$

$$v_{o,raw} = i_d \times z_c, \tag{4b}$$

$$i_d = i_s \left(e^{\frac{v_d}{nV_T}} - 1 \right), \tag{4c}$$

where the diode voltage is v_d, the diode current is i_d, the equivalent load impedance is denoted as z_c, and i_s, n, V_T represent the diode characteristic parameters. Herewith, (4c) describes the diode's current-voltage curve within the Shockley diode model which properly approximates the non-linearity.

After the diode, the RC circuit functions as a filter passing baseband signal while blocking the other frequency components in order to convey $v_{o,raw}$ to a stable DC voltage. The RC filter is described by a cut-off frequency $f_{cut-off}$ at which the signal is attenuated by 3 dB from the nominal passband value in frequency domain. Denote the frequency domain response of $v_{o,raw}$ as $V_{o,raw}$ and the frequency mask of RC filter as $M\left(f_{cut-off}\right)$, the output DC voltage V_{out} is computed in frequency domain as

$$V_{out} = V_o \times M\left(f_{cut-off}\right). \tag{5}$$

2.3 Performance Metrics

In SWIPT, two performance metrics are mainly considered, which is link throughput and harvested power. The link throughput evaluates the information transfer quality, while the harvested power measures the power transfer efficiency. Since RF signals are reused as information and power carrier in SWIPT, these two performance metrics may easily conflict with each other in the overall optimization problems [2]-[4]. However, BW impact on both metrics has not been considered before to the best of our knowledge. We relate BW to these two metrics in this subsection.

The SNR of the received signal Y is computed as

$$SNR = \frac{\rho Y^2}{N}. \tag{6}$$

Hereby, the noise $N = kT \times BW$ is a function determined by thermal temperature T and signal bandwidth BW; ρ is the power ratio sent to the information decoder. Thus the total throughput is

$$C = BW \cdot \log_2\left(1 + SNR\right). \tag{7}$$

The effective harvested power depends on the DC output of the energy harvester (5). Denote the load resistance as R_L, the harvested power is

$$P_o = \frac{V_{out}^2}{R_L}. \tag{8}$$

This equation is valid when the antenna matching is properly designed. Otherwise, the actual harvested power is calculated by including reflections in (8). In fact, since V_{out} is a value derived by (4) and (5) which strongly depends on the input instantaneous signal (3), V_{out} is a BW dependent parameter.

3 Bandwidth Impacts

As demonstrated in 2.3, the transmission signal BW not only influences the received information but also has an impact on the received power. In order to investigate the BW influence on the SWIPT system in detail, Matlab simulations were performed based on the constructed model in 2. In this work, we consider an AWGN channel ($H(f) = 1$) without influence on the generality of this analysis.

We consider a run-time SWIPT system, which refers to a user with a priorly constructed system. In this case, no hardware modification is allowed so that the system is optimized through software design. In this simulation, the parameter settings are as follows: $f_{cut-off} = 0.3$ MHz, $P_{in} = -10$ dBm, $\rho = 10^{-10}$, and $R_L = 10$ kOhm. Additionally, we adopt the diode characteristics of the zero-bias Schottky diode HSMS 2850. Moreover, perfect matching network is assumed for all input signals.

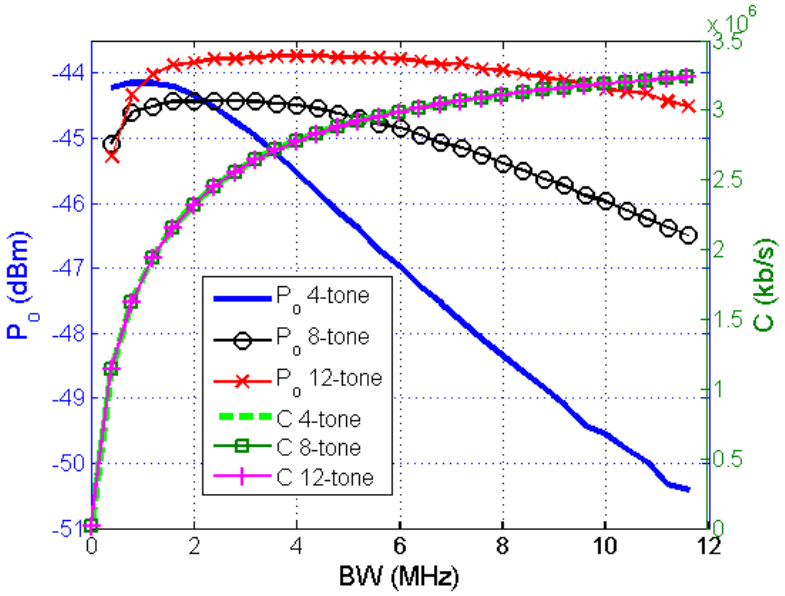


Figure 3: Output power and link variation with increasing BW in *run-time SWIPT*.

Fig. 3 shows how the output power and capacity change with increasing BW for a 4-tone, 8-tone, 12-tone transmission signal, respectively. The capacity is a concave function of BW. The capacity rises steeply in the BW limited regime until 10 MHz. This is because when the BW is small, it has more impact compared to SNR in (7). However, the capacity improvement saturates after 10 MHz, in the power limited regime, since noise is so huge that the power is not high enough to achieve a sufficient SNR. In addition, the power curves indicate the output DC power variation as a function of BW. A specific optimal BW value is observed for each curve depending on the number of tones.

Since the rectifier is a non-linear device, most of the output intermodulation products are accumulated at the frequency bin $\Delta f = \frac{BW}{N}$ for an equally spaced multi-sine input. For instance, the optimal BW for power transfer for a 12-tone signal excited SWIPT is about 3 MHz. The strongest intermodulation product occurs around 0.3 MHz in this case, which corresponds to the RC filter mask. When the signal BW ex-

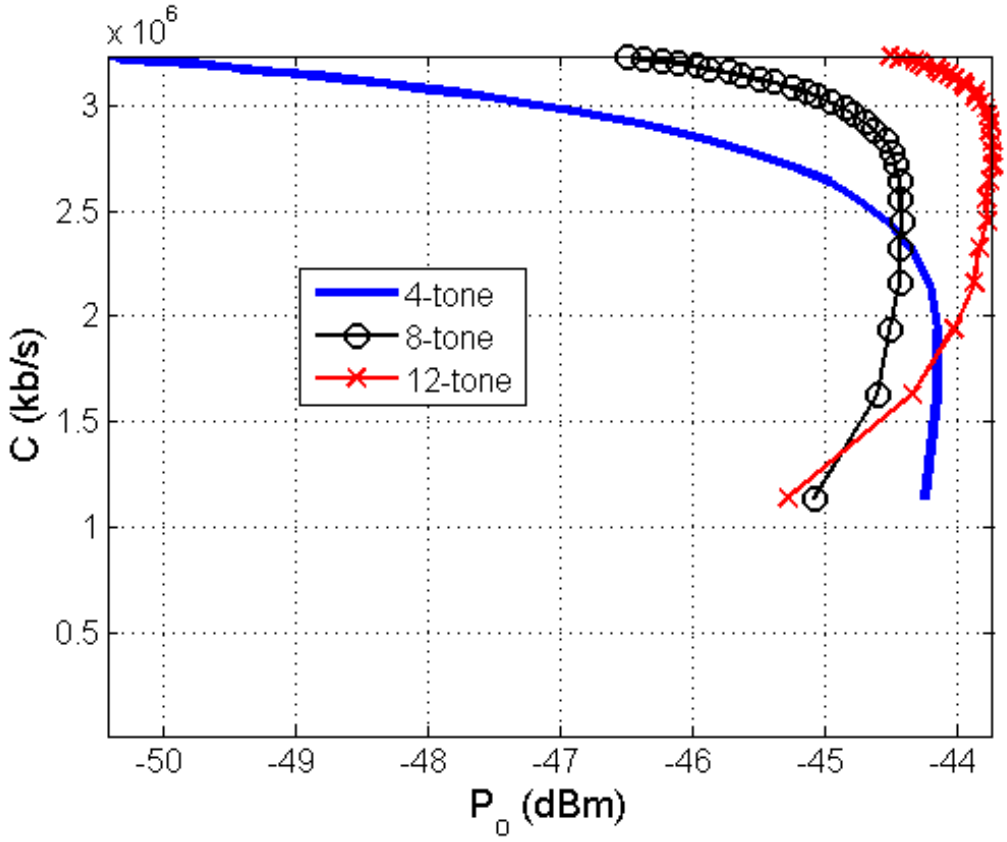


Figure 4: Output power and link throughput boundary of 4-tone, 8-tone, 12-tone signal with BW ranging from 0 to 12 MHz in *run-time SWIPT*.

ceeds 3 MHz, the first intermodulation product is filtered out by the RC filter resulting in a power transfer deterioration. Interestingly, a small rise before the optimal BW is observed in all power curves; with increasing number of tones, the power rises sharper in the beginning. This is because when the BW is small, the time period between high peaks is so large that the RC filter cannot charge to the saturation level, which is more severe for signals with larger number of tones. In short, a multi-signal consisting of more tones enables a larger optimal BW with the same RC filter as can be seen in Fig. 3.

Fig. 4 depicts the power and capacity bound of the 4-tone, 8-tone, and 12-tone excited SWIPT, respectively. It is observed that there is a trade-off between harvested power and link throughput. What's more, the SWIPT system with more multi-sines may have a higher harvested power while the link throughput remains the same.

Clearly, a proper BW choice will improve SWIPT performance when the system is constructed beforehand. In the BW limited regime, a larger BW produces more capacity, where a multi-sine with more tones is necessary to ensure enough converted power.

4 Conclusion

This paper established an integrated SWIPT model including the conventional wireless communication channel and the RF to DC channel. BW is related to the output power and link throughput through the constructed model. It is demonstrated that optimal BW and multi-sine signal choice, depending on the RC filter and tone number, improves the SWIPT performance significantly in the run-time SWIPT.

Acknowledgement

The acknowledgement goes to FWO and Hercules for the financial support on this research.

References

[1] S. Ulukus, A. Yener, E. Erkip, O. Simeone, M. Zorzi, P. Grover, and K. Huang, "Energy harvesting wireless communications: A review of recent advances," *IEEE Journal of Selected Areas in Communications*, vol. 33, no. 3, pp. 360–381, 2015.

[2] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Wireless Networks with RF Energy Harvesting: A Contemporaty Survey," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 2, pp. 757–789, May 2015.

[3] B. Clerckx, E. Bayguzina, D. Yates, and P. D. Mitcheson, "Waveform Optimization for Wireless Power Transfer with Nonlinear Energy Harvester Modeling," in *2015 IEEE Int. Symp. on Wireless Communications Systems*, Brussels, August 2015.

[4] E. Boshkovska, D. W. K. Ng, N. Zlatanov, and R. Schober, "Practical Non-Linear Energy Harvesting Model and Resource Allocation for SWIPT Systems," *IEEE Communications Letters*, vol. 19, no. 12, pp. 2082–2085, Sep. 2015.

[5] B. Clerckx, "Waveform optimization for swipt with nonlinear energy harvester modeling," in *20th Int. Workshop on Smart Antennas*, Munich, March 2016.

[6] M. S. Trotter, J. D. Griffin, and G. D. Durgin, "Power-Optimized Waveforms for Improving the Range and Reliability of RFID Systems," in *IEEE Int. Conf. on RFID Virtual J.*, April 2009, pp. 80–87.

[7] C.-C. Lo, Y.-L. Yang, C.-L. Tsai, C.-S. Lee, and C.-L. Yang, "Novel Wireless Impulsive Power Transmission to Enhance the Conversion Efficiency for Low Input Power," in *Int. Microwave Workshop Series on Innovative Wireless Power Transmission*, May 2011, pp. 55–58.

[8] A. Boaventura and N. B. Carvalho, "Maximizing DC Power in Energy Harvesting Circuits," in *IEEE MTT-S International Microwave Symp. Dig.*, June 2011, pp. 1–4.

[9] C.-L. Yang, C.-L. Tsai, Y.-L. Yang, and C.-S. Lee, "Enhancement of Wireless Power Transmission by Using Novel Multitone Approaches for Wireless Recharging," *IEEE Antennas Wireless Propag. Lett.*, vol. 10, pp. 1353–1357, Dec. 2011.

[10] A. Collado and A. Georgiadis, "Improving Wireless Power Transmission Efficiency Using Chaotic Waveforms," in *IEEE MTT-S International Microwave Symp. Dig.*, June 2012, pp. 1–3.

[11] C. Valenta and G. Durgin, "Rectenna Performance Under Power-optimized Waveform Excitation," in *IEEE Int. Conf. on RFID*, May 2013, pp. 237–244.

[12] A. Boaventura, A. Collado, A. Georgiadis, and N. Carvalho, "Spatial Power Combining of Multi-Sine Signals for Wireless Power Transmission Applications," *IEEE Trans. Microw. Theory Techn.*, vol. 62, no. 4, pp. 1022–1030, Apr. 2014.

[13] A. Collado and A. Georgiadis, "Optimal Waveforms for Efficient Wireless Power Transmission," *IEEE Microw. Compon. Lett.*, vol. 24, no. 5, pp. 354–356, May 2014.

[14] H. Sakaki, T. Kuwahara, S. Yoshida, S. Kawasaki, and K. Nishikawa, "Analysis of Rectifier RF-DC Power Conversion Behavior with QPSK and 16QAM Input Signals for WiCoPT System," in *Asia-Pacific Microwave Conference*, Sendai, Nov. 2014, pp. 603–605.

[15] C. R. Valenta, M. M. Morys, and G. D. Durgin, "Theoretical energy-conversion efficiency for energy-harvesting circuits under power-optimized waveform excitation," *IEEE Trans. Microw. Theory and Techn.*, vol. 63, no. 5, pp. 1758–1767, April 2015.

[16] N. Pan, A. S. Boaventura, M. Rajabi, D. Schreurs, N. B. Carvalho, and S. Pollin, "Amplitude and Frequency Analysis of Multi-sine Wireless Power Transfer," in *2015 IEEE Int. Workshop on Integrated Nonlinear Microwave and Millimetre-wave Circuits*, Taormina, Oct. 2015, pp. 1–3.

[17] C. E. Shannon, "A mathematical theory of communication," *The Bell Ssystem Technical Journal*, vol. 27, pp. 379–423, July 1948.

Generalized Optimal Pilot Allocation for Channel Estimation in Multicarrier Systems

François Rottenberg*, François Horlin**, Eleftherios Kofidis*** and Jérôme Louveaux*
*ICTEAM, Université catholique de Louvain, Belgium
{francois.rottenberg,jerome.louveaux}@uclouvain.be
**OPERA, Université libre de Bruxelles, Belgium
fhorlin@ulb.ac.be
***Department of Statistics and Insurance Science, University of Piraeus, Greece
kofidis@unipi.gr

Abstract

This paper* addresses the design of MSE-optimal preambles for multicarrier channel estimation under a maximum likelihood or minimum mean squared error criterion. The derived optimality condition gives insight on how to allocate the power of the pilots that compose the preamble. While many papers show that equispaced and equipowered allocation is optimal, the generalized condition demonstrates that there exist many different configurations that offer the same optimal performance. Furthermore, the condition applies not only to maximum likelihood but also to minimum mean squared error channel estimation. An application of the generalized condition in the presence of inactive subcarriers (virtual subcarriers problem) is shown such that a non equispaced allocation can achieve the same optimal performance as if an equispaced one could be used.

1 Introduction

Multicarrier systems aim at dividing a wideband signal into multiple narrowband signals centered around different subcarriers [1]. If the number of subcarriers is large with respect to the delay spread of the channel, each narrowband channel can be considered as frequency flat, which greatly simplifies the equalization task at the receiver. It is mainly for the their ability to effectively address the channel distortion that multicarrier systems have been so popular.

Channel estimation for multicarrier systems has been extensively studied; see, for instance, [2, 3] for recent review papers concerning channel estimation for orthogonal frequency division multiplexing (OFDM) or [4] for a review in offset QAM-based filterbank multicarrier (FBMC/OQAM) modulations. Specifically, the problem of optimizing the pilot allocation (including their position in frequency and their relative power) to minimize the mean squared error (MSE) of the channel estimate has been addressed in a number of works. Most of them focus on the case of least squares (LS) channel estimation, which corresponds to the maximum likelihood (ML) estimator under the Gaussian noise assumption and their main common conclusion is that the pilots should be equispaced and equipowered [5, 6]. The authors in [7] extend this result to the MIMO case and it is shown that optimal pilot sequences are equipowered, equispaced, and phase shift orthogonal.

Surprisingly, there are only few works addressing optimal pilot allocation for the minimum mean squared error (MMSE) estimator, even in the single-antenna case. In [8], the pilot allocation that maximizes the capacity for a MMSE channel estimator is shown to also correspond to equipowered, equispaced pilots. In [9], the authors derive the MSE-optimal training condition for the MMSE estimator in the context of a two-way relay OFDM-based network.

*The material in this paper has been partially submitted at IEEE SPAWC2016.

Very few papers actually investigate if other pilot configurations also satisfy the optimality condition and therefore offer the same performance as the equispaced configuration. In this paper, we find a sufficient and necessary condition for optimal pilot allocation that holds for both the maximum likelihood (ML) and the MMSE channel estimators. This result can be seen as generalizing the classical equispaced pilot configuration. Indeed, it is shown that a much wider family of pilot allocations may be optimal. Furthermore, the optimality condition can be directly used to solve the problem of allocating the pilots in the more realistic (and more challenging) case where inactive (virtual) subcarriers are present. The problem of virtual subcarriers is widely studied in the literature; see, for example, [10] for the single antenna case and [11] for the multiple antennas case. The originality of the work presented here is that it gives a methodology for finding optimal allocations, when the problem is feasible, which attains the same performance as if no null subcarriers were present.

The rest of the paper is organized as follows. Section 2 first serves as a reminder of channel estimation in multicarrier systems using the ML and the MMSE criteria. Section 3 derives a general optimality condition for pilot allocation that holds for the ML case and the MMSE estimator if the channel taps are assumed uncorrelated. Section 4 shows a possible application of the generalized condition to the virtual subcarriers problem. Simulation results are presented in Section 5. Section 6 concludes the paper.

Notations: Vectors and matrices are denoted by bold lowercase and uppercase letters, respectively. Superscripts *, ^{T} and ^{H} stand for conjugate, transpose and Hermitian transpose operators. tr, $\mathbb{E}$, $\Im$ and $\Re$ denote trace, expectation, imaginary and real parts, respectively.

2 Channel Estimation in Multicarrier Systems

We consider a multicarrier system with M subcarriers. Let us assume that the different subchannels can be considered flat and orthogonal to each other. A preamble of one multicarrier symbol is transmitted. The preamble vector composed of the transmitted pilots at the M subcarriers is denoted by $\mathbf{d} \in \mathbb{C}^{M\times 1}$. The channel impulse response $\mathbf{h} \in \mathbb{C}^{L_h\times 1}$ is assumed to be quasi-static[†] and L_h is the channel length. The vector of samples received after demodulation can be written as

$$\mathbf{y} = \mathbf{DF\Sigma h} + \boldsymbol{\eta}$$

where $\mathbf{F} \in \mathbb{C}^{M\times M}$ is the unitary discrete Fourier transform (DFT) matrix, $\mathbf{D} = \mathrm{diag}(\mathbf{d}) \in \mathbb{C}^{M\times M}$ contains the vector of transmitted pilot symbols $\mathbf{d}$ on its diagonal and $\boldsymbol{\eta} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_\eta)$ is additive Gaussian noise. Moreover, the noise is assumed to be white, i.e. $\mathbf{C}_\eta = \sigma^2\mathbf{I}_M$.[‡] $\mathbf{\Sigma} = \begin{pmatrix}\mathbf{I}_{L_h} & \mathbf{0}^H_{(M-L_h)\times L_h}\end{pmatrix}^H$ can be seen either as a selection matrix of the first L_h columns of $\mathbf{F}$ or as a zero padding matrix that appends $M - L_h$ 0's to the vector $\mathbf{h}$.

Note that this model fits both an OFDM system if the cyclic prefix is at least as long as the channel order [2] and an FBMC/OQAM system if the channel frequency selectivity is sufficiently mild and the noise correlation is neglected [4]. Then $\mathbf{d}$ represents the so-called pseudo-pilots (for fully loaded preambles) or the pilots (otherwise) [12–14].

[†]We assume, as usual, that the channel remains invariant in the duration of a multicarrier symbol.

[‡]The whiteness of the noise samples makes sense in an OFDM system while this is a stronger assumption in an FBMC system where correlation exists in both time and frequency [4].

2.1 Maximum likelihood channel estimator

We here assume that $\mathbf{h}$ is deterministic and unknown. Denoting by $f(\mathbf{y}|\mathbf{h})$ the conditional probability density function of $\mathbf{y}$ given $\mathbf{h}$, the ML estimator of the channel is given by [15]

$$\begin{aligned}\hat{\mathbf{h}}_{\mathrm{ML}} &= \arg\max f(\mathbf{y}|\mathbf{h}) \\ &= \left(\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{D}^H\mathbf{y}\end{aligned}$$

where $\mathbf{P} = \mathbf{D}^H\mathbf{D}$ is a diagonal matrix containing the power of each pilot on its diagonal. We define $\mathbf{p} \in \mathbb{C}^{M\times 1}$ as the vector containing the power of each pilot, such that $p_k = |d_k|^2$ and $\mathbf{P} = \mathrm{diag}(\mathbf{p})$. One can note that the ML estimator coincides here with the weighted LS estimator. The MSE is given by

$$\mathrm{MSE}_{\mathrm{ML}}(\mathbf{P}) = \sigma^2\mathrm{tr}\left[\left(\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\boldsymbol{\Sigma}\right)^{-1}\right].$$

This expression only depends on the power of the transmitted pilots $\mathbf{p}$ and not on their phase, which can be appropriately chosen for other purposes such as e.g., peak-to-average-power-ratio (PAPR) reduction [16].

2.2 MMSE channel estimator

We here consider that $\mathbf{h}$ follows a zero mean Gaussian distribution with correlation matrix denoted by $\mathbf{C}_h$, i.e. $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_h) \in \mathbb{C}^{L_h\times 1}$. The MMSE estimate $\hat{\mathbf{h}}_{\mathrm{MMSE}}$ is the one that minimizes $\mathbb{E}(\|\mathbf{h} - \hat{\mathbf{h}}_{\mathrm{MMSE}}\|^2)$. It can be shown to be equal to [15]

$$\begin{aligned}\hat{\mathbf{h}}_{\mathrm{MMSE}} &= \mathbb{E}(\mathbf{h}|\mathbf{y}) \\ &= \left(\mathbf{C}_h^{-1} + \frac{1}{\sigma^2}\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{\sigma^2}\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{D}^H\mathbf{y}\end{aligned}$$

and the corresponding MSE is

$$\mathrm{MSE}_{\mathrm{MMSE}}(\mathbf{P}) = \sigma^2\mathrm{tr}\left[\left(\sigma^2\mathbf{C}_h^{-1} + \boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\boldsymbol{\Sigma}\right)^{-1}\right]. \tag{1}$$

One can check that $\mathrm{MSE}_{\mathrm{MMSE}}(\mathbf{P}) \leq \mathrm{MSE}_{\mathrm{ML}}(\mathbf{P})$ and they will eventually converge at high signal-to-noise ratio (SNR), i.e. when $\sigma^2 \to 0$.

3 Optimality Condition for Channel Estimation

This section aims at deriving a sufficient and necessary condition for pilot allocations that solve the following pilot allocation problem subject to a training power constraint:

$$\min_{\mathbf{p}} \quad \mathrm{MSE}_{\mathrm{ML/MMSE}}(\mathbf{P}) \quad \text{s.t.} \quad \mathrm{tr}\left[\mathbf{P}\right] = P_T. \tag{2}$$

The following proposition gives an alternative necessary and sufficient condition that characterizes the optimal power allocations for the ML case. The condition also holds for the MMSE case under the (common) assumption that the channel taps are uncorrelated, i.e. $\mathbf{C}_h$ is a diagonal matrix with elements $\mathbf{C}_h = \mathrm{diag}(\lambda_h^1, \ldots, \lambda_h^{L_h})$.

Proposition 3.1. *Under the previous assumptions, any pilot allocation* $\mathbf{p} \in \mathbb{R}_+^{M\times 1}$ *is optimal in the sense of the minimum* MSE *for the ML and the MMSE estimators under a training power constraint if and only if (iff)* $\mathbf{p}$ *satisfies*

$$\sqrt{M}\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{p} = \begin{pmatrix} P_T \\ \mathbf{0} \end{pmatrix}. \tag{3}$$

Proof. The matrix $\mathbf{F}^H\mathbf{P}\mathbf{F}$ is circulant and has thus equal diagonal elements that we denote by x. Given the training power constraint, the value of x is independent of the structure of $\mathbf{P}$ and always equals

$$\begin{aligned} \operatorname{tr}\left[\mathbf{F}^H\mathbf{P}\mathbf{F}\right] &= Mx \\ x &= \frac{P_T}{M} \end{aligned}$$

where we used the cyclic trace property and the fact that $\operatorname{tr}[\mathbf{P}] = P_T$. The matrix $\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\boldsymbol{\Sigma}$ is the upper left $L_h \times L_h$ submatrix of $\mathbf{F}^H\mathbf{P}\mathbf{F}$ and therefore has diagonal elements equal to x. Using the fact that for a positive definite matrix $\mathbf{A}$, the following inequality [15, p. 65] holds,

$$\operatorname{tr}\left(\mathbf{A}^{-1}\right) \geq \sum_{i=1}^{L_h} a_{ii}^{-1} \tag{4}$$

where a_{ii} is the i-th diagonal element of $\mathbf{A}$ and equality holds iff $\mathbf{A}$ is diagonal, (1) can be lower bounded by

$$\mathrm{MSE}_{\mathrm{MMSE}}(\mathbf{P}) \geq \sum_{l=1}^{L_h} \left(\frac{1}{\lambda_h^l} + \frac{1}{\sigma^2}\frac{P_T}{M}\right)^{-1} \tag{5}$$

where equality holds iff $\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\boldsymbol{\Sigma}$ is diagonal. This condition can be rewritten as

$$\begin{aligned} [\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\boldsymbol{\Sigma}]_{l,m} &= \frac{1}{M}\sum_{k=0}^{M-1} p_k e^{j\frac{2\pi}{M}k(l-m)} \\ &= \begin{cases} \frac{P_T}{M} & \text{if } l = m \\ 0 & \text{if } l \neq m \end{cases} \end{aligned}$$

where $l = 0, 1, \ldots, L_h - 1$ and $m = 0, 1, \ldots, L_h - 1$. $\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\boldsymbol{\Sigma}$ is a Toeplitz Hermitian matrix. Then, it is sufficient to impose that its first column is 0 except for its first entry, i.e.

$$\sum_{k=0}^{M-1} p_k e^{j\frac{2\pi}{M}kl} = 0 \quad \forall l = 1, 2, \ldots, L_h - 1$$

which means that the inverse Fourier transform of the power allocation should be zero at indexes between 1 and $L_h - 1$ or in matrix form,

$$\sqrt{M}\boldsymbol{\Sigma}^H\mathbf{F}^H\mathbf{p} = \begin{pmatrix} P_T \\ \mathbf{0} \end{pmatrix}.$$

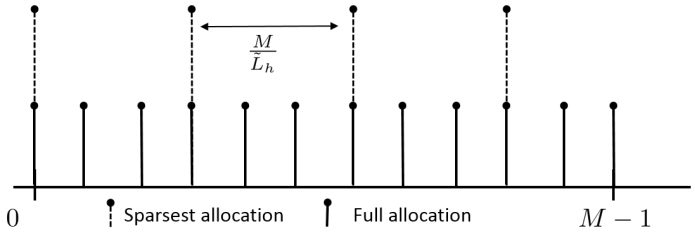


Figure 1: Two examples of optimal allocation: equispaced and equipowered allocation and full equipowered allocation.

To conclude, if $\mathbf{p}$ satisfies (3), $\mathbf{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\mathbf{\Sigma}$ is diagonal and reaches the lower bound in (5) and hence, $\mathbf{p}$ is optimal. In the other direction, if $\mathbf{p}$ is optimal, the lower bound in (5) should be satisfied and $\mathbf{\Sigma}^H\mathbf{F}^H\mathbf{P}\mathbf{F}\mathbf{\Sigma}$ has to be diagonal (by (4)), which is achieved only if (3) is satisfied. The previous derivations can be particularized to the case of the ML estimator setting $\mathbf{C}_h^{-1} = \mathbf{0}$ and the same result holds. This concludes the proof. □

The result of Proposition 3.1 and the following remarks can be related to the work in [16] which takes also the CP energy and PAPR into account. However, the approach in [16] is only concerned with the LS estimator whereas here the MMSE case is also addressed. Some remarks related to Proposition 3.1:

- ◇ If $\mathbf{C}_h$ is diagonal, the optimal pilot allocation is independent of the power delay profile (PDP) $\lambda_h^1, \ldots, \lambda_h^{L_h}$ and the noise level σ^2.
- ◇ Condition (3) imposes that the inverse Fourier transform (IFT) of the pilot allocation should only have zero coefficients between indexes 1 and $L_h - 1$ while the coefficient at index 0 represents the training power. This can somehow be seen as a requirement that $\mathbf{p}$ should not vary too slowly over the subarriers, except for the average value which represents the total training power. For instance, adding a cosine wave of frequency $\frac{2\pi l}{M}$ (and of amplitude such that the power allocation remains positive) to an optimal allocation will not affect the optimality if $l > L_h - 1$. Moreover, every optimal pilot allocation can be cyclically rotated arbitrarily in the frequency domain since this would simply correspond to multiplying by a complex exponential in the other domain, not affecting condition (3).
- ◇ There may be an infinite number of pilot allocations depending on L_h and M. Two classical allocations (see Fig. 1) are first the sparsest equispaced and equipowered allocation. Define $\tilde{L}_h$ as the smallest integer that divides M and such that $\tilde{L}_h \geq L_h$. Then, the sparsest allocation is given by $p_k = \frac{P_T}{\tilde{L}_h}, \forall k = 0, \frac{M}{\tilde{L}_h}, \ldots, (\tilde{L}_h - 1)\frac{M}{\tilde{L}_h}$ which is optimal since its IFT will have all elements at indexes smaller than $\tilde{L}_h$ equal to 0 (with $\tilde{L}_h \geq L_h$) except for the first one. A second is the full equipowered pilot allocation, i.e. $p_k = \frac{P_T}{M}, \forall k = 0, 1, \ldots, M-1$ since its IFT is a delta at 0. Furthermore, let us denote the two last power allocations $\mathbf{p}_1$ and $\mathbf{p}_2 \in \mathbb{R}_+^{M\times 1}$ respectively. Then, the convex combination $\mathbf{p}_3 = 0.5(\mathbf{p}_1 + \mathbf{p}_2)$ is an optimal allocation too.

4 Application in The Virtual Subcarriers Problem

In typical systems, a frequency mask should be respected imposing a limited out-of-band radiation. To ensure that the spectrum respects the mask, it is usual to keep

a number of subcarriers inactive at the edges of the band. Furthermore, in LTE-like systems, the time-frequency resources are "boxed" into different physical channels that may be transmitted simultaneously [17]. All of this imposes constraints on the possible pilot locations. Due to those multiple constraints, an equispaced pilot allocation may not always be possible, which complicates the problem [10, 11]. We here show that thanks to our generalized condition, optimality can still be reached in certain situations, which means that we can reach the same performance as if an equispaced or full equipowered allocation was possible.

Let us consider a system where no pilots can be transmitted at certain frequencies, i.e. $p_k = 0, \forall k \in \mathcal{K}$ where $\mathcal{K}$ is the set of inactive subcarriers. The number of remaining pilot positions is denoted by $N = M - |\mathcal{K}|$ and $\mathbf{p}_f \in \mathbb{R}^{N\times 1}$ is a vector made of the powers transmitted at those available subcarriers. Taking the virtual subcarriers into consideration, the optimality condition can then be rewritten as

$$\mathbf{A}\mathbf{p}_f = \mathbf{b} \quad \text{s.t.} \quad \mathbf{p}_f \in \mathbb{R}_+^{N\times 1} \tag{6}$$

where $\mathbf{A} = \sqrt{M}\mathbf{\Sigma}^H\mathbf{F}^H\mathbf{S}$, $\mathbf{b} = \begin{pmatrix} P_T \\ \mathbf{0} \end{pmatrix}$ and $\mathbf{S} \in \mathbb{R}^{M\times N}$ is formed by an identity matrix $\mathbf{I}_M$ where we removed the $|\mathcal{K}|$ columns corresponding to the virtual subcarriers frequencies. The constraint ensures that any element of $\mathbf{p}_f$ is real and greater than or equal to 0. Note that too many virtual subcarriers or badly placed may increase the ill-conditionning of the channel sensing [12]. In that case, there may not exist a solution to (6), i.e. an allocation that can still reach the optimal MSE. Finding a possible solution to (6) can be seen as a classical feasibility problem in the linear programming literature [18, Chap. 10]. Solving (6) is similar to finding an initial feasible point of a linear program. This can be solved efficiently in polynomial time outputting either a feasible point or the infeasibility of the problem.

5 Simulation Results

Fig. 2 shows an example of the virtual subcarriers problem. The simulation parameters are $M = 128$ subcarriers, a channel length $L_h = 10$ and the following sets of subcarriers, [0,5], [22,30], [43,48], [57,65], [123,M-1] are inactive, i.e. they cannot be used for training. The feasibility problem of (6) was solved using the Matlab function *linprog* and the solution is plotted in Fig. 2.

In the right figure of Fig. 2, the performance of the optimal pilot allocation with the ML and MMSE estimators is plotted. For both estimators, the power allocation is optimized so as to meet (3). A uniform PDP, i.e. $\lambda_h^l = \frac{1}{L_h}$, and an uncorrelated exponentially decaying delay profile is considered, i.e. $\lambda_h^l = \alpha 10^{-\frac{2(l-1)}{L_h}}$ such that $\mathrm{tr}\left[\mathbf{\Lambda}_h\right] = 1$ and $\mathbf{C}_h = \mathbf{\Lambda}_h$ in both cases. As expected, the MMSE estimator outperforms the ML one and the gap decreases at high SNR. The performance gap is larger for the exponentially decaying PDP than for the uniform PDP. This comes from the larger regularization effect of $\mathbf{C}_h^{-1}$ in (1) for the exponentially decaying PDP.

6 Conclusion

In summary, this paper addressed optimal pilot allocation for multicarrier systems for channel estimation under the ML and MMSE criteria. The obtained condition generalizes the commonly adopted equispaced pilot configuration, allowing for a much wider family of optimal allocations. This proves useful in practice, where null (virtual)

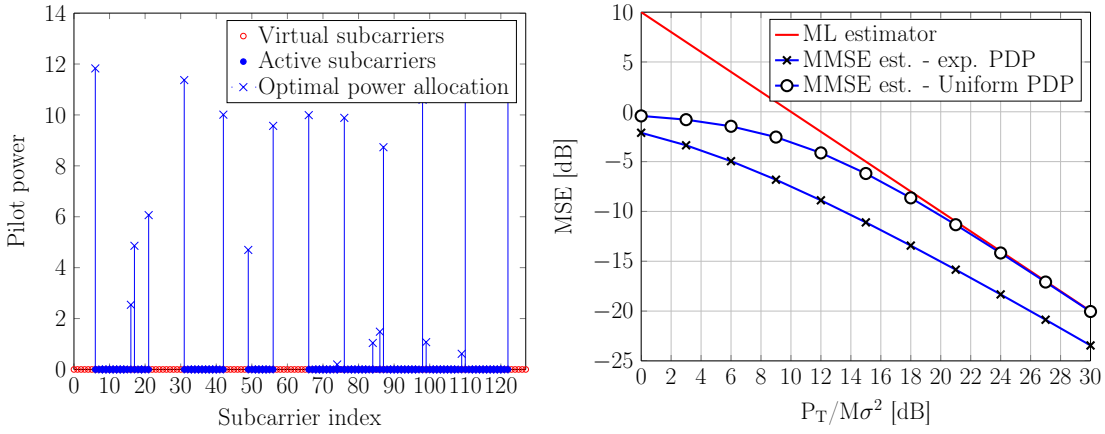


Figure 2: Due to system constraints, a large part of the band can not be allocated for training. However, there is still a feasible optimal allocation.

subcarriers are present as well. The reported simulation results demonstrated the value of the optimality condition in such a context.

This paper was concerned with single antenna systems. Extending these results to the multiple antennas case is a possible subject of future research.

Acknowledgment

The research reported herein was partly funded by Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture (F.R.I.A.).

References

[1] Y. G. Li and G. L. Stüber, *Orthogonal frequency division multiplexing for wireless communications*. Springer Science & Business Media, 2006.

[2] M. Ozdemir and H. Arslan, "Channel estimation for wireless OFDM systems," *IEEE Communications Surveys & Tutorials*, no. 9, pp. 18–48, 2007.

[3] Y. Liu, Z. Tan, H. Hu, L. Cimini, and G. Li, "Channel estimation for OFDM," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1891–1908, Fourth Quarter 2014.

[4] E. Kofidis, D. Katselis, A. Rontogiannis, and S. Theodoridis, "Preamble-based channel estimation in OFDM/OQAM systems: A review," *Signal Process.*, vol. 93, no. 7, pp. 2038–2054, July 2013. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2013.01.013

[5] R. Negi and J. Cioffi, "Pilot tone selection for channel estimation in a mobile OFDM system," *IEEE Transactions on Consumer Electronics*, vol. 44, no. 3, pp. 1122–1128, Aug 1998.

[6] J. Rinne and M. Renfors, "Pilot spacing in orthogonal frequency division multiplexing systems on practical channels," *IEEE Transactions on Consumer Electronics*, vol. 42, no. 4, pp. 959–962, 1996.

[7] I. Barhumi, G. Leus, and M. Moonen, "Optimal training design for MIMO OFDM systems in mobile wireless channels," *IEEE Transactions on Signal Processing*, vol. 51, no. 6, pp. 1615–1624, 2003.

[8] S. Ohno and G. B. Giannakis, "Capacity maximizing MMSE-optimal pilots for wireless OFDM over frequency-selective block Rayleigh-fading channels," *IEEE Trans. Information Theory*, vol. 50, no. 9, pp. 2138–2145, 2004.

[9] M. T. Tran, J. S. Wang, I. Song, and Y. H. Kim, "Channel estimation and optimal training with the LMMSE criterion for OFDM-based two-way relay networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, pp. 1–11, 2013.

[10] S. Ohno, E. Manasseh, and M. Nakamoto, "Preamble and pilot symbol design for channel estimation in OFDM systems with null subcarriers," *EURASIP Journal on Wireless Communications and Networking*, vol. 2011, no. 1, pp. 1–17, 2011.

[11] E. G. Larsson and J. Li, "Preamble design for multiple-antenna OFDM-based WLANs with null subcarriers," *IEEE Signal Processing Letters*, vol. 8, no. 11, pp. 285–288, 2001.

[12] F. Rottenberg, Y. Medjahdi, E. Kofidis, and J. Louveaux, "Preamble-based channel estimation in asynchronous FBMC-OQAM distributed MIMO systems," in *12th International Symposium on Wireless Communication Systems (ISWCS)*, 2015.

[13] L. Caro, V. Savaux, D. Boiteau, M. Djoko-Kouam, and Y. Louët, "Preamble-based LMMSE channel estimation in OFDM/OQAM modulation," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*. IEEE, 2015, pp. 1–5.

[14] D. Katselis, E. Kofidis, A. Rontogiannis, and S. Theodoridis, "Preamble-based channel estimation for CP-OFDM and OFDM/OQAM systems: A comparative study," *IEEE Trans. Signal Processing*, vol. 58, pp. 2911–2916, May 2010.

[15] S. M. Kay, *Fundamentals of Statistical Signal Processing. Vol. I: Estimation Theory.* Prentice-Hall, 1993.

[16] D. Katselis, "Some preamble design aspects in CP-OFDM systems," *IEEE Communications Letters*, vol. 16, no. 3, pp. 356–359, 2012.

[17] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband.* Academic Press, 2013.

[18] P. Bürgisser and F. Cucker, *Condition: The Geometry of Numerical Algorithms.* Springer Science & Business Media, 2013, vol. 349.

Co-existence of Cognitive Satellite Uplink and Fixed-Service Terrestrial

Jeevan Shrestha, and Luc Vandendorpe
ICTEAM institute, Université catholique de Louvain, Louvain La Neuve, 1348, Belgium
Email:{*jeevan.shrestha, luc.vandendorpe*}@uclouvain.be

***Abstract*—This paper considers an optimal and suboptimal frequency band allocation algorithms of non-exclusive/primary frequency bands for fixed transmitted power in the context of a cognitive satellite network. We consider N available primary frequency bands to be allocated toM cognitive satellite users such that the rate achieved by moving to the non-exclusive band is higher than the rate achievable in the exclusive band. After a cognitive user is moved to the primary band, the rate achieved by remaining users in the exclusive band is increased proportionally. Two suboptimal and an optimal frequency allocation algorithms are provided to address this varying requirement of the rate in each user switching in order to maximize achievable rate.**

The employment of cognitive network comes at a cost of implementing additional spectrum sensing unit. So it is natural to investigate whether it is sensible to implement spectrum sensing network in a certain frequency band or not. To address this issue, this paper also provides a condition in terms of the decision signal-to-noise ratio (SNR) threshold to check whether it is beneficial for a Satellite Communication user to implement spectrum sensing in a given primary band or not. This condition is primarily based on the probability of terrestrial transmitter being idle, the signal-to-noise ratio , and the detection probability. Numerical results show that the decision SNR threshold decreases with increase in source idle probability and increase in a number of cognitive satellite users. We provide two suboptimal (greedy) frequency allocation algorithms and compare their performances with baseline case. Simulation results are provided to demonstrate the performance of different algorithms.

I. Introduction

Satellite Communication (SatCom) is considered as a key element to offer seamless connectivity and communication service to any device, anywhere, anytime for future generation (5G) communication networks. The pervasive coverage achieved by SatCom makes them a suitable option to offload the data traffic of terrestrial 5G networks in geographical locations where implementing wired or wireless network is not economically and physically feasible. The ever-increasing demand for high data rate in broadband and broadcast services and limited availability of the usable spectrum have been a critical issue for SatCom network. To address this issue, although a number of high rate Ka-band satellite systems implementing multi-beam communication have already been employed, there still remains a substantial gap in meeting the spectral efficiency requirement of the next generation Terabit/s 2020 horizon [1]. In this context, cognitive satellite communication can be an attractive solution, which helps relieve the conventional spectrum scarcity and improve the utilization of the existing spectrum. A typical practice adopted by spectrum governance policies warrants exclusivity of spectrum use in a certain geographical area, but it has a fundamental limitation of low spectral utilization as reported by different measurement campaigns carried out at different parts of the world [2]. In principle, cognitive radio provides opportunistic spectrum access for dynamic spectrum usage with spectral co-existence among different networks. The potential advantages of cognitive radio come along with the compulsion to implement spectrum sensing unit as well as a prerequisite to providing adequate protection to the primary users to guarantee a certain quality of service (QOS).

Recently, a number of research projects have been commenced to address the spectrum sharing concept between two satellite systems or between satellite and terrestrial systems [4]. We consider the uplink of fixed satellite service (FSS) operating in the Ka band with terminals reusing frequency bands of Fixed-Service (FS) terrestrial microwave links (incumbent systems) as depicted in figure 1. We consider an interweave paradigm for cognitive SatCom network, such that the secondary user (satellite uplink) exploits spectrum holes or white spaces that exist in the primary user (terrestrial) spectrum. The primary objective is to maximize the achievable throughput of the SatCom networks by the adoption of CR techniques in the satellite uplink to allocate frequency.

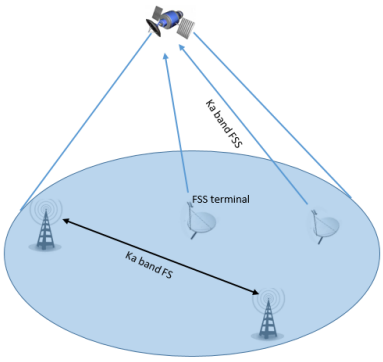


Fig. 1: Spectral coexistence of FSS uplink with the FS terrestrial link

In SatCom networks, usually the SatCom users are provided with exclusive bands in which they can communicate with the satellite in uplink transmission anytime. So it is instinctive that a user will implement cognitive radio (sensing network) only

if the rate achieved by moving to the non-exclusive band is higher than the rate achievable in the exclusive band. Even after obtaining the knowledge about spectrum availability, the white spectrum should be allocated to the users in an efficient manner. To address these issues, the contributions of this paper are twofold. First, a condition is provided to check whether it is beneficial for a SatCom user to implement sensing. This condition is primarily based on the probability of terrestrial transmitter being idle, the signal-to-noise ratio, and the detection probability. Second, an optimal frequency band allocation algorithm for fixed transmitted power is formulated to share the available spectrum holes amongst users efficiently.

II. System Model

III. Condition for Implementing Spectrum Sensing

In the existing cognitive model, transfer of a FSS user to a non-exclusive channel is done whenever the non-exclusive channel is detected free. It is beneficial to implement spectrum sensing for user k in an non-exclusive band μ only if the expected rate, $\mathrm{R}(\mathrm{k},\mu)$, when moving to that band is higher than the available rate, $\mathrm{R_E(k)}$, in the exclusive band, i.e.,

$$\mathrm{R}(\mathrm{k},\mu) \geq \mathrm{R_E(k)}. \tag{1}$$

Let $\mathrm{P_d}$, $\mathrm{P_f}$, and $\mathrm{P_s}(\mu)$ be the detection probability, the false alarm probability, the probability that the μ^{th} band is in idle state. By definition, we have $\mathrm{P_f} = \mathrm{P_r}(\hat{\mathrm{H}}_1/\mathrm{H}_0)$ and $\mathrm{P_d} = \mathrm{P_r}(\hat{\mathrm{H}}_1/\mathrm{H}_1)$, where $\mathrm{H}_0(\mathrm{resp.H}_1)$ is the event that the signal is actually absent (resp. present) and $\hat{\mathrm{H}}_0(\mathrm{resp.}\hat{\mathrm{H}}_1)$ the event that the signal is declared absent (resp. present). The expected rate, $\mathrm{R}(\mathrm{k},\mu)$ for user k, assuming that the user will switch to non-exclusive band μ whenever it senses the channel is idle is given by:

$$\begin{aligned}\mathrm{R}(\mathrm{k},\mu) &= \mathrm{R_{NE}}(\mathrm{k},\mu)\ \mathrm{P_r}(\hat{\mathrm{H}}_0/\mathrm{H}_0)\mathrm{P_r}(\mathrm{H}_0) \\ &\quad + 0 * \mathrm{P_r}(\hat{\mathrm{H}}_0/\mathrm{H}_1)\mathrm{P_r}(\mathrm{H}_1) + \mathrm{R_E(k)} \\ &\quad \Big(\mathrm{P_r}(\hat{\mathrm{H}}_1/\mathrm{H}_1)\mathrm{P_r}(\mathrm{H}_1) + \mathrm{P_r}(\hat{\mathrm{H}}_1/\mathrm{H}_0)\mathrm{P_r}(\mathrm{H}_0)\Big) \\ &= \mathrm{R_{NE}}(\mathrm{k},\mu)\ \mathrm{P_s}(\mu)(1-\mathrm{P_f}) \\ &\quad + \mathrm{R_E(k)}\ \left(\mathrm{P_d}(1-\mathrm{P_s}(\mu)) + \mathrm{P_f}\mathrm{P_s}\right), \end{aligned} \tag{2}$$

where we assume that if the cognitive FSS user transmits in certain non-exclusive band whenever it is busy, the rate achieved is zero. Then, the condition for implementing spectrum sensing for user k in μ^{th} non-exclusive band becomes

$$\begin{aligned}\mathrm{R_{NE}}(\mathrm{k},\mu) &\geq \frac{\mathrm{R_E(k)}\left(1-\mathrm{P_d}(1-\mathrm{P_s}(\mu)) + \mathrm{P_f}\mathrm{P_s}\right)}{\mathrm{P_s}(\mu)(1-\mathrm{P_f})} \\ \Leftrightarrow \mathrm{R_{NE}}(\mathrm{k},\mu) &\geq \mathrm{R_E(k)}\Phi(\mathrm{k},\mu), \end{aligned} \tag{3}$$

where $\Phi(\mathrm{k},\mu) = \frac{(1-\mathrm{P_d}(1-\mathrm{P_s}(\mu))+\mathrm{P_f}\mathrm{P_s})}{\mathrm{P_s}(\mu)(1-\mathrm{P_f})}$.

The condition for implementing spectrum sensing can be re-expressed in terms of SNR threshold, γ_{th}, as

$$\gamma_{\mathrm{k},\mu} \geq \gamma_{\mathrm{th}}, \tag{4}$$

where $\gamma_{\mathrm{th}} = 2^{\left(\frac{\mathrm{R_E(k)}\Phi(\mathrm{k},\mu)\mathrm{T}}{\mathrm{B_{NE}}(\mathrm{T}-\mathrm{T_s})}\right)} - 1$.

IV. Allocation of non-exclusive bands among different cognitive users

One of the major challenges for cognitive uplink satellite communications is to optimally assign available non-exclusive bands to multiple cognitive users. In the proposed approach, it is assumed that spectrum sensing is implemented only in the non-exclusive bands where the SNR is above the SNR threshold. Assuming a total number of M non-exclusive bands is free at any instant, the objective that we pursue is the maximization of the achievable rate in the satellite uplink.

In the current paper, we assume that each user is provided with the same rate in the exclusive band but extension for the case of different rates is quite straightforward. Let $\mathrm{a_{k\mu}} \in \{0,1\}$ denote the $\{\mathrm{k},\mu\}^{\mathrm{th}}$ element of an $\mathrm{M}\times\mathrm{N}$ non-exclusive band assignment matrix $\mathbf{A}$ with 1 denoting the assignment of μ^{th} (out of M) non-exclusive band to the k^{th} (out of N) cognitive FSS user. This way, $\mathbf{A}$ can be written as:

$$\mathbf{A} = \begin{pmatrix} \mathrm{a_{11}} & \dots & \mathrm{a_{1N}} \\ \vdots & \ddots & \vdots \\ \mathrm{a_{M1}} & \dots & \mathrm{a_{MN}} \end{pmatrix}.$$

It is assumed that only one non-exclusive band is assigned to a single FSS user at one time instant. Therefore, we have $\sum_{\mu=1}^{M} a_{k\mu} \leq 1$. Similarly, the $\mathrm{M}\times\mathrm{N}$ equivalent rate matrix in non-exclusive band can be written as

$$\mathbf{R_{eq}} = \begin{pmatrix} \mathrm{R_{eq}}(1,1) & \dots & \mathrm{R_{eq}}(1,\mathrm{N}) \\ \vdots & \ddots & \vdots \\ \mathrm{R_{eq}}(\mathrm{M},1) & \dots & \mathrm{R_{eq}}(\mathrm{M},\mathrm{N}) \end{pmatrix},$$

where the elements of the matrix for which sensing is not implemented is assigned to zero and $\mathrm{R_{eq}}(\mathrm{k},\mu)$ is given by

$$\mathrm{R_{eq}}(\mathrm{k},\mu) = \mathrm{R_{NE}}(\mathrm{k},\mu)(1-\mathrm{P_f}).$$

Then the optimization problem can be written as

$$\begin{aligned} \max_{\mathrm{a_{k\mu}},\mathrm{k}\in\{1,\dots,\mathrm{N}\},\mu\in\{1,\dots,\mathrm{M}\}} \quad & \sum_{\mathrm{k}=1}^{\mathrm{M}} \mathrm{a_{k\mu}} * \mathrm{R_{eq}}(\mathrm{k},\mu) \\ \text{subject to} \quad & \sum_{\mu=1}^{\mathrm{M}} \mathrm{a_{k\mu}} \leq 1, \end{aligned} \tag{5}$$

where the constraint means that each band can be allocated to at most one user.

This problem is of the integer assignment type. There exist well known algorithms such as the "Hungarian algorithm" to solve it efficiently. The throughput achieved by employing the Hungarian algorithm is also discussed in the performance analysis section. When the Hungarian algorithm is applied, all the available non-exclusive band are allocated to the FSS terminals in a single iteration. It is also worthy to consider an adaptive resource allocation algorithm where the rate available for FSS users in the exclusive band increases after moving a certain user to the non-exclusive band. Following subsection discuses two iterative algorithms for this adaptive resource allocation.

A. Greedy Algorithm I

This resource allocation algorithm is adaptive to the changing rate in exclusive band. The algorithm starts with identifying the FSS user that has the largest benefit to move. Let us denote this user by k_0. This user has one or several bands where the following condition is met:

$$\mathrm{R_{eq}}(\mathrm{k}_0, \mu) \geq \{\mathrm{R_E}(\mathrm{k}_0)\}_0 \tag{6}$$

where $\{\mathrm{R_E(k)}\}_0$ is the initial rate for FSS users in the exclusive band. User k_0 is moved to the non-exclusive band μ_0 corresponding to the highest $\mathrm{R_{eq}}(\mathrm{k}_0, \mu)$ in the first iteration, and band μ_0 becomes unavailable for the subsequent iteration. After this move, for other users the available rate in exclusive band becomes

$$\{\mathrm{R_E(k)}\}_1 = \frac{\mathrm{N}}{\mathrm{N}-1}\{\mathrm{R_E(k)}\}_0.$$

Actually the same procedure takes place at any iteration i: the FSS user that has the largest benefit to move into one of the remaining non-exclusive bands is identified and the rate for the users remaining in the exclusive band is updated as follows:

$$\{\mathrm{R_E(k)}\}_\mathrm{i} = \frac{\mathrm{N}-(\mathrm{i}-1)}{\mathrm{N}-\mathrm{i}}\{\mathrm{R_E(k)}\}_{\mathrm{i}-1}$$
$$\Leftrightarrow \ \{\mathrm{R_E(k)}\}_\mathrm{i} = \frac{\mathrm{N}}{\mathrm{N}-\mathrm{i}}\{\mathrm{R_E(k)}\}_0.$$

The iterations stop when there is no FSS user having benefit to move to the non-exclusive band. This algorithm is actually greedy since the FSS user with maximum profit is moved at any iteration but this may not be the best decision to take. The next algorithm is an optimal algorithm where switching of FSS user is done in such a way that the throughput is maximized.

B. Greedy Algorithm II

This algorithm starts with creating an index matrix where each element of the index matrix is a maximum integer value which satisfies following inequility:

$$\mathrm{i}(\mathrm{k}, \mu) \leq \mathrm{N} - \frac{\mathrm{N}\{\mathrm{R_E(k)}\}_0}{\mathrm{R_{eq}}(\mathrm{k}, \mu)}. \tag{7}$$

The inequility is straightforward extension of equation 7. The index matrix $\mathbf{I}$ is then can be written as:

$$\mathbf{I} = \begin{pmatrix} \mathrm{i}_{11} & \dots & \mathrm{i}_{1\mathrm{N}} \\ \vdots & \ddots & \vdots \\ \mathrm{i}_{\mathrm{M}1} & \dots & \mathrm{i}_{\mathrm{MN}} \end{pmatrix}$$

The $\{k, \mu\}^{th}$ element, $\mathrm{i}_{\mathrm{k}\mu}$, in the index matrix indicates upto which iteration it is benificial to move to μ^{th} non-exclusive band for the k^{th} user. The assignment of non-exclusive band to a user is done in reverse fashion. Let $\mathrm{i_{max}}$ be the maximul value of index in matrix $\mathbf{I}$. The index elements with greater of equal to certain value are grouped together.

Let U be the set of elements in matrix $\mathbf{I}$ such that $\mathrm{i}_{\mathrm{k}\mu} \geq \mathrm{i_n}$. Then a FSS user k^* is assigned to the μ^* exclusive band if $\{\mathrm{k}^*, \mu^*\} \in \mathrm{U}$ and $\mathrm{R_{eq}}(\mathrm{k}^*, \mu^*) \geq \mathrm{R_{eq}}(\mathrm{k}, \mu) \forall \{\mathrm{k}, \mu\} \in \mathrm{U}$. Then the particular user k^* and non-exclusive band μ^* is removed from the next iteration. The iteration goes in reverse manner and will be finished when iteration reches index value of 1. In this algorithm the user with maximum benifit is swithed at final iteration.

Parameter	Value
Non-exclusive Band	27.5-29.5 GHz
Exclusive Band	19.7-20.2 GHz (25 Users)
Parameters for FSS System	
EIRP	50dBW
$\mathrm{G_{Tx}^{FSS}}(0)$	42.1 dBi
D	35,786 km
$[\mathrm{G/T}]_{\mathrm{Rx}}^{\mathrm{SAT}}$	19.3dB/k
Sensing Network parameters	
Detection probability, $\mathrm{P_d}$	0.85
False detection probability, $\mathrm{P_f}$	0.2
Channel idle probability, $\mathrm{P_s}$	0.4
T	10μs
Total rate in exclusive band, $\mathrm{R_T}$	500Mbps

TABLE I: Simulation Parameters

V. Performance Evaluation

In this section, we present some numerical results to show the performance of the proposed resource allocation technique to give Ka-band cognitive uplink access to the FSS terminals reusing frequency bands of FS terrestrial microwave links with priority protection. It has been shown in BRIFIC FS database [7] that only around 60% of the FS Ka spectrum band (27.5 to 29.5 GHz) is occupied in terrestrial transmission. Hence, the uplink band could benefit from more than 800 MHz of additional bandwidth at any instant.

A. Simulation Parameters

We present link budget and numerical parameters for FSS users in Table I. For the performance analysis setup, the 40% of non-exclusive band (27.5-29.5 GHz) used for terrestrial transmission is divided into 23 sub bands in the interval of 35 MHz and the exclusive band (19.7-20.2 GHz) is assumed to be used by 25 exclusive users. We assume that the exclusive users are randomly distributed over circular area of radius 150 km. The corresponding SNR matrices were obtained considering all the carrier-user arrangements for non-exclusive cases. For exclusive band communication, it is assumed that all the user are provided with same rate with appropriate resource allocation. We consider a geostationary satellite orbit transmission where average distance from satellite to earth is 35786 km.

B. Numerical Results

At the first part of our numerical analysis, we calculated the SNR threshold for a given FSS user to implement spectrum sensing unit at given FS non-exclusive band. The SNR threshold is calculated by changing the channel idle probability and number of user in exclusive band. Figure 3 shows the variation of SNR threshold with respect to channel idle probability. All the other parameters of interest are kept fixed. The variation of SNR threshold with respect to number of user in exclusive

band is shown in figure 4. The SNR threshold decreases exponentially with increase in the source idle probability as well as increase in the number of FSS users in exclusive band. Also it is expected that for fixed parameters increase in the detection probability or decrease in false alarm probability decreases the SNR threshold.

In addition, the performance of different resource allocation algorithm is presented in figure 5. The rate achieved using Hungarian algorithm is maximum since is non adaptive to the change of rate in exclusive band. The reverse greedy algorithm outperforms greedy algorithm, although small gain is achieved at initial iterations. A cognitive SatCom network can either apply adaptive algorithm or static Hungarian algorithm depending on system design. But adaptive algorithm is robust to change in variation of availability of non-exclusive band.

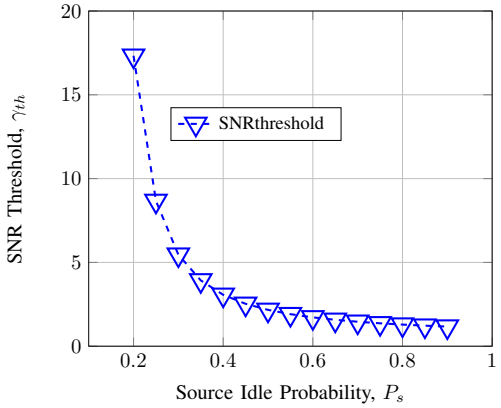


Fig. 2: SNR threshold for different source idle probability.

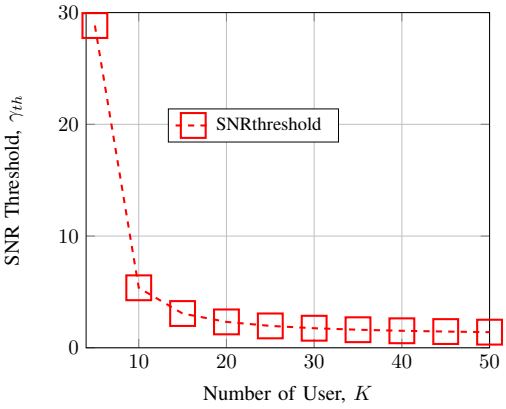


Fig. 3: SNR threshold for differentNumber of Cognitive user.

VI. Conclusion

In this paper, we analyze the performance of a cognitive FSS system in FS terrestrial Ka-band, where terrestrial FS is the incumbent user and satellite uplink FSS is the cognitive user. At first a condition in terms of SNR threshold is formulated to study practicability of implementing spectrum sensing unit for a given FSS user in certain non-exclusive FS Ka band. After

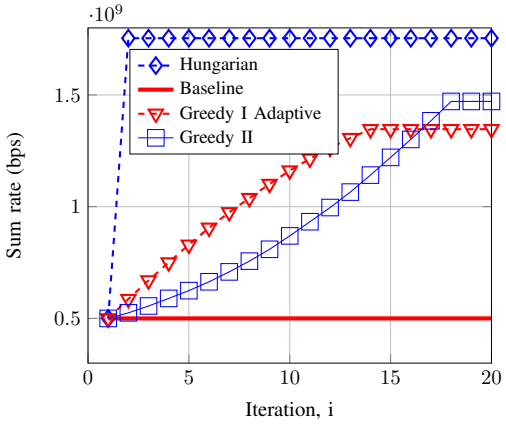


Fig. 4: Sum Rate Maximization in Cognitive Radio.

that two iterative resource allocation algorithm is developed to maximize the achievable throughput of cognitive FSS network. The algorithms we developed are iterative algorithms, which are adaptive to the variation in the change in rate in exclusive band. It is shown that, the optimal iterative algorithm achieves maximum throughput than greedy algorithm.

Acknowledgment

The authors would like to thank SES for the financial support of this work and BELSPO for financing the BESTCOM project in the framework of IAP program.

References

[1] Vidal, O.; Verelst, G.; Lacan, J.; Alberty, E.; Radzik, J.; Bousquet, M., "Next generation High Throughput Satellite system," in Satellite Telecommunications (ESTEL), 2012 IEEE First AESS European Conference on , vol., no., pp.1-7, 2-5 Oct. 2012

[2] Patil, K.; Prasad, R.; Skouby, K., "A Survey of Worldwide Spectrum Occupancy Measurement Campaigns for Cognitive Radio," in Devices and Communications (ICDeCom), 2011 International Conference on , vol., no., pp.1-5, 24-25 Feb. 2011

[3] Sharma, S.K.; Chatzinotas, S.; Ottersten, B., "Cognitive Radio Techniques for Satellite Communication Systems," in Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th , vol., no., pp.1-5, 2-5 Sept. 2013

[4] COgnitive RAdio for SATellite Communications - CoRaSat, European Commission FP7, Oct, 2012.

[5] Kourogiorgas, Charilaos; Panagopoulos, Athanasios D.; Liolis, Konstantinos, "Cognitive uplink FSS and FS links coexistence in Ka-band: Propagation based interference analysis," in Communication Workshop (ICCW), 2015 IEEE International Conference on , vol., no., pp.1675-1680, 8-12 June 2015.

[6] Bogale, T.E.; Vandendorpe, L., "Max-Min SNR Signal Energy Based Spectrum Sensing Algorithms for Cognitive Radio Networks with Noise Variance Uncertainty," in Wireless Communications, IEEE Transactions on , vol.13, no.1, pp.280-290, January 2014.

[7] ITU, "Brific for terrestrial services"," http://www.itu.int/en/ITU-R/terrestrial/brific", [Online; accessed 7-Sept-2015].

Low-Complexity Laser Phase Noise Compensation for Filter Bank Multicarrier Offset-QAM Optical Fiber Systems

T.-H. Nguyen, S.-P. Gorza, F. Horlin
Université libre de Bruxelles
OPERA department
1050 Brussels, Belgium
trung-hien.nguyen@ulb.ac.be

J. Louveaux
Université catholique de Louvain
ICTEAM institute
1348 Louvain-la-Neuve, Belgium
jerome.louveaux@uclouvain.be

Abstract

We report on a low-complexity modified blind phase search (BPS) for the carrier phase estimation in optical filter bank multicarrier offset quadrature modulation format (FBMC-OQAM) systems. More particularly, the distance calculations in the complex plane in state-of-the-art BPS method are replaced by simple multiplication-free operations in the real plane. Moreover, a modified phase searching strategy is also applied, leading to further reduction of the computational complexity. The proposed methods are numerically validated in a 5-subcarriers FBMC-16OQAM system. Similar to conventional BPS method, the tolerated normalized linewidth (the product of laser linewidth and symbol duration) of the proposed BPS method is 10^{-4} for a 1 dB SNR penalty of BER = 10^{-3} with about twice the reduction of computational complexity.

1 Introduction

Filter bank multicarrier offset quadrature modulation format (FBMC-OQAM) is an interesting alternative to orthogonal frequency-division multiplexing (OFDM) and has recently received much attention in optical fiber communication systems [1, 2, 3] thanks to its optimized spectral efficiency (SE) and its robustness to the channel variation in optical fibers. On each subcarrier of FBMC-OQAM systems, a half-symbol time delay between the in-phase and quadrature components of the signal is introduced, in order to achieve the orthogonality between adjacent subcarriers [4]. However, in the presence of the laser phase noise, inter-carrier interference appears between the OQAM subcarriers in addition to the corresponding rotation [5]. As a consequence, the frequently-used carrier phase estimation (CPE) methods for QAM constellations are no longer suitable, unless the laser linewidth is of several kHz [4]. Several CPE solutions for OQAM signals have been proposed so far. The pilot used in [4] can be exploited to the CPE, however, suffering from the reduced SE. An innovative phase tracking combined with the equalizer has been reported in [1], at the cost of extra feedback loop delay. Recently, the modified blind phase search (M-BPS) has recently been proposed in [5], taking the advantage of the feedforward operation [6]. However, the computation effort (CE) is increased associated with the increase of the OQAM modulation level.

In this paper, a modified BPS of lower complexity (LC-BPS) is proposed for FBMC-OQAM systems, in which the distance calculations in the complex plane in M-BPS method is replaced by simple multiplication-free operations in the real plane. Moreover, a modified phase searching strategy is also applied, leading to further reduction of the CE. The proposed methods are numerically validated in a 5-subcarriers FBMC-16OQAM system. Similar to M-BPS method, the tolerated normalized linewidth (the product of laser linewidth and symbol duration) of the LC-BPS method is 10^{-4} for a 1 dB SNR penalty of BER = 10^{-3} with about twice the reduction of CE.

2 Low-Complexity Feedforward Carrier Phase Estimation

The idea behind the modified BPS for OQAM signals is to pre-rotate the received samples with a number of phase tests before removing the half-symbol delay between the in-phase and quadrature components of the received signals [5]. Fig. 1 presents a block diagram of the modified BPS algorithm. As the conventional BPS, the input samples (before applying the standard OQAM equalization) at twice symbol-rate, r, are firstly rotated by test phase value $\phi_b = (b/B) \cdot \pi - \pi/2$, where $b = 1, 2, ..., B$ and B is the total number of the phase tests.

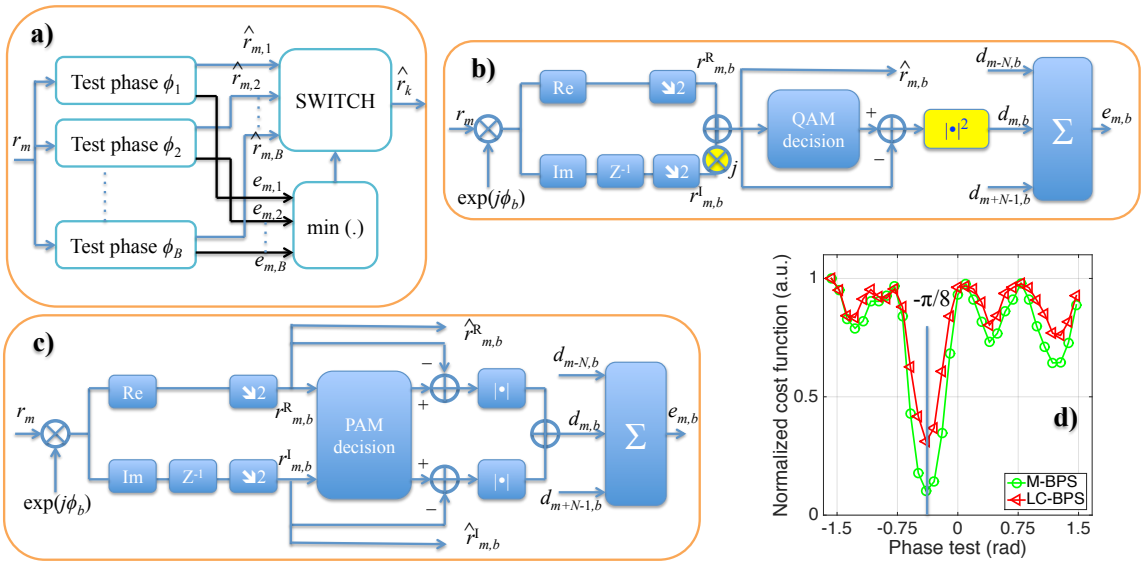


Figure 1: a) Block diagram of modified blind phase search (M-BPS) for OQAM signals. Test phase blocks for b) M-BPS and c) low-complexity LC-BPS without multiplier operator in the cost function calculation. d) Cost functions of M-BPS and LC-BPS for a phase noise value of $\pi/8$.

The rotated m-th sample of the k-th subcarrier can be represented by $r_{k,m,b} = r_{k,m} \cdot \exp(j \cdot \phi_b)$. To simplify the notation, the subcarrier index is omitted later. The rotated samples version (Fig. 1(a)) corresponding to each phase test is then sent to a switch, in order to select the best phase rotation to alleviate the phase noise impact. The rotated phase is chosen if it minimizes the cost function, $d_{m,b}$. The impact of the additive noise is reduced by averaging the cost function over $2N$ consecutive samples with the same phase test, $e_{k,b} = \sum_{n=-N}^{N-1} d_{k+n,b}$. Fig. 1(b) and (c) show the operations inside each test phase block for M-BPS and LC-BPS, respectively. The expected phase rotation with respect to the phase noise suppression can be expressed by

$$\begin{aligned} \hat{\phi}_{b,\mathrm{M-BPS}} &= \arg\min_{\phi_b} e_{k,b,\mathrm{M-BPS}} \\ &= \arg\min_{\phi_b} \textstyle\sum_{n=-N}^{N-1} \left(|r_{k+n,b} - DD\left(r_{k+n,b}\right)|^2 \right), \quad \text{for} \quad \mathrm{M-BPS}, \end{aligned} \tag{1}$$

and

Table 1: Computational complexity for M-BPS, MF-BPS and SLC-BPS

	Real multiplication	Real addition	Decision
M-BPS	$4NB$	$6NB$	$4NB$
LC-BPS	–	$6NB$	$4NB$
SLC-BPS	–	$3NB + 4N$	$2NB + 4N$

N - half of the summation block length; B - number of phase test values.

$$\begin{aligned}\hat{\phi}_{b,\mathrm{LC-BPS}} &= \arg\min_{\phi_b} e_{k,b,\mathrm{LC-BPS}} \\ &= \arg\min_{\phi_b} \sum_{n=-N}^{N-1} \sum_{l=1}^{2} \left(\left| r^l_{k+n,b} - DD\left(r^l_{k+n,b}\right) \right| \right), \quad \text{for} \quad \mathrm{LC-BPS},\end{aligned} \tag{2}$$

in which DD is the direct decision operator. $r^1_{m,b} = r^R_{m,b}$ and $r^2_{m,b} = r^I_{m,b}$ are the real and imaginary parts of the samples, respectively. Note that, the distance calculations in the complex plane in the M-BPS require 2 real-multiplications and 3 real-additions. It is clearly seen that the 2 real-multiplications are removed in the proposed LC-BPS, resulting in the reduced CE compared to the traditional M-BPS.
Fig. 1(d) shows the examples of cost functions for M-BPS and LC-BPS method at the phase noise value of $\pi/8$. It can be observed that the 2 cost functions exhibit the same minimum value at about $-\pi/8$. A new searching-strategy, referred to as SLC-BPS, can be used to reduce further the CE. More particularly, the coarse phase search (CPS) is firstly carried out with half of the required phase test numbers, the fine phase search (FPS) will compare the minimum cost function value (with respect to $\hat{\phi}_b$, found by CPS) to the two adjacent cost function values at the two adjacent phases (i.e. $\hat{\phi}_b \pm \pi/28$ explained later). As a result, the required total number of phase tests is $(B/2 + 2)$, or equivalently the total number of operators is reduced to $B/(B/2 + 2)$ times, approximated 2 times when B is large. Further CE reduction is feasible by cascading several CPS stages. Because the cost function exhibits several local minima (Fig. 1(d)), the number of cascaded CPS stages should be optimized but the detail investigation is out of the scope of this paper. Table 1 summarizes the required operations for different CPE methods in a block length, $2N$, with B phase test values. For a fair comparison, only the operations required for the cost function calculation are taken into account.

3 Results and Discussion

The performance of the proposed CPE methods is numerically studied with 5-subcarriers FBMC-16OQAM optical coherent systems. To focus on the laser phase noise impact, only one polarization is studied and other impairments (i.e. chromatic dispersion, frequency offset) are assumed to be completely compensated. Fig. 2 presents the block diagram used for the simulations. The 16OQAM signals on each subcarrier are generated by differential encoding, mapping the random binary sequences on the QAM constellation and staggering a half of symbol time between the in-phase and quadrature components. These signals are pulse-shaped with square-root raised cosine (SRRC, roll-off factor of 1) filters, $u(t)$, before imprinting onto the subcarriers. Note that, a $\pi/2$ phase difference is introduced between adjacent subcarriers to ensure the orthogonality. All the subcarriers are then summed up together to form a total 130 000 FBMC-16OQAM symbols. These symbols are then corrupted by additive Gaussian white noise (AWGN), $\gamma(t)$, and loaded with the laser phase noise, $\varphi(t)$. The phase

noise is modeled as a discrete random time walk $\varphi_m = \varphi_{m-1} + \Theta_m$, where Θ is a zero-mean Gaussian random variable with variance of $2\pi \cdot \Delta\nu \cdot T_S$. $\Delta\nu$ and T_S denote the laser linewidth and symbol duration, respectively. At the receiver (see Fig. 2), the subcarrier frequency is firstly removed and then compensated for the $\pi/2$ phase difference between adjacent subcarriers. After that, the signals are passed to the matched SRRC filter. Finally, the output signal samples are used to evaluate the effectiveness of proposed CPE algorithms.

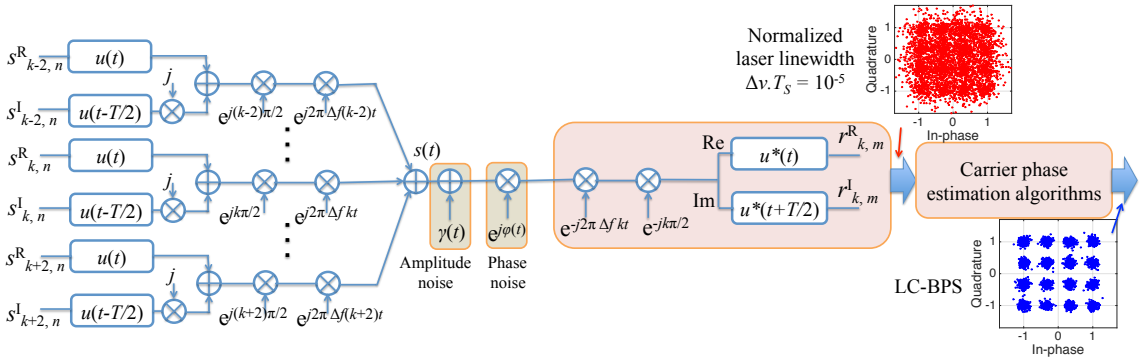


Figure 2: Block diagram of FBMC-16OQAM systems for characterizing the proposed carrier phase estimation on the k-th subcarrier.

Fig. 3(a) shows the calculated bit-error ratio (BER) as a function of the number of phase tests, B, at normalized linewidths, $\Delta\nu \cdot T_S$, of $5 \cdot 10^{-5}$ and $5 \cdot 10^{-6}$. It can be observed that the BER firstly deteriorates and then remains constant with the increase of B. The optimum number of phase tests for 16OQAM signals are 28 and are not affected by changing the normalized linewidth. For this reason, we use two adjacent phase test values of the minimum deduced phase $\hat{\phi}_b$ (in SLC-BPS) being $\hat{\phi}_b \pm \pi/28$. The impact of the summation block length, $2L$, is further studied in Fig. 3(b) for two normalized linewidths of $5 \cdot 10^{-5}$ and $5 \cdot 10^{-6}$. It can be seen that the minimum block length (defined as the shortest block length enabling to reach the target BER) for 16OQAM signals are 20 regardless of the laser linewidth.
In the next step, the evolution of BER at different SNRs is shown in Fig. 3(c) for different methods at the normalized linewidths of $5 \cdot 10^{-4}$ and $5 \cdot 10^{-5}$. The calculated BER decreases when increasing the SNR, all three methods have the same performance. However, at 10^{-2} BER, the algorithm performance for the high-normalized linewidth ($5 \cdot 10^{-4}$) case exhibits a 1.8 dB SNR penalty compared to that of the low-normalized linewidth ($5 \cdot 10^{-5}$) case. Note that, compared to the absence of the phase noise, there always exists a 1.2 dB SNR penalty after the CPE inherent to the traditional BPS algorithm [6]. The tolerated normalized linewidth of the algorithms at 12.2 dB SNR (which is 1 dB above the SNR reaching the target BER $= 10^{-3}$) is further investigated in Fig. 3(d). The BER remains constant as long as the normalized linewidth is smaller than $5 \cdot 10^{-5}$ and starts to increase for higher linewidths. Once again, the evolutions of all three algorithms are superimposed, confirming the effectiveness of the proposed low-complexity CPEs. It is clearly seen that the CPE algorithms for 16OQAM signals can tolerate the normalized linewidth of 10^{-4}, while our proposed methods reduce much computational effort, compared to the traditional M-BPS method.

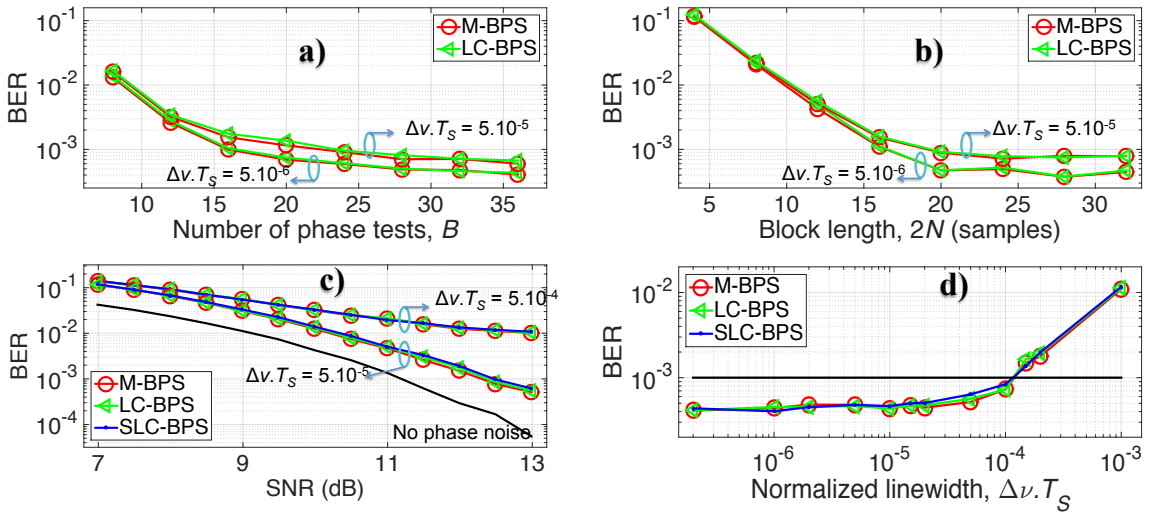


Figure 3: a) Calculated BER versus number of phase tests and b) calculated BER versus block length for two normalized linewidths of $5 \cdot 10^{-5}$ and $5 \cdot 10^{-6}$. c) BER as a function of SNR for two normalized linewidths of $5 \cdot 10^{-4}$ and $5 \cdot 10^{-5}$. d) BER versus normalized linewidth.

4 Conclusion

A low-complexity modified blind phase search (BPS) algorithm for OQAM signals has been proposed and numerically validated with 5-subcarriers FBMC-16OQAM systems. The proposed method provides a tolerated normalized linewidth of 10^{-4} for a 1 dB SNR penalty at BER = 10^{-3}, as the BPS with reduced computational effort.

5 Acknowledgment

This work is supported by the Belgian Fonds National de la Recherche Scientifique FNRS (PDR T.1039.15).

References

[1] J. Fickers, *et al.*, "Multicarrier offset-QAM for long-haul coherent optical communications," *J. Lightw. Technol.*, vol. 32, no. 24, pp. 4069-4076, Dec. 2014.

[2] S. Randel, *et al.*, "Generation of 224-Gb/s multicarrier offset-QAM using a real time transmitter," in *Proc. OFC 2012*, p. OM2H.2, Los Angeles, CA, USA, Mar. 2012.

[3] F. Horlin, *et al.*, "Dual-polarization OFDM-OQAM for communications over optical fibers with coherent detection," *Opt. Express*, vol. 21, no. 5, pp. 6409-6421, Mar. 2013.

[4] J. Zhao, *et al.*, "DFT-based offset-QAM OFDM for optical communications," *Opt. Express*, vol. 22, no. 1, pp. 1114-1126, Jan. 2014.

[5] H. Tang, *et al.*, "Feed-forward carrier phase recovery for offset-QAM Nyquist WDM transmission," *Opt. Express*, vol. 23, no. 5, pp. 6215-6227, Mar. 2015.

[6] T. Pfau, *et al.*, "Hardware-efficient coherent digital receiver concept with feedforward carrier recovery for M-QAM constellation," *J. Lightw. Technol.*, vol. 27, no. 8, pp. 989-999, Apr. 2009.

8-state unclonable encryption

Boris Škorić, TU Eindhoven (b.skoric@tue.nl)

Unclonable Encryption quantum-protects classical ciphertext so that it cannot be copied. We propose an improved variant which has better efficiency and noise tolerance. Our variant uses four cipherstate bases that are equally spaced on the Bloch sphere, instead of the usual + and × basis.

Unclonable Encryption (UE) was introduced by D. Gottesman [1] in 2003 and has since been largely ignored. There is a quantum channel from Alice to Bob, but no (cheap) channel from Bob to Alice. This scenario is relevant for instance when a message is sent into the future, or in the case of significant time lags in long-distance communication, or if multiple-round protocols are too costly. The aim of UE is to send a classical ciphertext to Bob, quantum-protected in such a way that one of the following two outcomes occurs: Either (i) Bob successfully recovers the plaintext and verifies its authenticity. Eve learns nothing about the ciphertext. Or (ii) Eve learns some of the ciphertext and Bob is not able to recover&verify the plaintext, i.e. the attack has been noticed.

Gottesman identified the implication chain *quantum authentication* $\Longrightarrow$ *UE* $\Longrightarrow$ *Quantum Key Distribution*. He gave a UE construction based on classical error correction (ECC) and single-qubit operations. Encryption consists of the following steps: append a Message Authentication Code (MAC) to the plaintext; encode using the ECC; encrypt with a One-Time Pad (OTP); prepare qubit states using Conjugate Coding (CC) in the +× bases. The scheme requires a MAC key, an OTP and a key (one bit per qubit indicating the basis choice) for the CC. Eve cannot copy the qubit states without knowing the CC key. If Bob detects no tampering, the MAC key and the CC key can be safely re-used.

Gottesman's scheme has the drawback that the CC allows Eve to obtain partial information about the qubit states. This necessitates Privacy Amplification (compression) and limits the amount of noise that can be tolerated on the quantum channel to 13.7% bit error rate. We replace the CC by 8-state encryption (Fig. 1), which works with 4 basis options and completely hides the qubit state from Eve. As a consequence, privacy amplification is not required, and any noise can be tolerated that is correctable by a classical ECC.
This result has potential implications for low-interaction QKD variants and commitment schemes that need unclonability.

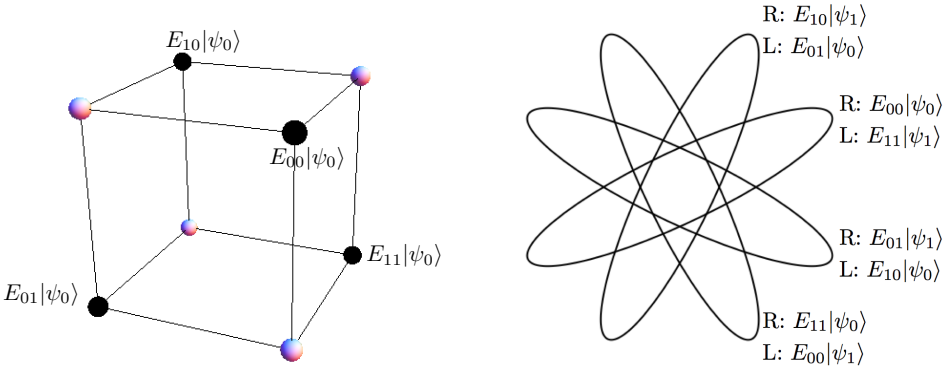


Figure 1: *The eight cipherstates* $|\psi_{uwg}\rangle = E_{uw}|\psi_g\rangle$ *shown (left) on the Bloch sphere, as corner points* $(\pm 1, \pm 1, \pm 1)/\sqrt{3}$ *of a cube; and (right) as elliptic polarisation states. 'R'=right, 'L'=left.*

References

[1] D. Gottesman. Uncloneable encryption. *Quantum Information and Computation*, 3(6):581–602, 2003.

[2] B. Škorić. Unclonable encryption revisited ($4 \times 2 = 8$), 2015. eprint.iacr.org/2015/1221.

Optimizing the discretization in Zero Leakage Helper Data Systems

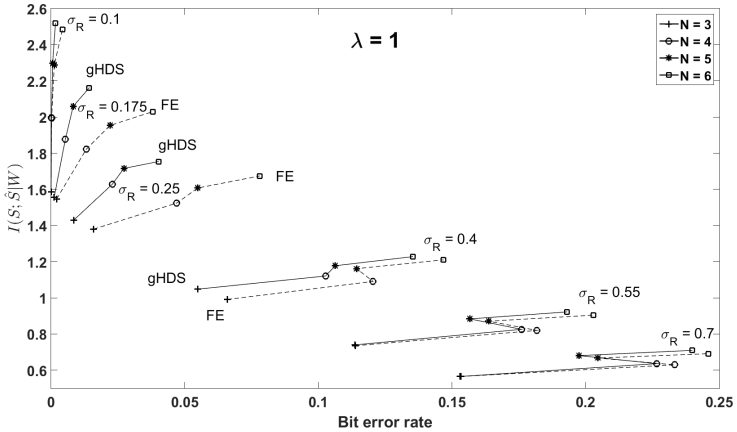
Taras Stanko, Fitria Nur Andini, and Boris Škorić. TU Eindhoven. t.stanko@tue.nl

Helper Data Systems (HDSs) are a cryptographic primitive that allows for the reproducible extraction of secrets from noisy measurements. Redundancy data called Helper Data makes it possible to do error correction while leaking little or nothing ('Zero Leakage') about the extracted secret string. We study the case of non-discrete measurement outcomes. In this case a discretization step is required. Recently de Groot et al. described a generic method to perform the discretization in a Zero Leakage manner. We extend their work and show how the discretization intervals should be set to maximize the amount of extracted secret key material when noise is taken into account.

Verbitskiy et al. [1] introduced a Zero Leakage Fuzzy Extractor (ZLFE) for $X \in \mathbb{R}$. They divided the space $\mathbb{R}$ into N intervals $A_0, \ldots, A_{N-1}$ that are equiprobable in the sense that $\Pr[X \in A_j] = 1/N$ for all j. At enrollment, if X lies in interval A_j then S is set to j. For the helper data they introduced a further division of each interval A_j into m equiprobable subintervals $(A_{jk})_{k=0}^{m-1}$. If the enrollment measurement X lies in interval A_{jk} then the index k is stored as helper data. The fact that all these subintervals are equiprobable leads to independence between the helper data and the secret.

De Groot et al. [2] took the limit $m \to \infty$ and showed that the resulting scheme is not just *a* ZLFE but *the* generic best performing ZLFE for $X \in \mathbb{R}$; other ZLFEs for $X \in \mathbb{R}$ can be derived from the generic scheme. Furthermore, de Groot et al. generalized the scheme of [1] from ZLFEs to general ZLHDSs by allowing intervals $A_0, \ldots, A_{N-1}$ that are not equiprobable. Several questions were left open regarding the `Rec` (reconstruction) procedure in general ZLHDSs and the performance of ZLHDSs compared to ZLFEs.

We derive an optimal `Rec` procedure that minimizes the probability of reconstruction errors. We obtain analytic formulas for Gaussian noise and Lorentz-distributed noise. Using this `Rec` we study the performance of ZLHDSs compared to ZLFEs. We define performance as the mutual information between the secret (S) and the reconstructed secret ($\hat{S}$) conditioned on the fact that the adversary knows the helper data (W). The mutual information $I(S;\hat{S}|W)$ represents the maximum amount of secret key material that can be extracted from X using a ZLHDS. The intricacies of the `Rec` procedure cause the mutual information to become a very complicated function of the choice of quantisation intervals $A_0, \ldots, A_{N-1}$. We have to resort to numerics. Our numerical results for Gaussian source and Gaussian noise show that optimization of the discretization intervals yields an improvement with respect to the ZLFE in terms of mutual information as well as reconstruction error probability. In most cases the gain in $I(S;\hat{S}|W)$ is modest (a few percent), but the reduction of the error rate can be substantial. We conclude that in practice it is better to use a ZLHDS than a ZLFE.



[1] E.A. Verbitskiy, P. Tuyls, C. Obi, B. Schoenmakers, and B. Škorić. Key extraction from general nondiscrete signals. *IEEE Transactions on Information Forensics and Security*, 5(2):269–279, 2010.

[2] J. de Groot, B. Škorić, N. de Vreede, and J.P. Linnartz. Quantization in continuous-source Zero Secrecy Leakage Helper Data Schemes. `https://eprint.iacr.org/2012/566`, 2012.

Zero-Leakage Multiple Key-Binding Scenarios for SRAM-PUF Systems Based on the XOR-Method

Lieneke Kusters Tanya Ignatenko Frans M.J. Willems
Eindhoven University of Technology
Dept. Electrical Engineering, SPS Group
P.O. Box 513, 5600 MB, Eindhoven
c.j.kusters@tue.nl t.ignatenko@tue.nl f.m.j.willems@tue.nl

Abstract

We show that the XOR-method based on linear error-correcting codes can be applied to achieve the secret-key capacity of binary-symmetric SRAM-PUFs. Then we focus on multiple key-bindings. We prove that no information is leaked by all the helper data about a single secret key both in the case where we use the same key all the time and when we use different keys. The notion of symmetry is crucial in these proofs.*

1 Introduction

The power-on state of static random-access memory (SRAM) can be used as a physical unclonable function (PUF)[7][3]. A PUF results from randomness in the production process of a device, and is hard to predict or to copy while at the same time the response is reliable. These properties make PUFs useful for various security systems.

In this paper, we investigate the XOR-method for secret-key binding to an SRAM-PUF. We show that as a result of symmetry in the state of the SRAM cells there is no leakage in multiple key-binding scenario's.

2 Secret-Key Binding Scheme

It is our goal to share a secret key S, assuming values in $\{1, 2, \cdots, |S|\}$, between an encoder and a decoder using a secret-key binding scheme as in Figure 1. Both parties observe the same SRAM-PUF, resulting in binary observation vectors $X^N = (X_1, X_2, \cdots, X_N)$ and $Y^N = (Y_1, Y_2, \cdots, Y_N)$ respectively, corresponding to the start-up values of the N cells of the SRAM. First, the secret key S is generated uniformly at random and presented to the encoder. The encoder then binds this secret key to the observation vector X^N by generating helper data M, hence $M = E(S, X^N)$. Now the decoder can construct an estimate $\hat{S}$ of the secret key using the helper data M and his own observation vector Y^N, hence $\hat{S} = D(M, Y^N)$. Ideally, the helper data M provides sufficient information to correctly reconstruct the secret key, and then $\hat{S} = S$. At the same time M should not leak any information about the secret to a third party, hence $I(S; M) = 0$. We assume that the SRAM-PUF is binary-symmetric, hence

$$\Pr\{(X^N, Y^N) = (x^N, y^N)\} = \prod_{n=1,N} Q(x_n, y_n),$$

*Research carried out in the E! 9629 PATRIOT project, which is funded by the Eurostars-2 joint programme with co-funding from the European Union Horizon 2020 research and innovation programme.

where $Q(0,1) = Q(1,0) = p$ and $Q(0,0) = Q(1,1) = 1-p$, for a parameter p that satisfies $0 \leq p \leq 1/2$. It is our goal to maximize the achievable secret-key rate for our system.

Definition 1 *A secret-key rate R_s with $R_s \geq 0$ is said to be achievable if for all $\epsilon > 0$ and for all N large enough, there exist encoders and decoders such that:*

$$\begin{aligned} \Pr\{\hat{S} \neq S\} &\leq \epsilon, \\ \frac{1}{N}\log_2|S| &\geq R_s - \epsilon, \\ I(S;M) &= 0. \end{aligned}$$

The maximum secret-key rate that is achievable is called the secret-key capacity and is denoted by C_s.

Theorem 1 *The secret-key capacity of a binary-symmetric SRAM-PUF with parameter p is*

$$C_s = 1 - h(p),$$

where the binary entropy function $h(\cdot)$ is defined as $h(\alpha) = \alpha \log_2 \frac{1}{\alpha} + (1-\alpha)\log_2 \frac{1}{1-\alpha}$, for $0 \leq \alpha \leq 1$.

In the next section, we first show that secret-key rates larger than $C_s = 1 - h(p)$ are not achievable (converse). After that, we show that secret-key rate $C_s = 1 - h(p)$ is achievable. We achieve this result by applying the so called code-offset method or the XOR-method [4].

3 Proof of Theorem 1

3.1 Converse result

We show here that there exists no secure and reliable scheme that meets the conditions in Definition 1 when $R_s > 1 - h(p)$. Consider the information in the secret key

$$\begin{aligned} \log_2|S| &= H(S) \\ &= I(S;M,Y^N) + H(S|M,Y^N) \\ &\overset{(a)}{=} I(S;M,Y^N) + H(S|M,Y^N,\hat{S}) \\ &\leq I(S;M,Y^N) + H(S|\hat{S}) \\ &\overset{(b)}{\leq} I(S;M,Y^N) + h(P_e) + P_e \log_2|S| \\ &= I(S;M) + I(S;Y^N|M) + h(P_e) + P_e\log_2|S| \\ &\overset{(c)}{=} 0 + H(Y^N|M) - H(Y^N|M,S) + h(P_e) + P_e\log_2|S| \\ &\leq H(Y^N|M) - H(Y^N|M,S,X^N) + h(P_e) + P_e\log_2|S| \\ &\overset{(d)}{\leq} H(Y^N) - H(Y^N|X^N) + h(P_e) + P_e\log_2|S| \\ &= I(X^N;Y^N) + h(P_e) + P_e\log_2|S| \\ &\overset{(e)}{=} N(1-h(p)) + h(P_e) + P_e\log_2|S|, \end{aligned}$$

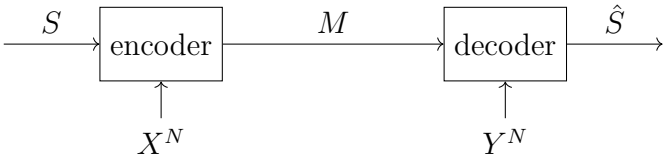


Figure 1: The secret-key binding scheme.

where (a) follows from the fact that $\hat{S} = D(M, Y^N)$, (b) from Fano's inequality where $P_e = \Pr\{\hat{S} \neq S\}$, (c) from the requirement $I(S; M) = 0$, (d) from the Markov relation $(S, M) - X^N - Y^N$, and (e) follows from the fact that all bit-pairs (X_n, Y_n) in the observation vectors are identically distributed according to $Q(\cdot, \cdot)$.

From this derivation we can conclude that

$$\frac{1}{N} \log_2 |S| \leq \frac{1 - h(p) + h(P_e)/N}{(1 - P_e)},$$

and a rate R_s is now achievable if

$$R_s - \epsilon \leq \frac{1}{N} \log_2 |S| \leq \frac{1 - h(p) + h(\epsilon)/N}{(1 - \epsilon)},$$

for all $\epsilon > 0$ and all N large enough. By letting $\epsilon \downarrow 0$ and $N \to \infty$ we can conclude that achievable rates satisfy $R_s \leq 1 - h(p)$ and consequently $C_s \leq 1 - h(p)$.

3.2 Achievability

We have shown that the secret-key capacity C_s of the secret-key binding model cannot exceed $1 - h(p)$. Here we will show that the XOR-method [4], which is based on error-correcting codes, see Figure 2, can be used to achieve capacity $C_s = 1 - h(p)$.

3.2.1 XOR-method and linear coding

The secret key S^K is now assumed to be a K-bit uniformly chosen binary vector. It is mapped to a codeword C^N by an encoder E_L corresponding to a linear error-correcting code, hence $C^N = E_L(S^K)$. Subsequently the helper data M^N, a binary vector of length N, is generated by adding modulo-2 the codeword C^N to the observation vector X^N. Similarly, at the receiver side, a decoder D_L transforms a noisy version codeword $\tilde{C}^N$ into an estimate of the secret key, i.e., $\hat{S}^K = D_L(\tilde{C}^N)$. The noisy version $\tilde{C}^N$ of the codeword results from adding, modulo-2, the helper data M^N to the observation vector Y^N.

Clearly, since $\tilde{C}^N = C^N \oplus X^N \oplus Y^N$, bit-errors occur with probability p. Therefore we can model the behavior of the XOR-scheme as a binary-symmetric channel with

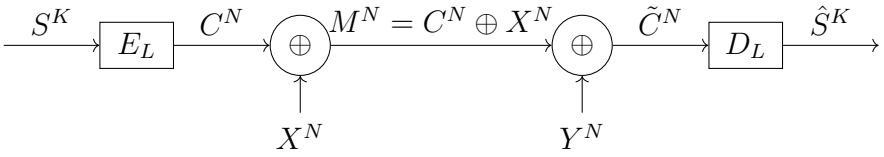


Figure 2: The XOR-method for secret-key binding.

cross-over probability p. We show that for this scheme a linear error-correcting code exists that achieves secret-key capacity $C_s = 1 - h(p)$.

We will use a random coding argument but will here generate a linear code at random, see [2]. We assume that a binary generator matrix $G^{K\times N}$ is generated uniformly at random. Now a codeword corresponding to a secret S^K is constructed by matrix multiplication as follows

$$C^N = S^K G^{K\times N} = \sum_{k=1}^{K} S_k G_k^N,$$

where G_k^N corresponds to the k^{th} row of the matrix and the additions are modulo-2. The receiver decodes a received codeword $\tilde{C}^N$ by evaluating all 2^K possible secret keys and decodes $\hat{S}^K$, if $\hat{S}^K$ is the unique key satisfing

$$d_H(\hat{S}^K G^{K\times N}, \tilde{C}^N) < \delta N, \tag{1}$$

where $d_H(\cdot,\cdot)$ denotes Hamming distance, and δ is appropriately chosen. Note that minimizing the Hamming distance is the best strategy for finding the most probable secret, since the secrets are equally likely and the channel is binary-symmetric.

In the presented linear coding scheme an error can occur in two cases. An error of the first kind occurs when the channel output $\tilde{C}^N$ is not close enough to the channel input C^N, and therefore the Hamming distance in (1) is too large for the correct secret key S^K. The probability of occurrence of this event is

$$\begin{aligned}\Pr\{d_H(S^K G^{K\times N}, \tilde{C}^N) \geq \delta N\} &= \Pr\{d_H(C^N, \tilde{C}^N) \geq \delta N\} \\ &= \Pr\{d_H(X^N, Y^N) \geq \delta N\} \overset{(a)}{\leq} 2^{-Nd(\delta||p)},\end{aligned} \tag{2}$$

where (a) follows from the Chernoff bound (see Appendix A), with $\delta > p$.

An error of the second kind occurs when the channel output $\tilde{C}^N$ is close to possible channel input $\check{C}^N = \check{S}^K G^{K\times N}$, for some $\check{S}^K \neq S^K$. The probability that this occurs is

$$\begin{aligned}&\Pr\{d_H(\check{C}^N, \tilde{C}^N) < \delta N\} \\ &\qquad = \Pr\{d_H(\check{S}^K G^{K\times N}, S^K G^{K\times N} \oplus X^N \oplus Y^N) < \delta N\} \overset{(b)}{\leq} 2^{-Nd(\delta||1/2)},\end{aligned}$$

where (b) follows from the fact that $\check{C}^N$ and $\tilde{C}^N$ are independent and Chernoff's bound (see Appendix A) can be applied, with $\delta < 1/2$. Furthermore, the independence of $\check{C}^N$ and $\tilde{C}^N$ follows from the fact that $\check{S}^K$ and S^K differ in it least one bit, and therefore the resulting codewords $\check{C}^N$ and C^N differ at least by one randomly generated matrix row and are thus independent.

There are $2^K - 1$ messages $\check{S}^K \neq S^K$, and therefore the probability of occurrence of an error of the second kind, using the union bound, is bounded as

$$\sum_{\check{S}^K \neq S^K} \Pr\{d_H(\check{C}^N, \tilde{C}^N) < \delta N\} \leq (2^K - 1)2^{-Nd(\delta||1/2)} \leq 2^{-N(1-h(\delta)-K/N)}, \tag{3}$$

It follows from the equations (2) and (3) that the error probability converges to zero for large N, as long as $K/N < 1 - h(\delta)$ and $p < \delta < 1/2$. Therefore, K/N can get arbitrarily close to $C_s = 1 - h(p)$ for $\delta \to p$.

The above results show, that averaged over the random linear code, the error probability vanishes for $N \to \infty$ for all $K/N = \frac{1}{N}\log 2^K$ arbitrarily close to C_s. Therefore, we can conclude that there are linear codes that achieve the capacity C_s, although we should still verify that $I(M^N; S^K) = 0$.

3.2.2 Information leakage by the helper data

We use the XOR-scheme (Figure 2) to bind a secret key S^K to an SRAM-PUF by creating helper data M^N, based on the observation vector X^N. In order for the scheme to be secure, we require that the (public) helper data M^N by itself does not leak any information about the secret key S^K. We can easily show that this is true for one secret S^K and helper data M^N:

$$\begin{aligned} I(M^N; S^K) &\overset{(a)}{\leq} I(M^N; C^N) \\ &= I(C^N \oplus X^N; C^N) \\ &= H(C^N \oplus X^N) - H(C^N \oplus X^N | C^N) \\ &\leq N - H(X^N) = N - N = 0, \end{aligned}$$

where (a) follows from the data-processing inequality, see Cover and Thomas [1], p. 35. This concludes the achievability part of the proof of Theorem 1.

4 Multiple SRAM-PUF Observations

4.1 Introduction

So far we have considered a single run of the key-binding scheme only, i.e. a binding step in which a key is bound to an SRAM observation followed by a reconstruction of the secret from a second observation of the SRAM and the public helper data.

However, it is not so obvious whether information leakage occurs when the same SRAM-PUF is used to bind secret keys multiple times. In this case the encoder constructs T helper data sequences $\left\{M_1^N, M_2^N, \cdots, M_T^N\right\}$ by binding each secret key to another observation of the SRAM-PUF $\left\{X_1^N, X_2^N, \cdots, X_T^N\right\}$. In this section we prove that even if all helper data is observed, no information is leaked about any single secret.

However, before we can investigate the information leakage resulting from multiple key-binding using the XOR-scheme, we first analyze the relation between multiple observations of the same SRAM-PUF.

4.2 SRAM model

An SRAM is composed of N cells of which the start-up values can be observed multiple times. We assume that all observations of cell n are i.i.d. where the probability of a 1 being observed is given by the cell's state θ_n. The states $\theta^N = (\theta_1, \theta_2, \cdots, \theta_N)$ of the cells are i.i.d. given by a random variable Θ_n which has probability density function (see e.g. [6], [5]):

$$p_{\Theta_n}(\theta) = p_\Theta(\theta), \quad 0 \leq \theta \leq 1. \tag{4}$$

In the following, we analyze multiple observations of a single cell with index $n \in \{1, 2, \cdots, N\}$. For this purpose we define the cell-observation vector $\mathtt{X}^T = (\mathtt{X}_1, \mathtt{X}_2 \cdots \mathtt{X}_T)$, with $\mathtt{X}_t$ the t^{th} observation of the n^{th} SRAM cell. Now, we can start our analysis. First, we use (4) and the fact that the observations are i.i.d. to find that for some cell-observation vector $\mathtt{X}^T$

$$\Pr\{\mathtt{X}^T = \mathtt{x}^T\} = \int_0^1 (1-\theta)^{T - w_H(\mathtt{x}^T)} \theta^{w_H(\mathtt{x}^T)} p_\Theta(\theta) d\theta,$$

where $w_H(\mathtt{x}^T)$ is the Hamming weight, or the number of ones in vector $\mathtt{x}^T$. Note that the probability of the vector $\mathtt{x}^T$ only depends on its Hamming weight $w_H(\mathtt{x}^T)$. Therefore, there are $T+1$ elementary probabilities

$$\pi(w) \triangleq \int_0^1 (1-\theta)^{T-w}\theta^w p_\Theta(\theta)d\theta,$$

for $w = 0, 1, \cdots, T$, that determine the cell vector probabilities

$$\Pr\{\mathtt{X}^T = \mathtt{x}^T\} = \pi(w_H(\mathtt{x}^T)).$$

4.3 Symmetric SRAM model

Next, we assume that for SRAM-PUFs the density $p_\Theta(\theta)$ of the state θ is symmetric, that is

$$p_\Theta(\theta) = p_\Theta(1-\theta), \quad 0 \le \theta \le 1.$$

It turns out that the symmetry of the probability density function of the state of the SRAM-PUF cells is an essential feature needed in our later proofs. As a result the elementary probabilities also show symmetric behavior:

$$\begin{aligned}\pi(T-w) &= \int_0^1 (1-\theta)^w\theta^{T-w}p_\Theta(\theta)d\theta = \int_0^1 (1-\theta)^w\theta^{T-w}p_\Theta(1-\theta)d\theta \\ &= \int_0^1 (\theta')^w(1-\theta')^{T-w}p_\Theta(\theta')d\theta' = \pi(w),\end{aligned}$$

for $w = 0, 1, \cdots, T$.

Lemma 1 *A consequence of the symmetry is that for the probability of the cell vector*

$$\begin{aligned}\Pr\{\mathtt{X}^T = \mathtt{x}_1, \cdots, \mathtt{x}_T\} &= \pi(w_H(\mathtt{x}_1, \cdots, \mathtt{x}_T)) \\ &= \pi(T - w_H(\mathtt{x}_1, \cdots, \mathtt{x}_T)) \\ &= \pi(w_H(\overline{\mathtt{x}_1}, \cdots, \overline{\mathtt{x}_T})) = \Pr\{\mathtt{X}^T = \overline{\mathtt{x}_1}, \cdots, \overline{\mathtt{x}_T}\},\end{aligned}$$

where $\overline{\mathtt{x}}$ *corresponds to the inverse of a bit* $\mathtt{x}$*.*

From this lemma we can conclude for any $t \in \{1, 2, \cdots, T\}$ and $\mathtt{x}_1, \mathtt{x}_2, \cdots, \mathtt{x}_T$

$$\begin{aligned}\Pr\{\mathtt{X}^T = \mathtt{x}_1, \cdots, \mathtt{x}_T\} &= \pi(w_H(\mathtt{x}_1, \cdots, \mathtt{x}_T)) \\ &= \pi(w_H(\mathtt{x}_1 \oplus \mathtt{x}_t, \cdots, \mathtt{x}_t \oplus \mathtt{x}_t, \cdots, \mathtt{x}_T \oplus \mathtt{x}_t)) \\ &= \Pr\{\mathtt{X}_1 = \mathtt{x}_1 \oplus \mathtt{x}_t, \cdots, \mathtt{X}_t = 0, \cdots, \mathtt{X}_T = \mathtt{x}_T \oplus \mathtt{x}_t)\}.\end{aligned}$$

Until now, we have looked at T observations $\mathtt{X}^T$ of a single cell with index $n \in \{1, 2, \cdots, N\}$. However, recall that we are interested in the relation between two or more observation vectors, that each correspond to an observation of all the N cells in the SRAM-PUF. Therefore, we now define T observation vectors $\{X_1^N, X_2^N, \cdots X_T^N\}$, each corresponding to an observation of all N SRAM cells, and we can state the following corollary.

Corollary 1 *For all* $x_1^N, x_2^N, \cdots, x_T^N$ *we have that for any* $t \in \{1, 2, \cdots, T\}$

$$\begin{aligned}&\Pr\{X_1^N = x_1^N, \cdots, X_t^N = x_t^N, \cdots, X_T^N = x_T^N\} \\ &\qquad = \Pr\{X_1^N = x_1^N \oplus x_t^N, \cdots, X_t^N = 0^N, \cdots, X_T^N = x_T^N \oplus x_t^N\}.\end{aligned}$$

4.4 Information leakage for multiple key-binding

We use the results from Corollary 1 to prove that no information leakage is resulting from observing multiple helper data corresponding to multiple key-binding with the same SRAM-PUF. For key binding of the t^{th} key S_t^K with $t \in \{1, 2, \cdots, T\}$, helper data M_t^N is generated from the codeword C_t^N and SRAM observation X_t^N, according to $M_t^N = C_t^N \oplus X_t^N$, see also Figure 2.

Different Keys: First, we look at multiple key-binding with different secret keys S_t^K and thus different codewords $C_t^N = E_L(S_t^K)$, for which we find that for any of the codewords C_t^N, with $t \in \{1, \cdots T\}$

$$
\begin{aligned}
&\Pr\{M_1^N = m_1^N, \cdots, M_T^N = m_T^N | C_t^N = c_t^N\} \\
&= \sum_{c_1^N \in \mathcal{C}^N} \cdots \sum_{c_{t-1}^N \in \mathcal{C}^N} \sum_{c_{t+1}^N \in \mathcal{C}^N} \cdots \sum_{c_T^N \in \mathcal{C}^N} \frac{1}{|\mathcal{C}^N|^{T-1}} \Pr\{X_1^N = m_1^N \oplus c_1^N, \cdots, X_T^N = m_T^N \oplus c_T^N\} \\
&\overset{(a)}{=} \sum_{c_1^N \in \mathcal{C}^N} \cdots \sum_{c_{t-1}^N \in \mathcal{C}^N} \sum_{c_{t+1}^N \in \mathcal{C}^N} \cdots \sum_{c_T^N \in \mathcal{C}^N} \frac{1}{|\mathcal{C}^N|^{T-1}} \Pr\{X_1^N = m_1^N \oplus m_t^N \oplus c_1^N \oplus c_t^N, \cdots \\
&\qquad\qquad \cdots, X_t^N = 0^N, \cdots, X_T^N = m_T^N \oplus m_t^N \oplus c_T^N \oplus c_t^N\} \\
&= \sum_{c_1'^N \in \mathcal{C}^N} \cdots \sum_{c_{t-1}'^N \in \mathcal{C}^N} \sum_{c_{t+1}'^N \in \mathcal{C}^N} \cdots \sum_{c_T'^N \in \mathcal{C}^N} \frac{1}{|\mathcal{C}^N|^{T-1}} \Pr\{X_1^N = m_1^N \oplus m_t^N \oplus c_1'^N, \cdots \\
&\qquad\qquad \cdots, X_t^N = 0^N, \cdots, X_T^N = m_T^N \oplus m_t^N \oplus c_T'^N\} \\
&= f(m_1^N, \cdots, m_T^N),
\end{aligned}
$$

where $f(\cdot)$ is some function that does not depend on c_t^N. Moreover in (a) we used Corollary 1. We can now conclude that

$$
\begin{aligned}
&\Pr\{M_1^N = m_1^N, \cdots, M_T^N = m_T^N\} \\
&\qquad = \sum_{c^N \in \mathcal{C}^N} \Pr\{M_1^N = m_1^N, \cdots, M_T^N = m_T^N | C_t^N = c^N\} \Pr\{C_t^N = c^N\} \\
&\qquad = \sum_{c^N \in \mathcal{C}^N} f(m_1^N, \cdots, m_T^N) \Pr\{C_t^N = c^N\} \\
&\qquad = f(m_1^N, \cdots, m_T^N) = \Pr\{M_1^N = m_1^N, \cdots, M_T^N = m_T^N | C_t^N = c_t^N\},
\end{aligned}
$$

for any c_t^N and any $t \in \{1, 2, \cdots T\}$. Now $I(C_t^N; M_1^N, \cdots, M_T^N) = H(M_1^N, \cdots, M_T^N) - H(M_1^N, \cdots, M_T^N | C_t^N) = 0$. By the data-processing inequality

$$I(S_t^K; M_1^N, \cdots, M_T^N) \leq I(C_t^N; M_1^N, \cdots, M_T^N) = 0,$$

and thus observation of the helper data does not leak any information about any of the secrets S_t^K, with $t \in \{1, 2, \cdots T\}$.

Single Key: Similarly for multiple bindings with a single secret-key S^K and thus a single codeword C^N, we can show that

$$
\begin{aligned}
&\Pr\{M_1^N = m_1^N, \cdots, M_T^N = m_T^N | C^N = c^N\} \\
&= \Pr\{X_1^N = m_1^N \oplus c^N, \cdots, X_T^N = m_T^N \oplus c^N\} \\
&\overset{(a)}{=} \Pr\{X_1^N = m_1^N \oplus m_t^N, \cdots, X_t^N = 0^N, \cdots, X_T^N = m_T^N \oplus m_t^N\} = g(m_1^N, \cdots, m_T^N),
\end{aligned}
$$

where $g(\cdot)$ is some function that does not depend on c^N, and (a) follows from Corollary 1. Following similar reasoning as used for different keys, we can conclude that $I(S^K; M_1^N, \cdots, M_T^N) = 0$, and thus observation of the helper data does not leak any information about the secret S^K.

5 Conclusion

We have shown that the secret key S^K can be shared, using the XOR-scheme shown in Figure 2 and linear coding, with a rate R_s arbitrary close to $1 - h(p)$, where p is the cross-over probability between two observations of the same SRAM-PUF cell. Furthermore, we have shown that the presented scheme is secure in the sense that no information is leaked by the helper data M^N about the secret key S^K, even if the same scheme (and SRAM-PUF) is used multiple times to bind either the same or different secret keys. This result follows from the key-assumption that the probability density function of the states of the SRAM cells is symmetric.

A Chernoff bound

Let $U_1, \cdots, U_N$ be independent random variables with $U_i \in \{0, 1\}$ and $\Pr\{U_i = 1\} = p$, for $i = 1, \cdots, N$. Then, for any $a \in (p, 1]$ and $b \in [0, p)$, we have

$$\Pr\left\{\sum_{i=1}^{N} U_i \geq aN\right\} \leq 2^{-Nd(a||p)}, \text{ and } \Pr\left\{\sum_{i=1}^{N} U_i \leq bN\right\} \leq 2^{-Nd(b||p)},$$

with $d(x||y) \triangleq x \log_2 \frac{x}{y} + (1 - x) \log_2 \frac{1-x}{1-y}$.

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* John Wiley & Sons, 2nd edition, 2006.

[2] Peter Elias. Coding for noisy channels. *IRE Conv. Rec.*, Pt. 4:37–46, 1955.

[3] Daniel E. Holcomb, Wayne P. Burleson, and Kevin Fu. Power-Up SRAM state as an identifying fingerprint and source of true random numbers. *IEEE Trans. Comput.*, 58(9):1198–1210, 2009.

[4] Ari Juels and Martin Wattenberg. A fuzzy commitment scheme. In *6th ACM Conf. Comput. Commun. Secur. - CCS '99*, pages 28–36, New York, New York, USA, 1999. ACM Press.

[5] Roel Maes. An Accurate Probabilistic Reliability Model for Silicon PUFs. In *Cryptogr. Hardw. Embed. Syst. - CHES 2013 15th Int. Work. St. Barbar. CA, USA*, pages 73–89. 2013.

[6] Roel Maes, Pim Tuyls, and Ingrid Verbauwhede. A soft decision helper data algorithm for SRAM PUFs. In *IEEE Int. Symp. Inf. Theory*, pages 2101–2105, 2009.

[7] Vincent van der Leest, Geert-Jan Schrijen, Helena Handschuh, and Pim Tuyls. Hardware intrinsic security from D flip-flops. In *Fifth ACM Work. Scalable Trust. Comput. - STC '10*, page 53, New York, New York, USA, 2010. ACM Press.

Localization in Long Range Communication Networks Based on Machine Learning

Hazem Sallouha Sofie Pollin
KU Leuven
Dept. Elektrotechniek, TeleMic
Kasteelpark Arenberg, 10 - 2444
hazem.sallouha@esat.kuleuven.be sofie.pollin@esat.kuleuven.be

Abstract

As the number of devices connected to internet is rapidly increasing, it is expected that by 2025, every device will have a wireless connection, hence leading to trillions of wirelessly connected devices. Therefore, internet of things (IoT) with long range, low power and low throughput (e.g., Sigfox and LoRa) are raising as a new paradigm enabling to connect those trillions of devices efficiently. In such networks with low power and throughput, localization became more challenging. However, in most of IoT applications (e.g., asset tracking) we are interested in localizing the nodes within a certain area, rather than estimating the exact position with global positioning system (GPS) coordinates. Therefore, the problem can be simplified to estimate the node's sector. In this paper we propose a localization mechanism based on machine learning and assuming that some nodes in the sector are integrated with a GPS. By using these GPS-nodes as a reference the network can learn the position of the other nodes. The results revealed that using the user back-end measurements (e.g., received signal strength (RSS), number of base stations and the end-to-end delay) nodes can be divided in sectors. Then, the GPS-node is used to define the coordinates of the sector. Moreover, a trade off between number of messages and localization accuracy is illustrated.

1 Introduction

Wireless technologies penetrate in all layers of our daily live. Concepts such as "Internet of Things (IoT)" and "Location-Based Services" are rising as a new communication paradigms. The idea behind IoT is to have the objects of our daily life equipped with microcontrollers, transceivers for digital communication, and suitable protocol stacks that will make them able to communicate and becoming an integral part of the Internet [1]. In such networks location information can be exploited in different layers from communication aided perpouses to application level where location information are needed to meaningfully interpret any physical measurements collected by sensor nodes (SN) [1], [2].

In order to enable connectivity for hundreds of SN, the currently deployed IoT networks are using long-rage, low power and low throughput communications [4]-[5]. These characteristics made the localization problem more challenging. In one hand it is too expensive to integrate a GPS receiver in each SN and on the other hand, ranging-based localization techniques face lack of accuracy because of low bandwidth and long distances [7]. An alternative promising localization method is fingerprinting-based localization [8]-[10]. Fingerprinting localization usually works in two phases: offline phase and online phase. In offline phase measurements are collected form SN at different locations and stored as a training data. During the online phase, the measurements of

the SN are compared with the training data to estimate their location. The comparing process is usually done using machine learning algorithms [9]. A comparison between various machine learning algorithms is presented in [11]. It has been shown that decision tree J48 [12] is one of the top accurate algorithms. A localization method for IoT is introduced in [13] which can satisfy diverse requirements for indoor and outdoor scenarios. However, the long range communication is not considered.

In this work we propose a localization algorithm for long range IoT networks based on machine learning. Assuming some nodes are equipped with a GPS receiver the wide served region can be split up into sectors based on the GPS sensor nodes (GPS-SNs). Then, the other SN will be classified to one of the sectors using messages fingerprinting.

The rest of the paper is organized as follows, section 2 present localization in long range IoT networks. First, the problem is defined then the details of the proposed algorithm is described. Section 3 demonstrates the experimental results. Finally, Section 4 concludes this paper.

2 Localization in Long range IoT Networks

An IoT network is basically a network where massive number of nodes is connected to the same cloud. One kind of IoT networks is the long range networks [4]-[5]. This kind of networks enable connectivity for hundreds of nodes using relativity low number of base stations [3]. Example of these networks is Sigfox network [4]. Sigfox network is cellular based where the uplink data flow from nodes to base stations is assumed to be 97% of the data traffic. However, in some cases a hybrid networks can be used where nodes forward data to a getaway node which in turn send the data to base stations [1].

2.1 Problem Definition

Since long range IoT networks are cellular based, one may suggest to use multilateration ranging-based techniques for localization. Although they are cellular based, there are two main differences that make ranging-based localization inaccurate. First of all, since each base station covers a relatively large area (long range base station can cover be up to 40 km [3]), using multilateration will lead to a relatively large prediction zone. Secondly, multilateration is a ranging based techniques. Therefore, the accuracy is depend on the estimated distance between base station and nodes. The distance between base station and nodes can be calculated using received signal strength (RSS) and/or time of arrival (ToA). The Cramr-Rao lower bound (CRLB) of a distance estimate $\widehat{d}$ derived from RSS estimation provides the following inequality [6]

$$\sqrt{Var(\widehat{d})} \geqslant \frac{\ln 10}{10} \frac{\sigma_{sh}}{n_p} d \tag{1}$$

where $Var\{\cdot\}$ is the variance, d is the distance between the base station and the node, n_p is the path loss factor, and σ_{sh} is the standard deviation of the zero mean Gaussian random variable representing the log-normal channel shadowing effect. From Eq. (1) we observe that increasing the distance d, which is the case in long range communication, will decease the estimation accuracy by increasing the variance. On the other hand, the best achievable accuracy of a distance estimate $\widehat{d}$ derived from ToA estimation satisfies the following inequality [6]

$$\sqrt{Var(\widehat{d})} \geqslant \frac{c}{2\sqrt{2}\pi\sqrt{SNR\beta}} \tag{2}$$

where c is the speed of light, SNR is the signal-to-noise ratio, and β is the effective (or root mean square) signal bandwidth. Unlike RSS ranging technique, the accuracy of ToA is based on the signal bandwidth. However, long range communication is mainly based on ultra-narrow-band (UNB) (100Hz) transmission [4]. Therefore, this UNB property leads to extremely inaccurate location estimates using time-based techniques.

2.2 Proposed Sector-Based Localization

Consider an outdoor long range communication network (i.e., Sigfox) with a massive number of nodes distributed in a wide area (>40km) with two types of nodes: (i) normal sensor node (SN), without GPS, use to collect data and forward it to base stations, and (ii) sensor node that has an extra GPS receiver (GPS-SN), therefore, its messages have a GPS coordinates included. The GPS-SNs with their known location can be used to virtually partition the wide coverage region to a smaller *sectors* as shown in Fig. 1 where each sector is assumed to has at least one GPS-SN. Note that, Fig. 1 does not give any indication about the cell coverage since every message being sent form any sector $\{C_{i,j}\}$ will be received by at least 3 base station.

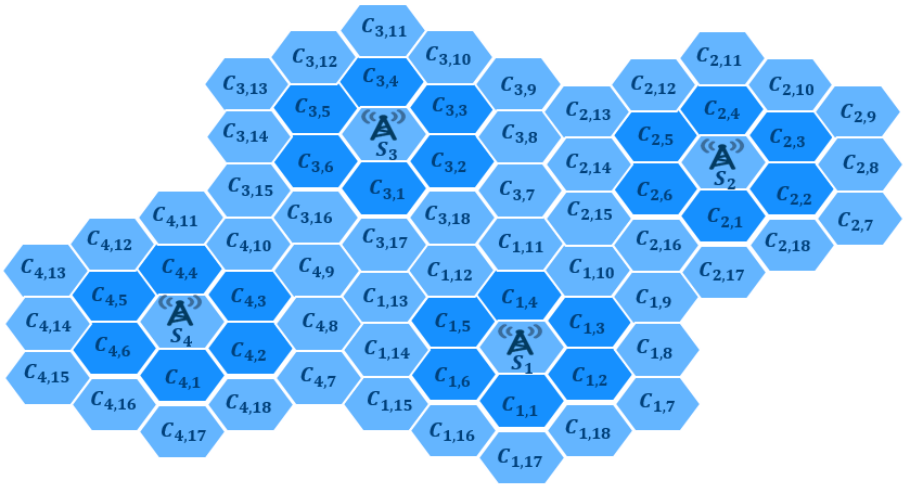


Figure 1: Illustrative example of how a wide region can be partitioned to a smaller regions each with a GPS-SN. Each cluster $C_{i,j}$ has a GPS-SN

We propose a localization algorithm based on machine learning that uses fingerprinting (RSSI, received SNR and base stations that have received the message) of the GPS-SN messages as a reference training data for classifying the other SN to one of the sectors. The machine learning algorithm (i.e., Decision Tree [12], Nave Bayes [14]) will assign each SN to one of the GPS-SN's sector based on their fingerprinting. Let N be the number of GPS-SNs in a defined region and M be the number of base stations that have received a message form a SN. We define region around the GPS-SN as a *sector* and denote a given sector by $\{C_{i,j}\}$. Then, the complete set of all sectors are given by

$$\mathbf{C} \quad = \quad \cup\{C_{i,j}\}, \forall i,j \tag{3}$$

where $i \in \{1,2,..,M\}$ and $j \in \{1,2,..,N\}$. The proposed algorithm shown in flow chart in Fig. 2 has two main steps:

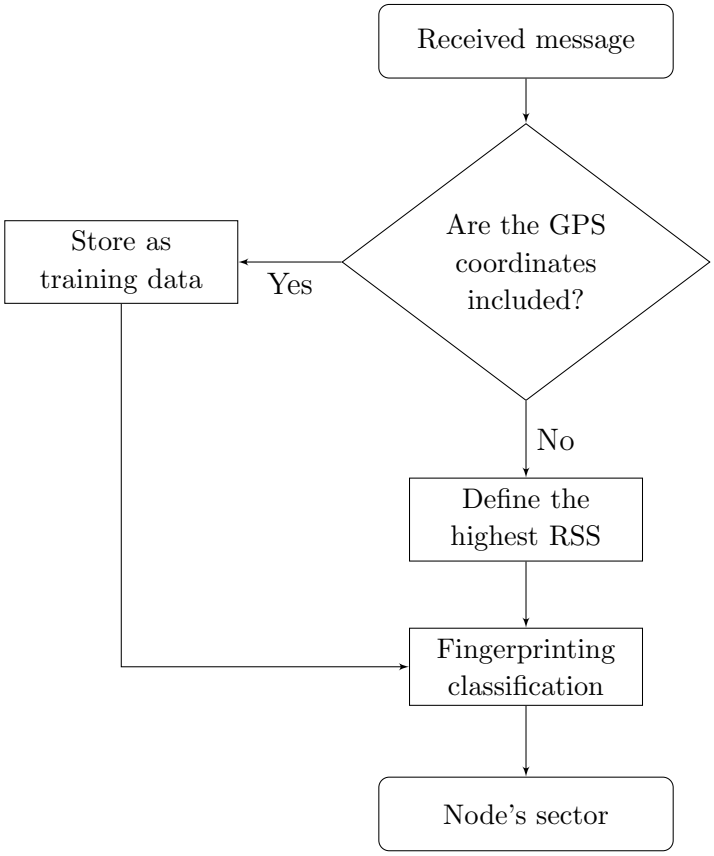


Figure 2: Proposed algorithm flow chart

1. Define the base station that received the message with highest RSS. For instance, assuming that base station S_1 received the highest RSS. Then, the sector of which the message has been transmitted form will be $\{C_{1,j}\}$ where $j \in \{1, 2, .., N\}$.

2. Use machine learning algorithm to classify messages fingerprinting in order to defined the exact sector of the node. This can be done by classifying each SN to one of the GPS-SN sectors. In long range communication networks the fingerprinting of message will include RSS, base stations that received the message, received SNR, and the number of message's replica.

The proposed algorithm can implicitly overcome the shadowing effect, this is due to the fact that the GPS-SN has a precise location and the other SN in the same sector will suffer relatively the same shadowing effect.

3 Measurements and Discussion

3.1 System Model

The system used is based on Sigfox network. Sigfox network has been deployed in real world and covered a huge part of Belgium with decent amount of base stations [4]. Sigfox is a standard developed by the telecom operator Sigfox. The main innovation of Sigfoxs network is the UNB (100Hz) transmission being used. This technology, with

the low spectral bandwidth, guarantees a long range communication between nodes and base stations while maintaining a limited transmission power. Binary phase shift keying (BPSK) is used as modulation technique. Because of BPSK spectral efficiency the bit rate is fixed to 100 bits per second.

Many Sigfox modules has been developed by different vendors. Some of these modules are integrated with a GPS receiver (i.e., TD1024 by TD next). Two modules with GPS receiver have been used in the measurements. These two modules will divide test area into two sectors.

3.2 Test Setup

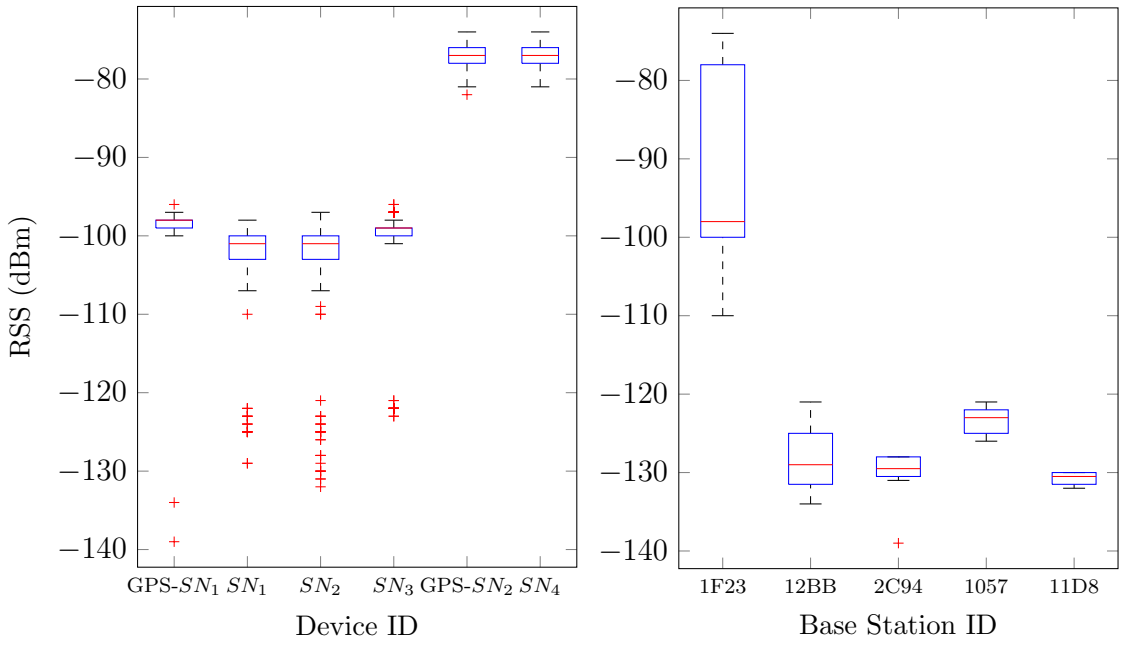
Six nodes in different position have been used in the measurements. A map of the node's positioned on top of two different buildings in Arenberg campus is shown in Fig. 3. Using this setup the proposed algorithm can be tested in order to localize the sector of SN where each GPS-SN will represent a sector. Moreover, with the node on Mechanical engineering department we can check how different their measurements are form the other nodes on ESAT building. Each node has sent 240 messages, along with each message information such as base station IDs that received the message, RSS at each base station and SNR can be seen on the user back end. As illustrated in the algorithm flow chart the messages received from the GPS-SNs will be used as a training data in order to classify the other received messages.

Figure 3: Map of the Arenberg campus with the position of the nodes used in the measurements

3.3 Results Analysis

This section details some initial simulation results of the proposed sector-based localization. WEKA software has been used during measurements analysis [15]. Fig. 4 present the highest RSS measurements verses both device ID and base station ID. Obviously, since the measurements have been done in a relatively small region, the same base station (1F23) has mostly received the highest RSS from the different SN. Moreover, Fig. 4a shows that the RSS values of the two nodes GPS-SN_1 and GPS-SN_2 are separated with 23 dB.

One can say that, based on the two figures Fig. 4a and Fig. 4b, we can easily defined the nodes into two sectors. However, in practical scenario, this will require a large number of messages which is challenging in terms of the delay required to collect



(a) RSS vs Device ID

(b) RSS vs Base station ID

Figure 4: The highest RSS versus base station ID and device ID

large number of messages. For instance, Sigfox networks allows each individual SN to send only 140 message per day (1 message every 10 minutes) [4]. Therefore, a machine learning (Decision tree, C4.5 [12]) algorithm is used to consider all the possible measurements at the same time. Then, a low number of training data will be required to correctly localize the other messages one by one upon arrival. As shown in Fig. 5 only 3 training messages are enough to get a 95% correctly classifying 240 messages from 3 SN and the other 237 from the two GPS-SNs, while after 10 training messages we can get almost 100% correctly classifying. However, as the density of nodes increase the minimum number of message will increase since more training data will be needed to learn the sector of each node.

Fig. 6 present a more challenging case, in which SN_1 is replaced by GPS-SN_2. Then, we used the proposed algorithm in order to classify the other two nodes on ESAT roof. Since the nodes are close to each other, more training messages are needed. As shown in Fig. 6 in case of messages by message classification, more than 100 training messages are required in order to get a classification accuracy over 90%. On the other hand, a clear improvement in accuracy is achieved by averaging the RSS of the messages 10 by 10. Based on the observations from Fig. 5 and Fig. 6 there is trade-off between localization accuracy which is represented by the density of GPS-SNs and delay which is represented by the number of messages.

4 Conclusion

A sector-based localization algorithm is proposed for long range IoT networks where the training data are collected by GPS-SN. Assuming GPS-SNs are distributed in the wide region, using the proposed algorithm we can learn the location of the other nodes. Simulation results show that there is a trade-off between the required number of training data to accurately localize the SN and localization accuracy. Furthermore,

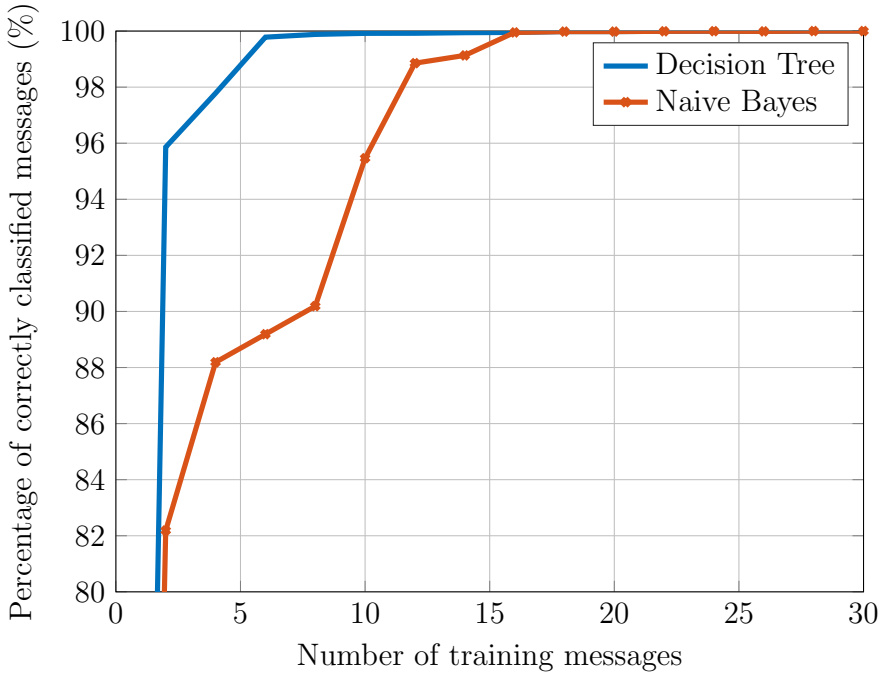


Figure 5: Percentage of correctly classified messages using two different machine learning algorithms

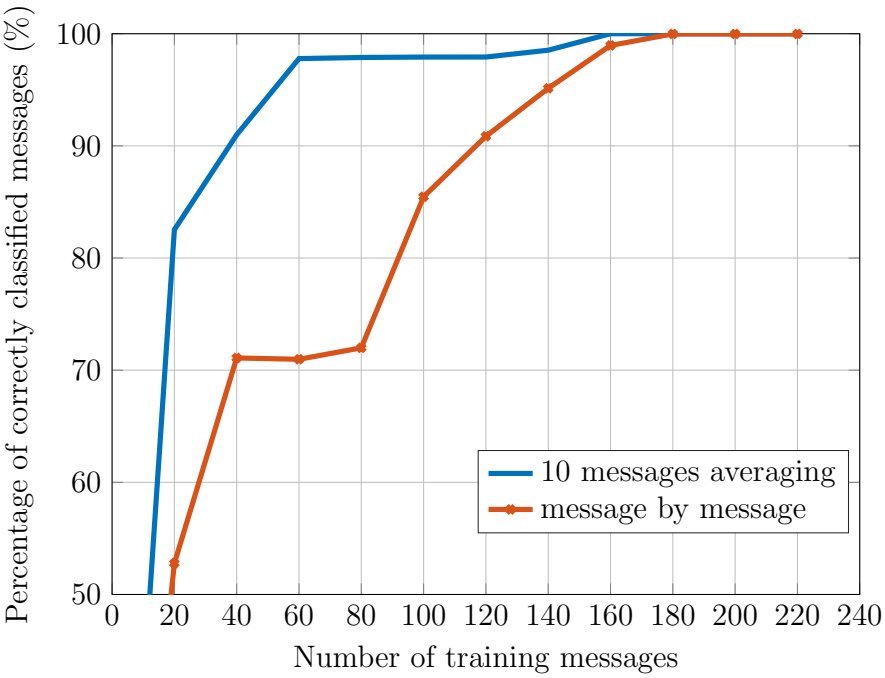


Figure 6: Percentage of correctly classified messages using decision tree machine learning algorithm in case of message by message and 10 messages averaging

the results show that averaging the messages finger printing measurements can improve classification accuracy and decrease the number of training messages required.

References

[1] A. Zanella, N. Bui, A. Castellani, and L. Vangelista, "Internet of Things for Smart Cities," IEEE Internet Things J., vol. 1, no. 1, pp. 2232, Feb. 2014.

[2] R. D. Taranto, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks,"IEEE Signal Process. Mag., vol. 31, no. 6, pp. 102112, Nov. 2014.

[3] B. Reynders, W. Meert, and S. Pollin, "Range and Coexistence Analysis of Long Range Unlicensed Communication," 23rd International Conference on Telecommunications (ICT), pp. 390-395, May 2016.

[4] Sigfox, "http://www.sigfox.com," 2014, accessed: May 2016.

[5] Lora alliance, http://lora-alliance.org, 2015, accessed: May 2016.

[6] S. Gezici, Z. Tian, G.B. Giannakis, H. Kobayashi, A.F. Molisch, H.V. Poor, and Z. Sahinoglu, "Localization via ultra-wideband radios," IEEE Signal Processing Mag., vol. 22, no. 4, pp. 7084 2005.

[7] S. Halder1 and Ghosal , "A survey on mobile anchor assisted localization techniques in wireless sensor networks," Springer, Wireless Networks, pp. 120, 2015.

[8] S. Fang, T. Lin, and K. Lee, "A Novel Algorithm for Multipath Fingerprinting in Indoor WLAN Environments," IEEE Trans, Wireless Comm., vol. 7, no. 9, pp. 3579-3588, Sep. 2008.

[9] K. Mikkel and Baun, "A taxonomy for radio location fingerprinting, Lecture Notes in Computer Science," Springer vol. 4718, pp. 139156, 2007.

[10] J. Torres-Sospedra, R. Montoliu, A. Martnez-Us, T. J., J. P., Avariento, M., Benedito Bordonau, and J., Huerta, "UJIIndoorLoc: A New Multi-building and Multi-floor Database for WLAN Fingerprintbased Indoor Localization Problems," 5th International Conference on Indoor Positioning and Indoor Navigation, 2014.

[11] S. Bozkurt1, G. Elibol, S. Gunal and U. Yayan "A Comparative Study on Machine Learning Algorithms for Indoor Positioning," INISTA, 2015 International Symposium, pp. 1-8, Sep. 2015.

[12] J. R. Quinlan, "C4.5: programs for machine learning," Elsevier, 2014.

[13] Z. chen, F. Xia, F. Bu and H. Wang, "A localization method for the Internet of Things," Springer Super computing, pp. 657674, Sep 2013.

[14] G. H. John, and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," 11th Conference on Uncertainty in Artificial Intelligence, pp., 338-345, 1995.

[15] http://www.cs.waikato.ac.nz/ml/weka/, WEKA.

Privacy-Preserving Alpha Algorithm for Software Analysis

Gamze Tillem Zekeriya Erkin Reginald L. Lagendijk
Delft University of Technology
Department of Intelligent Systems, Cyber Security Group
Delft, The Netherlands
G.Tillem@tudelft.nl Z.Erkin@tudelft.nl R.L.Lagendijk@tudelft.nl

Abstract

Validation in a big software system can be managed by analysis of its behaviour through occasionally collected event logs. Process mining is a technique to perform software validation by discovering process models from event logs or by checking the conformance of the logs to a process model. A well-known algorithm in process mining to discover process models is alpha algorithm. However, while utilising alpha algorithm is useful for software validation, the existence of some sensitive information in the log files may become a threat for the privacy of users. In this work, we propose a protocol for privacy-preserving alpha algorithm on encrypted data. Our protocol aims to generate process models for a software without leaking any information about its users. It achieves same computational complexity with the original algorithm despite the additional computation overhead.

1 Introduction

Software systems have an evolving nature which enables them to grow continuously with new updates and interactions. While growth of software systems is beneficial for its functionality, conversely, it complicates managing its validation. In the traditional approach software validation is maintained by analysing the conformance of pre-defined cases in the design time. However, for complex software systems which has interactions with several external tools, a priori prediction of cases is challenging. This challenge introduces a new approach for software validation which shifts the validation procedure to online phase, namely, analysis of software in the run time. The event logs that are generated during the execution of software, enable observation of behaviour and checking the conformance of design requirements.

A non-trivial technique to monitor software behaviour for validation is process mining. As a field between data mining and process modeling, the aim of process mining is to discover, monitor and improve the real processes by extracting information from the event logs [13]. Process mining utilises log information in three categories. The first category is process discovery which generates a process model from log data. The second category is conformance checking whose purpose is to indicate that the real behaviour of the system conforms to the model by comparing event log of a process with an existing process model. Finally, the third category is enhancement where an existing process model is improved by comparing it with event logs.

As the core component of process mining, log data has a crucial role to determine in which way the software behaviour is modelled. Software log can contain information about user, system settings (e.g. type of operating system, number of cores, memory usage), interactions with other components and the date or duration of execution. Aforementioned information is valuable for process miner to obtain knowledge about the software behaviour. On the other hand, the content of information is vulnerable

against privacy threats since it may contain sensitive information about user or system. An example of such a threat is recently experienced by GHTorrent platform [6]. Aiming to monitor GitHub events to simplify searching on them, GHTorrent does not consider removal of personal data from events. However, it appears that some users of the platform abused the personal data to send survey invitations to data owners [5]. Receiving hundreds of e-mails from external parties, the data owners has started complaining about their privacy in collected logs* which, in the end, required the platform developers to revise their privacy policy [6]. The case of GHTorrent shows that as log based software analysis gets popular, the importance of privacy in log files becomes prominent.

In our work, we aim to design a privacy-preserving process discovery protocol to generate process models from event logs while guaranteeing the privacy of event logs. As an initial step, alpha algorithm [14] is selected for process discovery since it clearly shows the steps for discovery of process models. Our protocol utilises encryption to guarantee the confidentiality of logs. To overcome difficulty of retrieving information from encrypted data, we use homomorphic encryption schemes which enable us to perform operations on ciphertext without using decryption mechanism. These schemes are useful especially in multiparty settings which requires prevention of information leakage to other parties while performing computations on encrypted data.

Privacy in software is investigated from different aspects by research community. Several works focus on providing privacy in released test data through anonymization techniques [7, 10] or machine learning techniques [8]. Some other works, e.g. [2, 3], are interested in controlling crash report generation to eliminate sensitive information in reports. Furthermore, preventing the leakage of sensitive data from running software is another concern in software privacy which is achieved by utilising information flow mechanisms in [4] and [15]. However, to the best of our knowledge, none of the existing works deals with the privacy of software validation under process mining. Our protocol is the first attempt to operate process discovery algorithms on software in a privacy-preserving manner.

In the rest of the paper, first we provide some preliminary knowledge (Section 2). Then, we introduce the protocol for privacy preserving alpha algorithm in Section 3 and continue with the complexity analysis in Section 4. Finally, in Section 5, we conclude our paper and explain the directions of future research.

2 Preliminaries

Prior to explain the protocol for privacy-preserving alpha algorithm, we provide some preliminary knowledge about alpha algorithm and cryptographic tools in this section.

2.1 Alpha Algorithm

Alpha algorithm is one of the first process discovery algorithms to discover process models from event logs. Since it covers basic steps of discovery, it is favourable as a starting point for process discovery. It takes an event log L as input and outputs a process model. The process model is represented as a *Petri net*, which is a modelling language used in process mining [14]. L is a set of traces and each trace is a set of activities. Formally, $L = [\sigma_1, \sigma_2, ...\sigma_x]$ where σ_i is a trace and $x \in Z^+$. For each $\sigma_i = \langle t_{1_i}, ..., t_{j_i} \rangle$, t_{j_i} is an activity where $1 \leq j_i \leq K$ and K is the maximum number of activities. Then,

$$L = [\langle t_{1_1}, ..., t_{j_1} \rangle, \langle t_{1_2}, ..., t_{j_2} \rangle, ..., \langle t_{1_x}, ..., t_{j_x} \rangle].$$

Alpha algorithm runs in 8 steps to generate a process model. In this section, the steps of alpha algorithm is explained through an example. Assuming that following event log L is collected from a software

*https://github.com/ghtorrent/ghtorrent.org/issues/32

$L = [\langle a, b, e, f\rangle, \langle a, b, e, c, d, b, f\rangle, \langle a, b, c, e, d, b, f\rangle, \langle a, b, c, d, e, b, f\rangle, \langle a, e, b, c, d, b, f\rangle]$,

the alpha algorithm proceeds through the following steps :

Step 1: Discovers distinct set of activities (T_L) in $L \implies T_L = \{a, b, c, d, e, f\}$.

Step 2: Discovers initial activities (T_I) in each $\sigma_i \implies T_I = \{a\}$ and assigns an initial place i_L.

Step 3: Discovers final activities (T_O) in each $\sigma_i \implies T_O = \{f\}$ and assigns a final place o_L.

Step 4: Groups activities using ordering relations (direct succession ($>$), causality ($\rightarrow$), parallel ($||$) and choice ($\#$)) [14] to create relation set X_L. The relations between each activity can be represented in a footstep matrix as in Figure 1. Then, using footstep matrix, X_L is $\implies$

$X_L = \{(\{a\}, \{b\}), (\{a\}, \{e\}), (\{b\}, \{c\}), (\{b\}, \{f\}), (\{c\}, \{d\}), (\{d\}, \{b\}),$
$(\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$.

Step 5: Removes pairs from X_L to create an optimised relation set (Y_L) $\implies$

$Y_L = \{(\{a\}, \{e\}), (\{c\}, \{d\}), (\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$.

Step 6: Determines set of places for process model (P_L) $\implies$

$P_L = \{p_{(\{a\},\{e\})},\ p_{(\{c\},\{d\})},\ p_{(\{e\},\{f\})},\ p_{(\{a,d\},\{b\})},\ p_{(\{b\},\{c,f\})},\ i_L,\ o_L\}$.

Step 7: Connects places P_L by introducing arcs F_L.

Step 8: Returns $\alpha(L) = (P_L,\ ,T_L,\ F_L)$ which is demonstrated as a Petri net in Figure 1.

	a	b	c	d	e	f
a	#	←	#	#	←	#
b	→	#	←	→	||	←
c	#	→	#	←	||	#
d	#	←	→	#	||	#
e	→	||	||	||	#	←
f	#	→	#	#	→	#

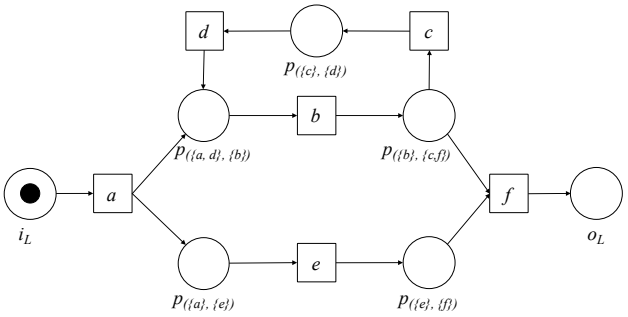


Figure 1: Footstep matrix and process model as Petri net for $E(L)$.

2.2 Cryptographic Tools

As stated in Section 1, we construct our protocol on homomorphic encryption schemes to prevent leakage of sensitive information during computations. Considering the trade-off between somewhat homomorphic and additively homomorphic schemes with respect to efficiency of operations and functionality of cryptosystem, we decide to analyse our protocol both on additive homomorphic scheme, Paillier cryptosystem [11] , and on somewhat homomorphic scheme, YASHE [1].

Paillier cryptosystem: Based on decisional composite residuosity problem [11], Paillier cryptosystem can encrypt a plaintext m on a modulus $N = p \cdot q$ where $p,\ q$ are large primes and $g = n + 1$ as $E(m) = g^m \cdot r^N \mod N^2$ where $r \in_R Z_N$. The cryptosystem enables to perform addition and scalar multiplication on encrypted text. Two encrypted plaintext $m_1,\ m_2$ can be added as $E(m_1) \cdot E(m_2) = E(m_1 + m_2)$. Scalar multiplication is performed as $E(m_1)^c = E(c \cdot m_1)$.

YASHE: While Paillier cryptosystem is constructed on integers, YASHE scheme is constructed on ideal lattices. The security of the scheme is based on Ring Learning with Errors assumption [1]. Because of page limitation, we refer readers to [1] for more details. Here we only summarise homomorphic properties of YASHE scheme.

We are given two ciphertexts c_1 and c_2 which are encryptions of m_1 and m_2 and $[\cdot]_a$ refers to reduction to modulus a. Then, homomorphic addition is achieved by adding c_1 and c_2 as $c = [c_1 + c_2]_q$ which is equal to the encryption of $[m_1 + m_2]_t$. On the other hand, homomorphic multiplication is performed in two phases. In the first phase an intermediate ciphertext $\hat{c} = [\lceil t/q \cdot c_1 \cdot c_2 \rfloor]_q$ is computed. Since this operation increases noise which prevents a correct decryption of ciphertext [1], in the second phase a Key Switching mechanism is applied to $\hat{c}$ to transform it into a decryptable ciphertext c.

3 Privacy-Preserving Alpha Algorithm

We now describe our protocol for privacy-preserving alpha algorithm on encrypted data. The protocol is based on semi-honest model with three entities which are User, Log Repository and Process Miner. User is the end user of a software who generates event logs and sends them to Log Repository in encrypted form. Log Repository is a semi-honest storage unit which is responsible for collecting and storing encrypted event logs. It can be either specific to a certain software product or a common repository which manages logs from different software products. Since Log Repository is not fully trusted, in our setting it is not allowed to see order relations between any two encrypted activities of event log L. Finally, Process Miner is a semi-honest third party which has capabilities to generate process models from encrypted event logs. To be able to generate process models, Process Miner has to learn the order relations in event logs. However, it cannot learn the content of log files.

The protocol is based on three main phases which are Set Up, Relation Discovery and Model Discovery which are demonstrated in Figure 2. As it is explained later in this section, Relation Discovery phase requires utilisation of secure equality tests to discover relations between encrypted activities. Thus, Secure Equality Check subcomponent is integrated to that phase. We show two efficient Secure Equality Check mechanisms [12, 9] here, but, other efficient mechanisms can also be adapted to the protocol. Rest of this section explains each phase of privacy-preserving alpha algorithm protocol in detail.

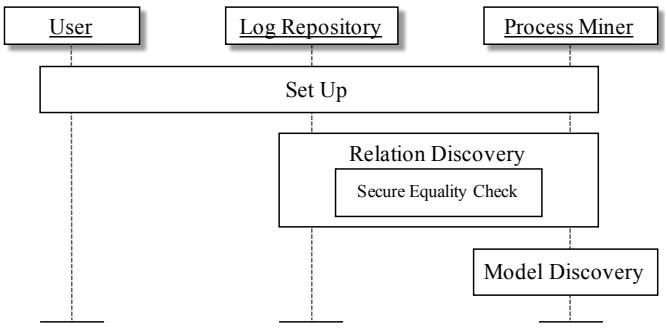


Figure 2: Overview of Privacy-Preserving Alpha Algorithm

3.1 Set Up

In Set Up, initially cryptographic keys are generated by a trusted third party and distributed to related entities. Since user is only responsible for generation of encrypted event logs, he is provided public key pk. Log Repository and Process Miner are given their secret shares sk_{LR} and sk_{PM}, respectively.

In the second part of Set Up phase, according to user's interaction with software an event log L is generated, as explained in Section 2. After generation of L, user

encrypts it under selected encryption scheme (Paillier or YASHE) using pk and outsources encrypted log $E(L)$ to Log Repository. Finally, Log Repository shares $E(L)$ with Process Miner which is going to discover process model in encrypted log data. The format of data that Process Miner retrieves is:

$$E(L) = [\langle E(t_{1_1}), ..., E(t_{j_1})\rangle, \langle E(t_{1_2}), ..., E(t_{j_2})\rangle, \ \dots\ , \langle E(t_{1_x}), ..., E(t_{j_x})\rangle].$$

3.2 Relation Discovery

The core of our protocol is to securely detect distinct activities, address initial and last activities in each trace and identify the relations between them. To that end, we construct a relation table RT whose indices correspond to encrypted activities. For each index, RT shows whether the activity is initial (Init) or last (Last) in its trace and it stores the list of direct successors (Direct Successor) for the activity. When an encrypted activity $E(t_{y_i})$ where $y \in [1, j]$ is retrieved, RT is searched to find a match for $E(t_{y_i})$ using secure equality checks. If there is no match for $E(t_{y_i})$, then it is inserted into RT as a new index. Figure 3 demonstrates the procedure of Relation Discovery phase.

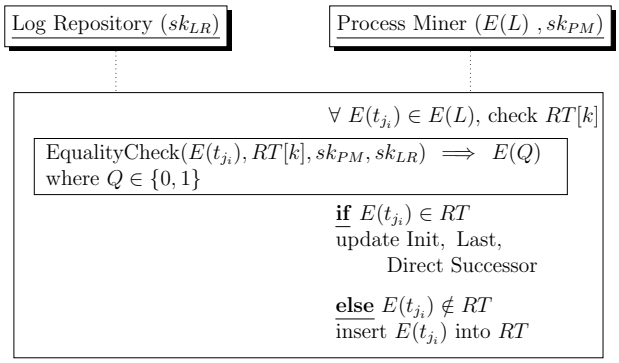


Figure 3: Overview of Relation Discovery phase

To clarify the procedure, we can construct RT by using the example log data in Section 2. Initially, Process Miner has the following encrypted log data:

$$E(L) = [\langle E(a), E(b), E(e), E(f)\rangle, \langle E(a), E(b), E(e), E(c), E(d), E(b), E(f)\rangle,$$
$$\langle E(a), E(b), E(c), E(e), E(d), E(b), E(f)\rangle, \langle E(a), E(b), E(c), E(d), E(e), E(b), E(f)\rangle,$$
$$\langle E(a), E(e), E(b), E(c), E(d), E(b), E(f)\rangle].$$

Starting from the first activity $E(a)$ in trace $\sigma_1 = \langle E(a), E(b), E(e), E(f)\rangle$, Process Miner scans RT to find a match for the current activity. Since initially the table is empty, $E(a)$ is directly added to RT (Table 1). For second activity, $E(b)$, one equality check should be performed to compare it with $E(a)$. Since $E(b) \neq E(a)$, $E(b)$ is inserted into RT as a new index (Table 2). Furthermore, since $E(b)$ directly follows $E(a)$, it is added into Direct Successor list of $E(a)$. When the same operations are applied for each encrypted activity, the relation table RT is completed as shown in Table 3.

3.2.1 Secure Equality Check for Relation Discovery

Construction of RT requires comparison of encrypted activities which has to be managed by secure equality check (SEC) mechanisms. Since proposing an equality check mechanism is not our main concern, we adapted two existing mechanisms to our protocol [9, 12]. Below, we briefly describe these mechanisms and refer the readers to [9, 12] for their detailed description.

	Index	Init	Last	Direct Successor
E(a)	0	+	+	

Table 1: RT with one element

	Index	Init	Last	Direct Successor
E(a)	0	+	+	1
E(b)	1	-	-	

Table 2: Inserting $E(b)$ into RT

	Index	Init	Last	Direct Successor
$E(a)$	0	+	+	1, 2
$E(b)$	1	-	-	2, 3, 4
$E(e)$	2	-	-	1, 3, 4, 5
$E(f)$	3	-	+	-
$E(c)$	4	-	-	2, 5
$E(d)$	5	-	-	1, 2

Table 3: Complete version of relation table RT

SEC by Toft [12]: Toft [12] proposes a SEC protocol by employing Jacobi symbol. The protocol requires a virtual trusted third party which is Arithmetic Black Box (ABB) to provide secure storage and to perform arithmetic computations. The equality of two values is tested by testing whether their difference d is equal to 0. For an encryption modulus M, if $d = 0$, then Jacobi symbol for $d + r^2$ where $r \in_R M$ is $J_{d+r^2} = \left(\frac{d+r^2}{M}\right) = 1$. Otherwise, if $d \neq 0$, $J_{d+r^2} = -1$. Although, Toft's scheme is efficient, the result is correct with 1/2 probability due to probabilistic nature of Jacobi symbol. Thus, reducing the probability to a negligible degree requires the repetition of protocol κ times with the same input.

SEC by Lipmaa & Toft [9]: Different from [12], Lipmaa and Toft [9] introduce a SEC protocol which utilises Hamming distance. Similar to [12], the protocol is based on zero check for difference d and ABB is responsible for secure storage and arithmetic computations. Hamming distance is computed between a random r and $m = r + d$. To reduce the complexity of operations in Hamming distance computation, an offline preprocessing phase to compute random values, random inverses and random exponents is proposed. Furthermore, for online phase Lagrange interpolation is used. Although the result of the protocol is deterministic, it has drawback of computational complexity which is bounded by the bit length of encryption modulus M.

3.3 Model Discovery

After discovery of the order relations between encrypted activities, the final phase of our protocol generates the process model using the information in RT. In the original algorithm, a footstep matrix is constructed to demonstrate casual, parallel and choice relations between activities based on direct successions (see Section 2). In the same manner, our protocol constructs the footstep matrix using Direct Successor lists in RT. Finally, the process model as a *Petri net* is generated using the ordering relations in footstep matrix as it is showed in Figure 4.

4 Complexity Analysis

Utilising encryption is advantageous to maintain the confidentiality of log files. However, a qualified scheme for software analysis should also consider the efficiency of computations for practicability. Thus, in this section we analyse the complexity of our protocol.

To evaluate performance of our protocol, initially we have to investigate the complexity of original alpha algorithm. Since in the original algorithm computations are

	0 (a)	1 (b)	2 (e)	3 (f)	4 (c)	5 (d)
0 (a)	#	→	→	#	#	#
1 (b)	←	#	||	→	→	←
2 (e)	←	||	#	→	||	||
3 (f)	#	←	←	#	#	#
4 (c)	#	←	||	#	#	→
5 (d)	#	→	||	#	←	#

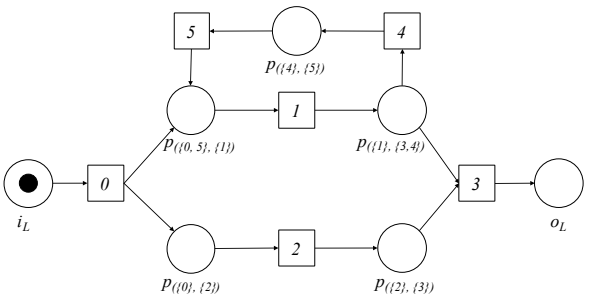


Figure 4: Footstep matrix and process model as Petri net for $E(L)$.

handled by one party (Process Miner), the complexity can be analysed only in terms of computational cost. Step 4 and 5 of the algorithm dominate computational complexity. Construction of footstep matrix in 4^{th} Step requires $\mathcal{O}(xK^2)$ comparisons to find order relations where x is total number of traces and K is maximum number of activities in one trace. In Step 5, $\mathcal{O}(K^2)$ comparisons are performed to find maximal relation sets. Consequently, the overall computational complexity of original algorithm is $\mathcal{O}(xK^2)$.

In our protocol, the computational complexity is dominated by construction of relation table RT. Similar to the original alpha algorithm, this process necessitates $\mathcal{O}(xK^2)$ comparisons in the worst case. However, each comparison is performed by running a secure equality check protocol rather than integer or string comparisons as in the scheme with plaintext. Therefore, despite in theoretical bounds our protocol has the same complexity with the original scheme, it is useful to analyse the cost of one equality check protocol to understand additional cost of encryption in the protocol. Table 4 overviews the complexity of computations for SEC protocols [9, 12] which are used in Relation Discovery phase. To comply with notations in ABB schemes, the complexity as the number of ABB operations is also provided.

	Using SEC from [9]	Using SEC from [12]
ABB operations	$\mathcal{O}(\ell)$	$\mathcal{O}(\kappa)$
Paillier based implementation (num of multiplications and exponentiations)	$\mathcal{O}(\ell)$	$\mathcal{O}(\kappa)$
YASHE based implementation (num of additions and multiplications)	$\mathcal{O}(\ell)$	$\mathcal{O}(\kappa)$

Table 4: Overview of complexity in SEC protocols

The analysis results shows that the cost of computation is bounded by bit size in [9] where $\ell = \lceil \log_2 M \rceil$. The operations in the preprocessing phase dominates the complexity of protocol. On the other hand, in [12] computation cost is determined by correctness parameter κ. Although, one run of protocol is handled by constant number of operations, since the result is probabilistic, it requires κ repetitions. Finally, both additive and somewhat homomorphic setting have same theoretical bounds, but in reality the number of operations for somewhat homomorphic based implementation is less than the number of operations in additive homomorphic implementation. However, it does not necessarily imply that somewhat homomorphic setting is more efficient than additive homomorphic since the bit size of modulus and the complexity of operations differ in two settings.

5 Conclusion and Future Work

In this work we have addressed for the first time the privacy in software analysis under process mining techniques. Specifically, we presented a naive protocol for privacy-preserving alpha algorithm to generate process models from encrypted event logs. Our protocol achieves same theoretical bounds with the original algorithm. However, it requires usage of secure comparison protocols which imposes additional computations

with larger bit sizes. In the future, we continue exploring privacy issues in different algorithms of process mining for software analysis. Furthermore, we extend our research with implementation of algorithms and utilisation of other privacy enhancing technologies.

References

[1] Joppe W Bos, Kristin Lauter, Jake Loftus, and Michael Naehrig. Improved security for a ring-based fully homomorphic encryption scheme. In *Cryptography and Coding*, pages 45–64. Springer, 2013.

[2] Pete Broadwell, Matt Harren, and Naveen Sastry. Scrash: A system for generating secure crash information. In *Proceedings of the 12th conference on USENIX Security Symposium-Volume 12*, pages 19–19. USENIX Association, 2003.

[3] Miguel Castro, Manuel Costa, and Jean-Philippe Martin. Better bug reporting with better privacy. In *ACM Sigplan Notices*, volume 43, pages 319–328. ACM, 2008.

[4] William Enck, Peter Gilbert, Seungyeop Han, Vasant Tendulkar, Byung-Gon Chun, Landon P Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N Sheth. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Transactions on Computer Systems (TOCS)*, 32(2):5, 2014.

[5] Arnoud Engelfriet. Is it legal for ghtorrent to aggregate github user data? `https://legalict.com/privacy/is-it-legal-for-ghtorrent-to-aggregate-github-user-data/`, 2016. Accessed May 3, 2016.

[6] Georgios Gousios. The issue 32 incident _ an update. `http://gousios.gr/blog/Issue-thirty-two`, 2016. Accessed May 3, 2016.

[7] Mark Grechanik, Christoph Csallner, Chen Fu, and Qing Xie. Is data privacy always good for software testing? In *Software Reliability Engineering (ISSRE), 2010 IEEE 21st International Symposium on*, pages 368–377. IEEE, 2010.

[8] Boyang Li. Enhancing utility and privacy of data for software testing. In *Software Testing, Verification and Validation Workshops (ICSTW), 2014 IEEE Seventh International Conference on*, pages 233–234. IEEE, 2014.

[9] Helger Lipmaa and Tomas Toft. Secure equality and greater-than tests with sublinear online complexity. In *Automata, Languages, and Programming*, pages 645–656. Springer, 2013.

[10] David Lo, Lingxiao Jiang, Aditya Budi, et al. kbe-anonymity: test data anonymization for evolving programs. In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, pages 262–265. ACM, 2012.

[11] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *Advances in cryptology–EUROCRYPT–99*, pages 223–238. Springer, 1999.

[12] Tomas Toft. Sub-linear, secure comparison with two non-colluding parties. In *Public Key Cryptography–PKC 2011*, pages 174–191. Springer, 2011.

[13] Wil Van Der Aalst, Arya Adriansyah, Ana Karla Alves de Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, Jagadeesh Chandra Bose, Peter van den Brand, Ronald Brandtjen, Joos Buijs, et al. Process mining manifesto. In *Business process management workshops*, pages 169–194. Springer, 2011.

[14] Wil Van der Aalst, Ton Weijters, and Laura Maruster. Workflow mining: Discovering process models from event logs. *Knowledge and Data Engineering, IEEE Transactions on*, 16(9):1128–1142, 2004.

[15] David Yu Zhu, Jaeyeon Jung, Dawn Song, Tadayoshi Kohno, and David Wetherall. Tainteraser: protecting sensitive data leaks using application-level taint tracking. *ACM SIGOPS Operating Systems Review*, 45(1):142–154, 2011.

A framework for processing cardiac signals acquired by multiple unobtrusive wearable sensors

Silviu Dovancescu [1] Attila Para [1,2] Dan Stefanoiu [2]
[1] Philips Research
Eindhoven, The Netherlands
silviu.dovancescu@philips.com

[2] Politehnica University of Bucharest
Dept. of Automatic Control and Systems Engineering
Bucharest, Romania
attila.para@stud.acs.upb.ro
dan.stefanoiu@acse.pub.ro

Abstract

Emerging wearable technologies, such as smartwatches or band aid-like sensors, enable unobtrusive continuous monitoring of cardiac signals anytime and anywhere. Users of these technologies generate an increasing demand for personal health applications to support disease prevention as well as a healthy lifestyle. As a first step in the development of personal health applications for wearables, researchers usually perform studies which include data collection using wearable physiological sensors in parallel with clinically accepted monitoring systems that serve as a reference. Handling the resulting large amounts of heterogeneous data remains challenging. In this paper, we present a framework for cardiac signal processing that facilitates research into personal health applications based on wearable sensors. We showcase the use of the framework in a feasibility study of unobtrusive blood pressure (BP) monitoring based on photoplethysmographic (PPG) and electrocardiographic (ECG) signals from wearable sensors.

1 Introduction

Cardiac signals represent a rich source of information on the general health status and the well-being of individuals. Monitored continuously over long periods of time, these signals may reveal concealed cardiac disorders or indicate early stages of health degradation.

Traditionally, the electrical activity of the heart, reflected by the electrocardiogram (ECG), has been the target of long term monitoring. ECG monitoring consists of measuring the potential between two or more points on the body. Beside the in-hospital use, ECG is also widely used both in home health care (e.g. Holter monitor, peel-and-stick patch) and sports applications (e.g. chest strap).

The photoplethysmogram (PPG) is a non-invasive circulatory signal which reflects the total volume changes in all blood vessels. It uses a single point measurement with optical skin contact, which is less obtrusive compared to the traditional ECG. PPG requires a LED to irradiate the tissue and a photodetector, which measures the light intensity. There are two different techniques for PPG monitoring: transmission mode and reflexive mode PPG. In transmission mode the LED and the photodetector are on the opposite side of a thinner part of the body (e.g. finger, earlobe), while in reflexive mode the photodetector is placed alongside the LED. In the latter case the light is reflected by deeper structures and absorbed in blood vessels [1].

The unobtrusive continuous monitoring of cardiac signals, e.g. ECG or PPG, was facilitated in the last decade by wearables such as smart bracelets, watches or band aid-like sensors which targeted lifestyle applications. However, the data generated by these devices is not yet exploited up to its potential for medical purposes. Users of wearables may be interested in new personal health applications that provide comprehensive analyses of their vital signs and potentially support them in their interaction with healthcare provider.

In order to develop reliable personal health applications for wearables, it is necessary to first conduct research studies that include data collection using wearable physiological sensors in parallel with clinically accepted monitoring systems which serve as a reference. These studies generate large amounts of heterogeneous data. To facilitate the research, we have designed and implemented a modular framework for cardiac signal processing. The framework enables processing and aggregation of multiple cardiac signals, acquired by multiple devices.

In this paper, we present the design of the cardiac signal processing framework, with focus on the modules for PPG processing. We showcase the use of the framework in a feasibility study aimed at unobtrusive monitoring of blood pressure (BP) using wearable sensors.

2 Cardiac signal processing framework

The proposed framework (Fig. 1) includes layers for importing, processing and synchronizing signals recorded by multiple sensors as well as for quantitative evaluation. Traditionally, the most used cardiac signals are from measurements like electrocardiogram (ECG), blood pressure (BP) and respiration, but other signals, such as the photoplethysmogram (PPG) or the impedance cardiogram (ICG) are also used in research. The signal processing layer includes modules for beat detection, extraction of beat morphology features and feature conditioning applied to the cardiac signals.

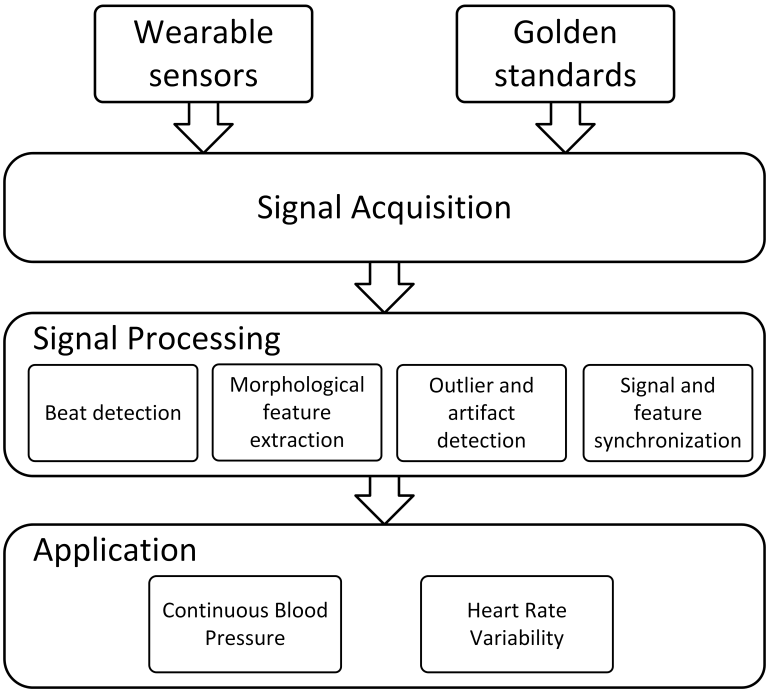


Figure 1: Framework for processing cardiac signals.

2.1 Signal acquisition

The signal acquisition layer consists of modules that interface between the cardiac signal analysis framework and external heterogeneous data sources. A dedicated module for each type of wearable or gold standard device, used in the research project, imports the raw data recorded by the device and converts it to a unified internal data structure.

2.2 Signal processing

2.2.1 Heartbeat detection

The heart beat detection module is the most important one, because the other processing modules are based mostly on signals segmented according to heartbeats. The PPG morphological features are extracted in a beat-to-beat manner and inter-beat-intervals are used for synchronization, both of them requiring accurate beat detection.

Heartbeats can be detected from several cardiac signals, including ECG and PPG. However, if available, ECG is the preferred signal, because the morphology of the QRS complexes makes an accurate and robust detection possible, even in the presence of noise. QRS detection was widely researched and many algorithms have been described in literature [2, 3]. We used the easy to implement and fast amplitude threshold method, because we are primarily interested in the detection of R-peaks.

2.2.2 Morphological features extraction

The PPG module extracts relevant PPG morphology features corresponding to each heartbeat.

The PPG waveform is composed of a DC component, which varies due to respiration and local autoregulation, and a pulsatile AC component. The AC component represents the changes of the blood volume in the capillary with each heartbeat [1, 4]. The rising edge (A-B) of the pulsatile component is concerned with systole while the falling edge (B-A_{next}) with diastole [5].

Fig. 2 outlines the characteristic points of the PPG waveform, represented with capital letters, as follows: A - foot, B - systolic peak, C - dicrotic notch, D - diastolic peak, Q - inflection point of the rising edge.

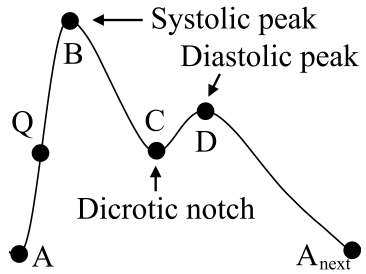
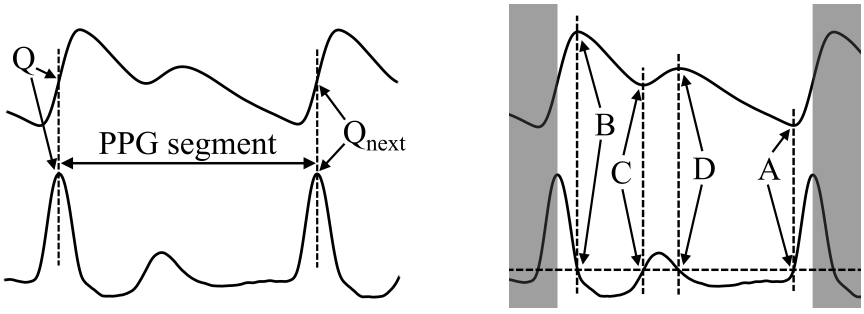


Figure 2: A typical waveform of the PPG during one heart cycle and its characteristic parameters.

The features were extracted separately for every heart cycle. Since the amplitude of the systolic peak is affected by many factors which can cause abrupt changes, the automatic segmentation process of the PPG signal was based on the signals first order derivate. This was used to find the inflection point (Q) of the rising edge. The inflection point is equivalent to the maximum value of the first derivative wave of the PPG, as shown in Fig. 3a.

The PPG signal was segmented according to the identified indexes of the inflection points. This operation was necessary for a correct detection of the systolic and diastolic peaks. The A and C points were identified in the first derivative wave for every segment as the two positive to negative zero-crossings of the time axis, while the B and D points are the two negative to positive zero-crossings. This method is illustrated in Fig. 3b.



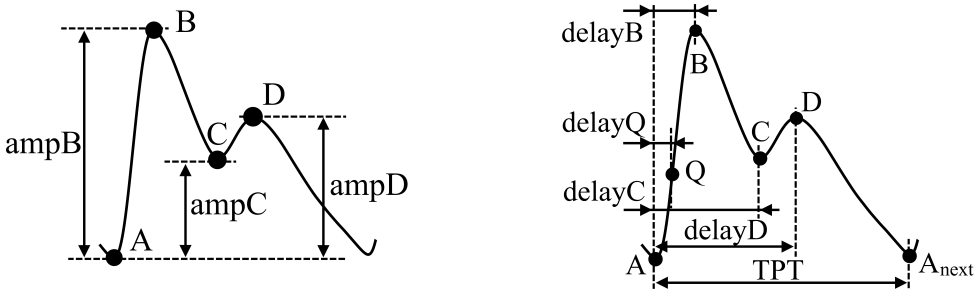
(a) Segmentation and inflection point of the rising edge.

(b) Characteristic point identification.

Figure 3: Fingertip PPG waveform and the first order derivative wave of it.

Amplitude indices The absolute amplitude of the PPG signal is influenced by multiple factors (e.g. skin compression under the photodetector), therefore all of the amplitude related features were defined as relative amplitudes. The amplitude of the foot (A) was selected as the base amplitude. The amplitudes of points B, C and D, as well as their rates (ampB/ampC, ampD/ampC and ampB/ampD) were included in the list of extracted features. Fig. 4a. illustrates these amplitudes.

Time span indices The A point serves as the reference point, also for time span indices. The total pulse time (TPT) was defined as the time span between two consecutive A points. The time delays of the B, C, D and Q points, as well as the ratios of B, C and D point delays to TPT were added to the list of extracted features. The time span indices are presented in Fig. 4b.



(a) Amplitudes.

(b) Time span indices.

Figure 4: Indices in PPG waveform.

Other features There are other features which were studied, including the area under the PPG curve and the slope of the rising edge (mAB).

2.2.3 Features related to pulse propagation

The pulse arrival time (PAT) was defined as the time span between the R-peak of the electrocardiogram (ECG) and the Q point in the PPG waveform of the same heart cycle, as show in Fig. 5. We preferred to choose the inflection point (Q) instead of the systolic peak (B) because the exact determination of the former one is more robust compared to the rounded systolic peak and it is less affected by artifacts. The multiplicative inverse of the PAT was also included in the list of analyzed features.

The distance between two inflection points (Q), identified in the same heart cycle but in two different PPG signals recorded on different parts of the arm, are referred as pulse transit time (PTT) as illustrated in Fig. 5. The difference between the PAT and PTT is that the PAT has two components (i.e. pre-ejection period (PEP) and vascular transit time (TT)), while the PTT consist only of the TT.

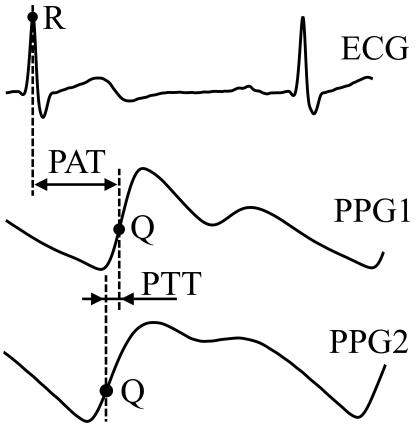


Figure 5: Definitions of PAT and PTT (a) ECG (b) fingertip PPG (c) PPG on upper-arm.

Once the characteristic points A, B, C, D and Q were identified, the presented features are calculated as follows:

$$B_{ampl} = X[B_i] - X[A_i] \tag{1}$$

$$B_{delay} = \frac{B_i - A_i}{F_S} \tag{2}$$

$$PAT = \frac{Q_i - R_i}{F_S} \tag{3}$$

$$Area = \frac{A_{i+1} - A_i}{2N} \sum_{n=1}^{N} (X[k] + X[k+1]) \tag{4}$$

where i is the number of the heart cycle, X represents the i^{th} segment of the PPG signal, N is the number of samples in the PPG curve, A_i, B_i, Q_i are the indexes of points A, B and Q in the PPG curve of the i^{th} heart cycle. The remainder of the amplitude and time span indices are calculated similarly.

2.3 Outlier removal

The PPG signal is affected by external factors (e.g. pressure between the skin and the sensor, movements), which can alter the morphology of the PPG waveform in a way that makes a correct segmentation impossible, corrupting extracted features. The feature sets reserved for the affected heart cycles need to be marked automatically as outliers and excluded from further processing. The outlier detection is based on inter-beat intervals (IBI), i.e. the time span between the same characteristic points of two consecutive heart cycles. If the morphology of the PPG waveform is altered, the difference between the IBIs, extracted from PPG and ECG for the same heart cycle, is considerably larger than measurement or processing errors.

In our application, we investigated the correlation between the extracted PPG features and BP variability during different activities. Therefore, higher frequency

variations (e.g. BP variations caused by the respiration) were filtered out. The BP and each feature were interpolated from an irregular, beat-to-beat sampling to a constant 10 Hz sampling rate using shape-preserving piecewise cubic interpolation. The resulting signals were than filtered with a 4th order low-pass digital Butterworth filter with a cutoff frequency of 0.02 Hz.

2.4 Feature synchronization

Wearable physiological sensors usually do not enable easy connection to clinically accepted monitoring systems for synchronous data acquisition. Therefore, heterogeneous data, resulted from studies which involve data collection using this type of devices, needs to be synchronized at a later stage.

In studies of cardiac signals, the most trivial parameter for synchronization is the beat-by-beat IBI data. In the same heart cycle the interval between two heartbeats should be the same, regardless of the measurement type. This value can be given directly by the device itself (e.g. blood pressure monitor) or it can be extracted from both ECG and PPG signals. If the resulted IBI signals are not affected by erroneous samples, that means there is only a time delay between them, which can be determined using cross correlation approach (Eq. 5).

$$n_{delay} = \arg\min_{n} \left(\sum_{i=0}^{N-1} f[i]g[i+n] \right) \tag{5}$$

where $N \in \mathbb{N}$ represents the length of the IBI time series.

However, signals collected during activities of daily life are often affected by noise and artifacts which prevent the correct segmentation of all heart cycles. In such cases, real IBI samples, with an error between them less than a fixed threshold, need to be matched, while outliers need to be eliminated. We used the extended Longest Common Subsequence (LCSS) method, described by Vlachos et al. [6], to match the values that are within a certain range in space and time, unmatched samples being most likely outliers.

3 Framework evaluation

The framework was evaluated in a feasibility study of unobtrusive blood pressure (BP) monitoring, using wearable sensors and a reference device. Within cardiovascular diseases (CVD), hypertension is one of the leading risk factors and causes 9.4 million deaths a year worldwide [7]. Currently available cuff-based and invasive instruments cause discomfort for the user and are inconvenient for long-term and continuous use. The ability to monitor BP in an unobtrusive way using signals recorded by wearables would be essential for an early diagnosis of hypertension.

In the last decade the relationship between the PPG and BP has been widely researched, and it was shown that the pulse arrival time (PAT) is highly correlated to BP changes [8, 9, 10, 11]. Wong et al. [12] studied furthermore the correlation between the SBP and the two components of the PAT: the pre-ejection period (PEP) and vascular transit time (TT) and showed that the PEP is more correlated to the SBP compared to TT. Beyond the PAT, there are other features of the PPG signal that are likely to be correlated with BP [13].

In an exploratory BP monitoring study, we used a multi-signal device to record an ECG and multiple PPG signals at different wavelengths and different locations on the body. BP variations were induced through activities e.g. relaxation, breathing exercise and physical exercise. The Portapres (Finapres, Netherlands) portable monitoring

device was used to record the reference BP. The Portapres uses two finger cuffs in alternation to provide beat-by-beat systolic, diastolic and mean BP as well as IBI. We used our framework to aggregate, process and extract features from the acquired signals. An example output is the PAT that shows a high correlation with the corresponding beat to beat systolic BP (Fig. 6).

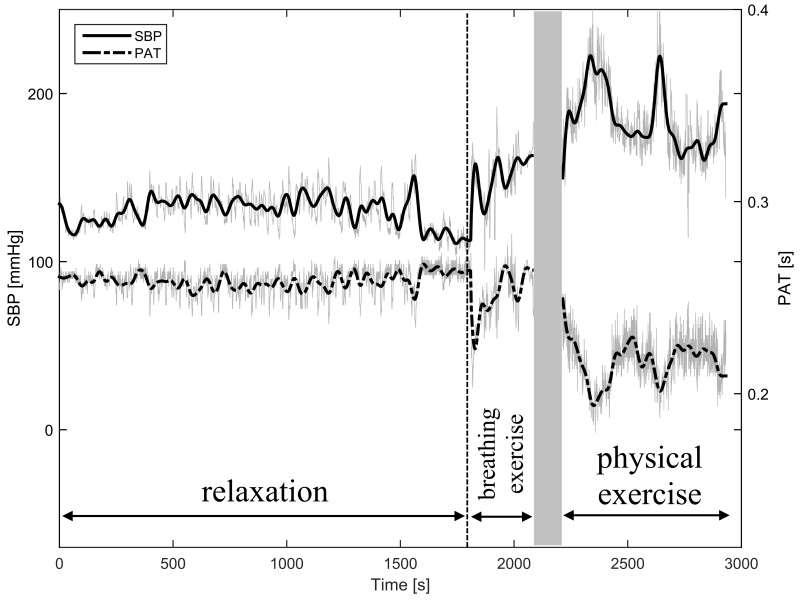


Figure 6: Fluctuation of SBP and PAT PPG feature of a subject during relaxation, breathing and physical workload.

Features extracted using our framework can facilitate research in unobtrusive BP monitoring and other applications of cardiac signals acquired by wearable sensors. According to our observations during this short evaluation, the PPG feature with the highest correlation to the BP can alter from subject-to-subject, which suggests the necessity of an adaptive feature selection module. Moreover, the framework can be further extend with modules that may improve feature extraction reliability (e.g. data provided by the accelerometer incorporated in devices).

4 Conclusion

Emerging wearable technologies have a huge potential in cardiac signal processing, but the personal health application development for wearables requires extensive research and studies that involve processing of large amount of data. In this paper we presented an extendable, modular framework for cardiac signal processing, focusing on a PPG processing module. The proposed framework facilities the research of wearable applications and it was evaluated in a study of unobtrusive BP monitoring.

Acknowledgments

This contribution was partially supported by the UEFISCDI (Romania), as part of the HEART project, contract nr. 130/2012.

References

[1] A. Reisner, P. Shaltis, D. McCombie, and H. Asada, "Utility of the photoplethysmogram in circulatory monitoring," Anesthesiology, vol. 108, no. 5, p. 950, 2008.

[2] P. Hamilton, "Open source ECG analysis," in Computers in Cardiology, pp. 101104, 2002.

[3] M. Elgendi, B. Eskofier, S. Dokos, and D. Abbott, "Revisiting QRS Detection Methodologies for Portable, Wearable, Battery-Operated, and Wireless ECG Systems," PLoS ONE, vol. 9, no. 1, p. e84018, Jan. 2014.

[4] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," Physiol. Meas., vol. 28, no. 3, pp. R1R39, Mar. 2007.

[5] M. Elgendi, "On the Analysis of Fingertip Photoplethysmogram Signals," Curr. Cardiol. Rev., vol. 8, no. 1, pp. 1425, Feb. 2012.

[6] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multidimensional time-series with support for multiple distance measures," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 216225, 2003.

[7] S. S. Lim, et al., "A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 19902010: a systematic analysis for the Global Burden of Disease Study 2010," Lancet, vol. 380, no. 9859, pp. 22242260, Dec. 2012.

[8] S. Ye, G.-R. Kim, D.-K. Jung, S. Baik, and G. Jeon, "Estimation of systolic and diastolic pressure using the pulse transit time," World Acad. Sci. Eng. Technol., vol. 67, pp. 726731, 2010.

[9] P. Fung, G. Dumont, C. Ries, C. Mott, and M. Ansermino, "Continuous noninvasive blood pressure measurement by pulse transit time," in Engineering in Medicine and Biology Society, IEMBS04. 26th Annual International Conference of the IEEE, vol. 1, pp. 738741, 2004.

[10] J. Lass, K. Meigas, D. Karai, R. Kattai, J. Kaik, and M. Rossmann, "Continuous blood pressure monitoring during exercise using pulse wave transit time measurement," in Engineering in Medicine and Biology Society, IEMBS04. 26th Annual International Conference of the IEEE, vol. 1, pp. 22392242, 2004.

[11] T. Ma and Y. T. Zhang, "A correlation study on the variabilities in pulse transit time, blood pressure, and heart rate recorded simultaneously from healthy subjects," in Engineering in Medicine and Biology Society, IEEE-EMBS 2005. 27th Annual International Conference of the, pp. 996999, 2005.

[12] M. Y. M. Wong, E. Pickwell-MacPherson, Y. T. Zhang, and J. C. Y. Cheng, "The effects of pre-ejection period on post-exercise systolic blood pressure estimation using the pulse arrival time technique," Eur. J. Appl. Physiol., vol. 111, no. 1, pp. 135144, Jan. 2011.

[13] Y. Li, Z. Wang, L. Zhang, X. Yang, and J. Song, "Characters available in photoplethysmogram for blood pressure estimation: beyond the pulse transit time," Australas. Phys. Eng. Sci. Med., vol. 37, no. 2, pp. 367376, Jun. 2014.

Proof of the Median Paths

Thijs Veugen
TNO
Technical Sciences
The Hague, The Netherlands
thijs.veugen@tno.nl

Abstract

We consider the problem of coding for discrete memoryless channels with noiseless feedback. When studying Horstein's sequential coding scheme, Schalkwijk in 1971 found a regular behaviour of the so-called median paths for certain channel error probabilities, which led to the development of repetition strategies. We proof that Schalkwijk's block decoding scheme exactly follows Horstein's regular median paths.

1 Introduction

We consider the problem of coding for discrete memoryless channels with noiseless feedback [4], which is strongly related to Ulam's game [3, p. 281]. In this game, person A knows a secret integer m, which person B should guess by constructing sets one by one, and asking whether m is in it or not. Person A is allowed to lie a fixed number of times, and B wants to minimise the number of required questions.

During the game, person B is maintaining a probability distribution on the possible integers of A. We will show that the median of this distribution plays a very important role. In 1963, Horstein [1] developed a sequential coding scheme for the binary symmetric channel with noiseless feedback, by closely following the median of the receiver's distribution. Schalkwijk [2] in 1971 found out that, for certain channel error probabilities, these medians exhibit a regular pattern, which led to the development of a simple block coding scheme, achieving channel capacity for those particular channel error probabilities.

However, this regular behaviour of the median, visualised by median paths, has never been proven. In this paper we present a proof, which has been known since 1997, but has never been published.

2 Median paths

In 1963, Horstein [1] developed a sequential coding scheme for the binary symmetric channel with noiseless feedback. Let p, $0 \leq p < \frac{1}{2}$, be the error probability of the binary symmetric channel, and $q = 1 - p$. Suppose the sender has an (infinite) binary message, presented as a point m in the message interval $[0, 1]$, which he would like to send towards the receiver. Assuming each message being equally likely, the initial message distribution of the receiver will be a uniform distribution over the interval $[0, 1]$. The idea of Horstein was as follows. He introduced a parameter a, $0 < a < 1$. Before transmitting the n^{th} bit, the sender will compute the currently expected message distribution, from the viewpoint of the receiver, and a message point m_n, such that $\Pr(m \geq m_n) = a$. In case $m \geq m_n$, the sender will transmit a 1, and a 0, otherwise. Because of the assumed noiseless feedback link, the sender learns which output bits have been received by the receiver, which enables him to compute the expected message distribution.

In case the receiver just got a 1, the message points in the $[m_n, 1]$ message interval become more likely, and the message points in the $[0, m_n)$ interval become less likely. The channel output distribution $\underline{\pi}$ can be computed as $\pi_0 = aq + (1-a)p$ and $\pi_1 = (1-a)q + ap$, and thereby the transmission rate

$$R(a) = I(X;Y) = H(Y) - H(Y|X) = h(\pi_1) - h(p),$$

where I represents mutual information, H information entropy, and h binary entropy.

From this equation, we derive that the transmission rate is maximised by setting $a = \frac{1}{2}$, because it leads to a uniform output distribution. The maximal transmission rate equals channel capacity $1 - h(p)$. This value of a corresponds with the *median* of all message points. In this case, after receiving a 1, the probability density of all messages in the upper interval $[m_n, 1]$ increases with a factor $\frac{q}{\pi_1} = 2q$, while in the lower interval $[0, m_n)$, the probability density will decrease with a factor $\frac{p}{\pi_0} = 2p$.

Figure 1 depicts an example of a median tree, showing all possible medians during 7 transmissions, given a binary channel with error probability $p = 0.1$. All medians have different values, and their pattern seems to be irregular. In 1971, Schalkwijk [2] discovered that, for certain channel error probabilities, specific median values keep recurring, and the consecutive medians follow a more regular pattern. An example of a regular median tree, for error probability $p = (3 - \sqrt{5})/4$, is depicted in Figure 2.

As explained earlier, after receiving a 1, the median moves up, and goes down, otherwise. The medians from Figure 2 have an additional property, namely that after receiving 100 or 011, the median returns to exactly the same value as three transmissions earlier. Not very surprisingly, this occurs when $2p \cdot 2q \cdot 2q = 1$, which explains the value $p = (3 - \sqrt{5})/4$. Schalkwijk observed that a receipt of 1000 leads to the same median as after receiving 0, which could be interpreted as an erroneously transmitted 0, followed by three additional 0's trying to correct this transmission error. This led to the birth of *repetition* strategies [2, 4].

3 Proof

For each integer value of k, $k \geq 3$, regular median trees can be found, where the median returns to its original value after receiving 10^{k-1} or 01^{k-1}. This occurs exactly for the solution p of

$$(2p) \cdot (2q)^{k-1} = 1.$$

In the corresponding repetition block-coding strategy, the sender is transmitting a fixed-length message, which does not contain the subsequences 10^k and 01^k. Each time a transmission error occurs, which can be detected because of the noiseless feedback link, the sender repeats the erroneously received symbol k times. This is applied in a recursive way, so a transmission error during a repetition sequence leads to k additional symbols to be sent.

The question is: why does an 'error correction', implemented as substituting, from right to left, any subsequence 10^k or 01^k in the received sequence by 0 or 1, respectively, not affect the values of the corresponding medians?

Because of the way (regular) median trees are constructed, we have the following **Property**:
Suppose we have an arbitrary receiver's distribution [1] to start with. Let m_i, $0 \leq i \leq k+1$, be the medians after receiving the first i symbols of the sequence 10^k. By definition of Horstein's scheme (assume the median goes up when receiving 1), it follows that $m_0 < m_1$, $m_1 > m_2 > \ldots > m_{k+1}$, and $m_k = m_0$. Furthermore, the receiver's distribution after transmission k is equal to the original distribution, when restricted to the interval $[0, m_0]$.

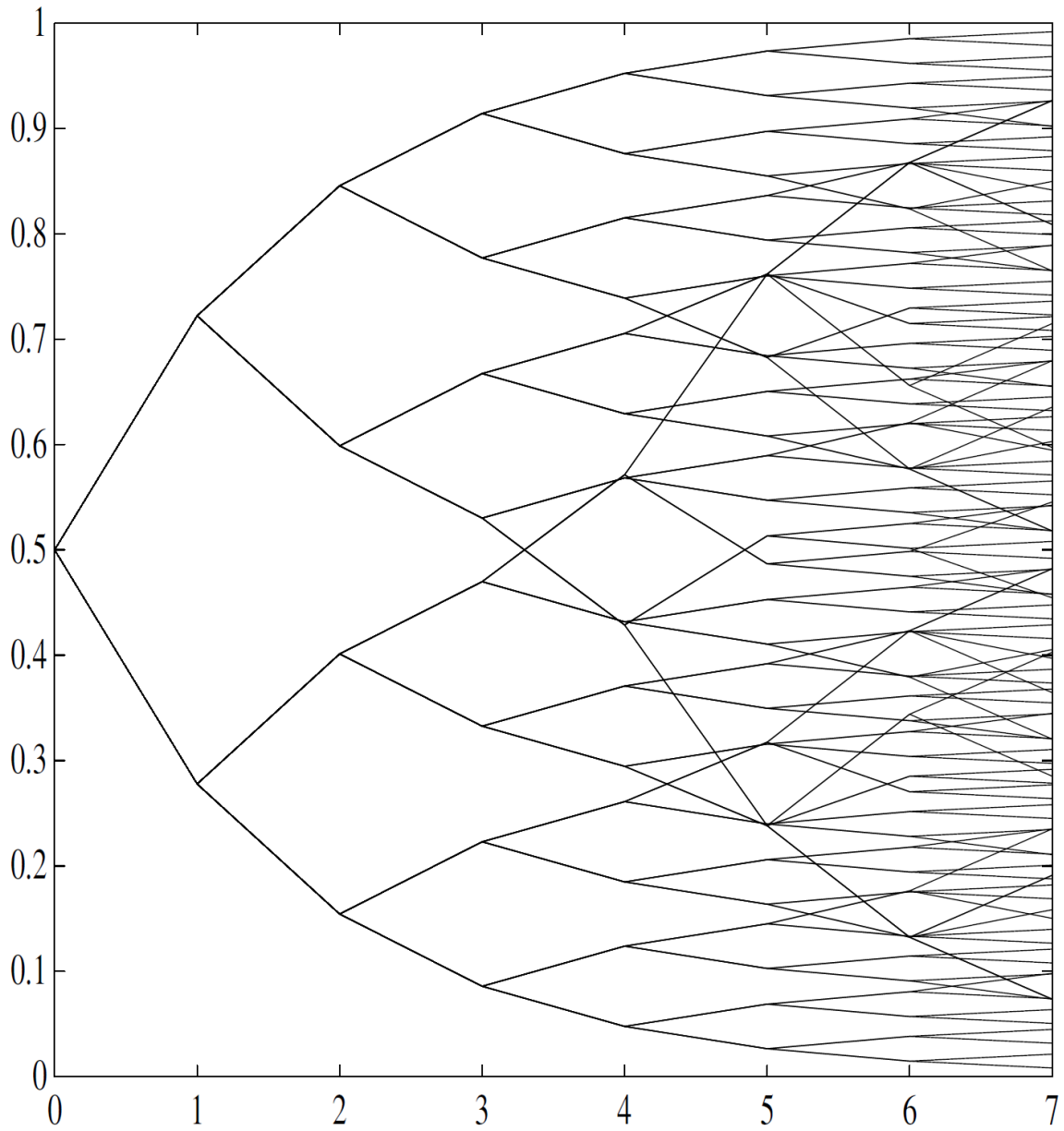


Figure 1: Median tree for $p = 0.1$

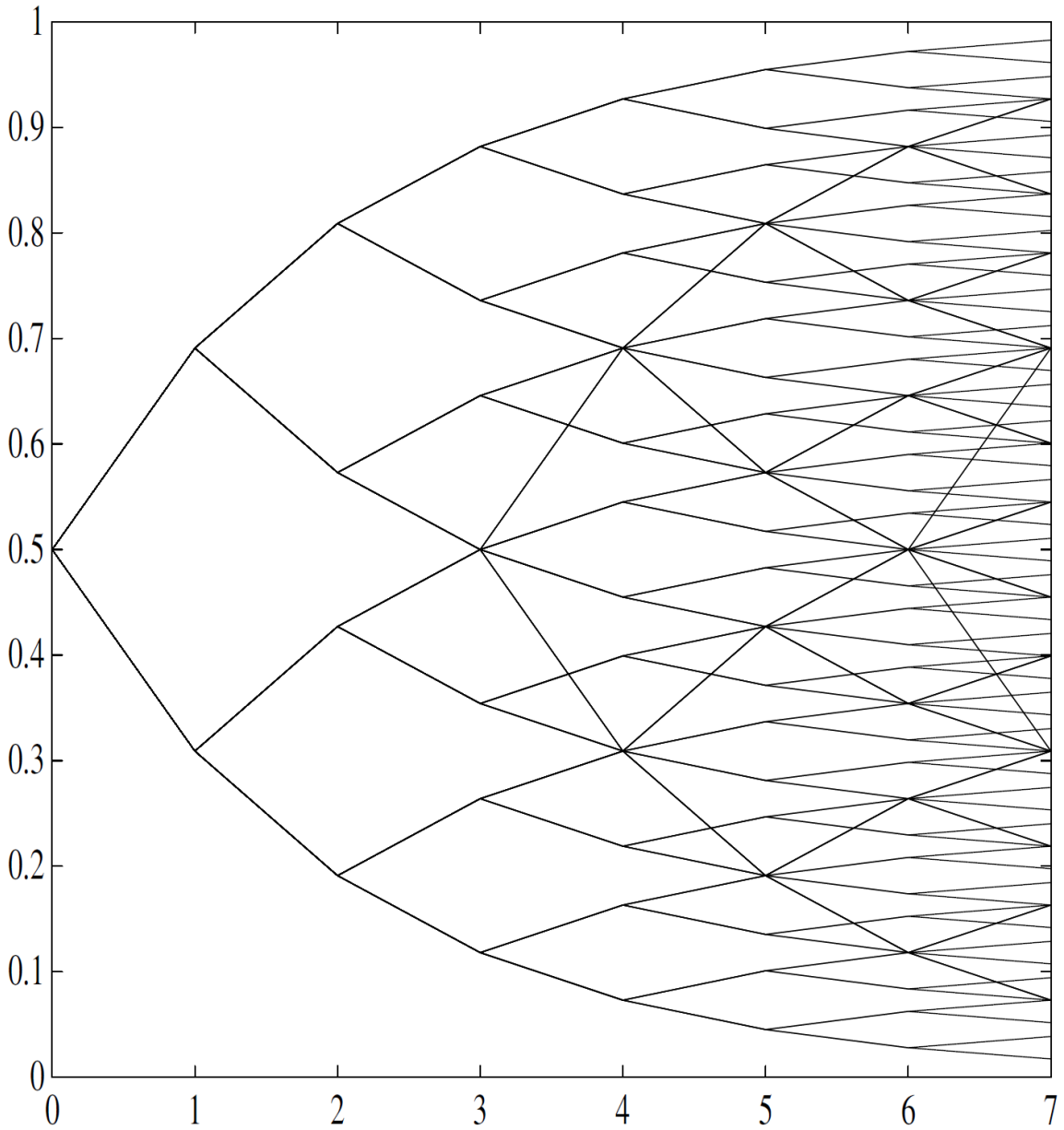


Figure 2: Median tree for $p = (3 - \sqrt{5})/4$

Using this property, we can proof the following lemma.

Lemma 1. *Let $\underline{x} = x_1 \dots x_N$ be an arbitrary received sequence without forbidden subsequences. Let m_i, $0 \leq i \leq N$, be the medians after receiving $x_1 \dots x_i$. If $x_1 = 0$, then $m_i \leq m_0$ for $0 \leq i \leq N$.*

Proof: Assume $x_1 = 0$. We proof by natural induction $m_i \leq m_0$ for $0 \leq i \leq N$. The basis is $n = 0$. The induction step is as follows. Suppose that for some n, $n \geq 0$, we have $m_i \leq m_0$ for all i, $0 \leq i < n$. Assume $m_n > m_0$. Since $m_{n-1} \leq m_0 < m_n$, it follows that $x_n = 1$. If $x_{n-1} = 0$, then $m_{n-2} \geq m_n$, according to the property above, which would contradict $m_{n-2} \leq m_0$, so $x_{n-1} = 1$. Similarly, it is shown that $x_{n-2} = 1$, and $x_{n-3} = 1$, etc., until $x_{n-k+1} = x_{n-k+2} = \dots = x_n = 1$. Now, because $\underline{x}$ contains no forbidden subsequences, we conclude $x_{n-k} = x_{n-k-1} = \dots x_1 = 1$, which contradicts our first assumption $x_1 = 0$. So it must be that $m_n \leq m_0$. □

This lemma is used in the proof of our main theorem, which answers the research question.

Theorem 1. *Let $\underline{x} = x_1 \dots x_N$ be an arbitrary received sequence. Let m_i, $0 \leq i \leq N$, be the medians after receiving $x_1 \dots x_i$. Let $x_n \dots x_{n+k}$ be the rightmost forbidden subsequence. Let $\underline{x'} = x'_1 \dots x'_{N-k}$ be the sequence after substitution of $x_n \dots x_{n+k}$ by x_{n+k} in $\underline{x}$. Let m'_i, $0 \leq i \leq N-k$, be the medians after receiving $x'_1 \dots x'_i$. Then*

$$m'_i = m_{i+k}, n \leq i \leq N-k.$$

Proof: Since $x_1 \dots x_{n-1} = x'_1 \dots x'_{n-1}$, it follows that $m_{n-1} = m'_{n-1}$. Furthermore, by the previous property, $m_{n+k-1} = m_{n-1}$. Assume w.l.o.g. $x_n = 1$. Since $x_n \dots x_{n+k}$ is a forbidden subsequence, $x_{n+k} = 0$. Because $x_{n+k} \dots x_N$ is a sequence without forbidden subsequences, it follows from our Lemma that $m_i \leq m_{n+k-1}$ for $n+k-1 \leq i \leq N$. Due to our property, the receiver's distribution after receipt of x_{n+k-1} is equal to the receiver's distribution after receiving x_{n-1}, when restricted to the interval $[0, m_{n+k-1}]$. Since each median after transmission $n+k-1$ lies in this interval, it is easily shown that $m'_i = m_{i+k}$ for $i = n-1, n, \dots, N-k$, respectively. □

The theorem explains that substituting the rightmost forbidden subsequence does not affect the corresponding median values. Therefore, the entire decoding process as a whole will not affect the median values of the original message (assuming all transmission errors have been corrected). This affirms that Schalkwijks repetition block coding strategies naturally fit into Horstein's sequential coding scheme.

4 Conclusion

We introduced Horstein's sequential coding scheme for the binary symmetric channel with noiseless feedback. We showed the existence of regular median trees, which led Schalkwijk to the discovery of repetition block coding strategies. Although at first sight, these repetition strategies seem to form a natural extension of Horstein's scheme, it is not straightforward to see that the behaviour of the medians is in line with the entire decoding process. We were able to prove this, affirming their conceptual coherence.

References

[1] Michael Horstein. Sequential transmission using noiseless feedback. IEEE Transactions on Information Theory, 9: 136-143, July 1963.

[2] J. Pieter M. Schalkwijk. A class of simple and optimal strategies for block coding on the binary symmetric channel with noiseless feedback. IEEE Transactions on Information Theory, 22(9): 1369-1373, May 1971.

[3] S.M. Ulam. Adventures of a mathematician. Charles Scribner's sons, 1976.

[4] Thijs Veugen. Multiple-repetition coding for channels with feedback. PhD thesis. Eindhoven, University of Technology. June 1997.

Enhancing privacy of users in eID schemes

Kris Shrishak
Cyber Security Group, Department of Intelligent Systems
Delft University of Technology, The Netherlands
k.s.sridaran@student.tudelft.nl

Zekeriya Erkin
z.erkin@tudelft.nl

Remco Schaar
UL Transaction Security
The Netherlands
remco.schaar@ul.com

Abstract

In todays world transactions are increasingly being performed over the Internet but require identification of users as in face-to-face transactions. In order to facilitate eGovernance as well as other eCommerce services Electronic Identification (eID) schemes, which intend to provide unique and reliable identification and authentication of the users, have been introduced. eID schemes commonly involve a Service Provider which provides a service, such as online shopping, to the user and an Identity Provider which verifies the users identity and facilitates the user to identify itself to the Service Provider. Every transaction made over the Internet reveals bits of information about the user which can be accumulated and abused, thus necessitating security and privacy in order to prevent misuse of data and invasion of personal privacy. In this work, five eID schemes which are in use or are proposed in EU countries is surveyed and the strengths and weaknesses of these schemes is investigated. All the schemes have given importance to security while only a few of them are designed with privacy in mind. Identity Providers in federated eID schemes are observed to be a privacy hotspot as they store user information and can uniquely identify the user. The use of homomorphic encryption and block chain in eID schemes is further explored in order to prevent the Identity Provider from becoming a privacy hotspot while fulfilling its role in the scheme.

1 Introduction

Long before the Internet came into existence, Governments have had public authentication schemes by issuing Identity documents to identify a person or verify aspects of a person's personal identity. These documents were trusted not only by the governments but also by businesses that required reliable authentication of users [1]. Today the number of services offered online is increasing rapidly and this growth has forced users to maintain multiple credentials for authentication and identification to service providers, which has caused security and usability issues. Password-based authentication mechanism employed by most service providers has lead to users reusing the same password or writing them on paper. Usage of other authentication mechanisms such as hardware security tokens is also not convenient. As a result, many countries in the European Union (EU) have either developed an electronic identification (eID) scheme or are in the process of developing one. All schemes have put security at the forefront but few have considered the privacy implications of a nation-wide single identification scheme. We consider it essential that designers of eID systems focus on privacy, specifically informational privacy of users which we interpret such that when a user mentions 'my information', the 'my'

> "is not the same as 'my' in 'my car' but rather the same as 'my' in 'my body' or 'my feelings'; it expresses a sense of constitutive belonging, not of external ownership, a sense in which my body, my feelings, and my information are part of me but are not my possessions [2]."

We survey five eID schemes in the EU and provide brief descriptions of two solutions which could be applied in eID schemes to improve privacy. Information privacy in eID schemes is investigated in terms of the following properties [3]:

1. Anonymity: Users may use a service without disclosing their identity.
2. Pseudonimity: Users may utilise a service by using pseudonyms.
3. Data minimization: Only the required information about the user must be shared in order to prevent misuse.
4. Unlinkability: User should be able to use resources and services without others being able to link these activities.
5. Unobservability: Users should be able to use services without being observed by others.
6. Transparency: User data should be obtained only when necessary and after user consent.

2 eID Schemes

eID systems intend to provide reliable identification and authentication of the users. The eID systems discussed in this section were designed for use by both public and private services. The parties involved in these systems differ widely but commonly involve the following:

- ◇ *User* - wants to authenticate her/himself to the Service Provider to access a resource.
- ◇ *Service Provider* (SP) - provides a service, such as online shopping or government tax services, and makes transaction decisions based upon the acceptance of a users authenticated credentials and attributes.
- ◇ *Identity Provider* (IDP) - verifies the user's identity or credentials and facilitates the user to authenticate her/himself to the SP. It improves the overall usability since the user does not need to remember multiple authentication credentials.

2.1 Belgian eID scheme

The Belgian eID scheme is a nation-wide Public Key Infrastructure (PKI) which requires each citizen to present her/himself at the municipality for strong user authentication during the issuing phase. Thus it can be inferred that the Belgian government has taken up the role of IDP. The user is issued a smart card and is required to buy a card reader for online use. The objective of the Belgian eID card has been to fulfil four functions, namely, citizen identification, authentication, digital signature and access control [4].

The eID card has the name, title, nationality, place and date of birth, gender, and a photo of its holder printed on it in addition to a hand written signature of its holder and of the civil servant who issued the card. All this information is also stored on the chip in an Identity file which is signed by the National Register (RRN). The chip also contains an address file which is kept independently as the address of its holder may change within the validity period of the card. The RRN signs the address file together with the identity file to guarantee the link between these two files. The corresponding signature is stored as the address file's signature.

Privacy Analysis

The main concern with the Belgium eID card is privacy. The user's identity is revealed for all transactions and (s)he does not have control over the data that is shared with the SPs while it remains possible for colluding SPs to link user activities as the authentication certificate contains the RRN. Verhaeghe et. al [5] provide further privacy and security threats if the card or the middleware is compromised.

2.2 GOV.UK Verify

GOV.UK Verify is the eID scheme of the United Kingdom. It is a federated identification infrastructure where an online central hub mediates user authentications between IDPs and SPs. The role of the hub is to ensure interoperable identification and authentication as well as provide privacy benefits by hiding the IDP from the SP. The eID scheme has been designed considering nine Identity Assurance Principles - user control, transparency, multiplicity, data minimization, data quality, service user access and portability, certification, dispute resolution and exceptional circumstances [6]. Users of the eID scheme are identified by a pseudonym ($\boldsymbol{u}$) at the hub. In addition to IDPs, SPs and hub, the scheme also includes (1) Attribute Providers (ATP), which are responsible for establishing attributes and (2) Matching Service (MS), which helps validate assertions from IDPs and derives a pseudonym $\boldsymbol{v}$ from $\boldsymbol{u}$ for the SPs it serves [7].

Privacy Analysis

In spite of claiming that privacy is one of the design criteria for the eID scheme, GOV.UK Verify has multiple privacy issues. The hub, which has full visibility of the user pseudonym and personal information of citizens can link interactions of the same user across different SPs as well as undetectably impersonate users at any SP without user authentication. The MS, which has the task of matching pseudonyms and the attributes into a local account, can link the user and the SPs that choose the same MS. Colluding SPs using the same MS can link user activities as the same pseudonym is used. Thus it can be inferred that GOV.UK Verify actually degrades the privacy of citizens instead of enhancing it. Finally, the hub can be used for undetectable mass surveillance.

2.3 Dutch eID scheme using polymorphic pseudonyms

The Dutch eID scheme utilises a federated eID infrastructure. Some variations of the Dutch eID scheme have been proposed during the planning stage. In the following, the version with polymorphic pseudonyms [8] is considered as it is the most privacy friendly version among those proposed. In addition to IDPs and SPs, the scheme also includes (1) a Pseudonym Provider (PP), which generates polymorphic pseudonym when the IDP sends a unique identifier($U\text{-}id$) (2) brokers, which mediate user authentication between the various eID parties, essentially IDPs and the many SPs. Other parties include Attribute Providers (ATP) and Authorization Providers, a Key Management Authority (KMA) and a investigation authority. The pseudonymization in this scheme involves three levels of pseudonyms, namely, (1) polymorphic pseudonyms (2) encrypted pseudonyms and (3) pseudonyms.

Privacy Analysis

The Dutch eID scheme with polymorphic pseudonyms has few of the privacy properties mentioned in section 1. It provides pseudonimity to the users such that SPs get a user specific pseudonym from the IDP, but the pseudonym is independent of the IDP. Meanwhile, the pseudonym obtained by different SPs for the same user is not the same, thus preventing colluding SPs from being able to link user activities. Finally, encrypted pseudonyms are randomized such that the broker cannot identify the same user on multiple occasions.

In spite of efforts made to provide privacy to the users, this scheme has countable issues. IDP is a privacy hotspot which knows the attributes of users, which SPs are visited by users and how often. This information might be considered very sensitive in some cases. The IDP does not need to know this information in order to perform its role. An alternative proposed in [8] is to store polymorphic pseudonyms in a smart card and use chip authentication as in the German eID scheme. The Dutch eID scheme is complex and it is possible that in the future SPs might outsource the decryption of encrypted pseudonyms to a third party allowing them to learn users visiting patterns.

2.4 German eID scheme

The German eID scheme uses a direct authentication eID infrastructure. The design goals of this scheme were data minimization, data security and transparency. In order to use this scheme, the user needs an eID card, a reader and the Ausweisapp software while the SP needs an authorization certificate, an eID-Server that handles authentication by communicating with the card. The eID card contains a chip which stores the information printed on the card as well as the fingerprints of the holder, if the holder wishes. The document number and the fingerprints can be read offline only by authorities who have machines certified by the Federal Office for Information Security (BSI).

The German eID scheme makes use of cryptographic protocols, to perform mutual identification, which are also used in EU passports [9]. Password Authenticated Connection Establishment (PACE) protocol provides secure communication and explicit password-based authentication of the eID card and the terminal while the Extended Access Control (EAC) protocol provides secure key establishment between a chip card and a terminal, using a PKI. It serves the purpose of limiting access to the sensitive data stored on the chip card. EAC comprises of terminal authentication and chip authentication. Finally Restricted Identification (RI) protocol generates a sector-specific identifier for each card, enabling the pseudonymous identification of the card-holder.

Privacy Analysis

The German eID scheme provides pseudonimity to the users and reduces sharing of excessive data with SPs. For instance, to verify if the age of the user is above 18, only a yes/no is sent instead of the age. By using secure channels, user data is not observable in transit. However the security of eID authentication relies on the tamper-resistance of the smart card chips. If an attacker manages to extract the chip authentication key from any eID card, then this attacker would be able to forge arbitrary identities. Also, user data that is transmitted after selective disclosure by the user does not contain any signature in order to verify if it is indeed the data that was originally issued by BSI and sent by a legitimate eID cardholder [1]. Only the context of the EAC protocols run and the secure channel thus established assure the eID-Server of the authenticity of the eID data. Finally, The eID-Server is recommended to be implemented by the SP but can also be implemented by third parties, thus creating a privacy risk as it serves more than one SP and can learn the visiting patterns of users as the attributes that are revealed by the users are identifying. Until 2012, eID servers implemented by only two third parties were being used by the SPs [10].

2.5 IRMA (I Reveal My Attributes)

IRMA (I Reveal My Attributes) is a attribute-based credentials (ABC) eID scheme. The credential issuer or the IDP issues credentials to the user and vouches for the validity of the attributes contained in the credential. After issuing the credentials, the IDP cannot recognise the credential as it is signed using blind signatures. This eliminates the possibility of the issuer tracking the card owner. The user can use the credentials to prove the possession of an attribute to the SP. Two important technologies that make use of an ABC approach are Microsoft's U-Prove [11] and IBM's Identity Mixer (Idemix) [12]. Idemix is built upon the concept of Camenisch-Lysyankaya signature scheme and its protocols [13]. IRMA is a partial implementation of Idemix that demonstrates the applicability of ABCs on smart cards [14]. The implementation includes privacy enhancing features of ABCs such as selective disclosure of attributes using zero-knowledge protocols.

Privacy Analysis

Users can perform transactions anonymously for transactions that do not need an identity but can be completed if the credentials are satisfactory. IRMA provides issuer

unlinkability as issuance involves creating a blind signature which conceals the resulting credential from the IDP. The user is also guaranteed that when a credential is verified multiple times by a SP, these sessions cannot be linked. By selective disclosure, the user can choose to reveal only a selection of the attributes and has greater control.

IRMA is certainly the most privacy friendly scheme discussed in this document. There still remains few issues that need to be addressed. If a card is lost or stolen, the lack of revocation procedure in the current implementation allows the possibility of misuse of credentials till it expires. A revocation scheme [15] for IRMA has been recently proposed but it introduces a new problem. The procedure uses a revocation value which is encoded by the credential such that the credential can be identified when it is revoked. Thus weakening the unlinkability argument.

A comparison of eID schemes in terms of their privacy properties is provided in Table 1.

Table 1: Comparison of eID schemes

Privacy Properties	eID Schemes				
	Belgian	UK	Dutch	German	IRMA
Anonymity / Pseudonimity		✓	✓	✓	✓
Data minimization				✓	✓
Unlinkability				✓	✓
Unobservability			✓	✓	✓
Transparency			✓	✓	✓

3 Privacy Enhancing Solutions

In section 2, a number of privacy issues in some of the existing/proposed eID schemes were identified. In this section, we briefly discuss two possible solutions to address the issue of privacy hotspots in federated eID systems.

3.1 Homomorphic Encryption for Privacy

Homomorphic encryption is a form of encryption which allows processing of ciphertexts and generate an encrypted result which, when decrypted, matches the result of operations performed on the plaintexts. Rivest, Adleman, and Dertouzos proposed the idea of homomorphic cryptosystems in their 1978 paper [16] and soon partially homomorphic cryptosystems such as unpadded RSA [17], ElGamal [18] and Paillier [19] were proposed. But it was only in 2009 that the first fully homomorphic encryption (FHE) scheme was proposed by Gentry [20]. The earliest FHE scheme which was based on ideal lattices was not suitable for practical implementation as it was computationally expensive and the ciphertext sizes were also very large [21]. Variants of the scheme based on different hardness assumptions such as Learning With errors (LWE) [22] or Ring-LWE [23] and integer-based or approximate Greatest Common Divisor (GCD) problem [24], also turned out to be impractical as the noise contained in the ciphertexts could not be managed after certain number of homomorphic operations and required an expensive bootstrapping step to refresh the ciphertext. But real world applications do not need to be able to handle all circuits. Thus a leveled homomorphic encryption scheme which can handle a circuit of low depth is sufficient. Many optimizations such as modulus switching, tensoring and re-linearization [25] have been proposed to make these schemes more practical.

The property of homomorphic encryption which allows computation on encrypted data can be utilised in federated eID schemes. In the Dutch eID scheme using polymorphic pseudonyms, the multiplicative homomorphic property of ElGamal cryptosystem

is utilised to transform pseudonyms. A similar approach could be used for attributes as well. But multiplicative homomorphism may not be sufficient if we want to prevent IDPs from having direct access to user data during authentication. Leveled homomorphic encryption schemes may find application in this scenario.

3.2 Block chain for Decentralization

Block chain was introduced by Satoshi Nakamoto as a timestamp server as part of the Bitcoin protocol [26]. A block chain is a public ledger shared by all nodes participating in a system based on the Bitcoin protocol [27]. A full copy of a block chain contains every transaction ever executed. Every block contains a hash of the previous block such that a chain of blocks is created from the first block of the chain, also known as genesis block, to the current block. This way the blocks are arranged in chronological order. It is also computationally infeasible to modify a block as every block that follows must also be regenerated.

Block chain allows to eliminate the necessity of a central party. But it introduces additional issues. A public ledger provides everyone in the network access to the data on the block chain. This means that instead of one central party having access to all the data, now all parties in the network have access. So block chain as it is, cannot be used to store private data and hence does not address the problem of privacy but merely shifts the problem. Even though Bitcoin protocol allows the usage of multiple pseudonyms or public keys, it is possible to link the activity of users [28]. Enigma [29], a decentralized computation platform which allows storage and computation of private data, uses off-chains in addition to the block chain to store private data. A similar approach has been used in [30] but it requires a minimally trusted manager.

As a primitive idea we propose that block chain could be used in federated eID systems by using off-chain consisting of IDPs. User data can be split up among the IDPs thus preventing any one IDP from having complete information about its users, thus decentralizing the role of IDP. This ideas requires further research and refinement before it can be considered for nation-scale eID systems.

4 Conclusion

We have surveyed five eID schemes in the EU and analysed their privacy properties. Belgian, GOV.UK and Dutch schemes use the approach of identifying entities by a unique number at the IDP. This approach is convenient for bookkeeping but it is not privacy friendly. Unique identifiers can be used to trace the user and her/his activities. IDPs have been identified as a privacy hotspot due to the large trove of user data that they store and process, which allows them to link user activities. Finally, we proposed two possible solutions - homomorphic encryption and block chain - to allow the IDP to perform its role of authenticating users to SPs without becoming privacy hotspots.

References

[1] A. Poller, U. Waldmann, S. Vowe, and S. Türpe, "Electronic identity cards for user authentication - promise and practice," *IEEE Security & Privacy*, vol. 10, no. 1, pp. 46–54, 2012.

[2] L. Floridi, "The ontological interpretation of informational privacy," *Ethics and Information Technology*, vol. 7, no. 4, pp. 185–200, 2005.

[3] ISO15408-2:2005, "Information technology - security techniques - evaluation criteria for it security - part 2: Security functional requirements," tech. rep., International Standard Organization, 2005.

[4] D. D. Cock, C. Wolf, and B. Preneel, "The belgian electronic identity card (overview)," in *Sicherheit*, vol. 77 of *LNI*, pp. 298–301, GI, 2006.

[5] P. Verhaeghe, J. Lapon, B. D. Decker, V. Naessens, and K. Verslype, "Security and privacy improvements for the belgian eid technology," in *SEC*, vol. 297 of *IFIP Advances in Information and Communication Technology*, pp. 237–247, Springer, 2009.

[6] "Identity assurance principles," tech. rep. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/361496/PCAG_IDA_Principles_3.1__4_.pdf.

[7] L. T. A. N. Brandão, N. Christin, G. Danezis, and anonymous, "Toward mending two nation-scale brokered identification systems," *PoPETs*, vol. 2015, no. 2, pp. 135–155, 2015.

[8] "Polymorphic pseudonymization." Available at https://www.idensys.nl/fileadmin/bestanden/idensys/documenten/basisdocumentatie/documentatieset/PP_Scheme_091.pdf.

[9] BSI, "Advanced security mechanisms for machine readable travel documents and eIDAS token part 2 protocols for electronic identification, authentication and trust services (eIDAS)," tr-03110-2, Bundesamt fur Sicherheit in der Informationstechnik, 2015.

[10] R. Bjones, I. Krontiris, P. Paillier, and K. Rannenberg, "Integrating anonymous credentials with eids for privacy-respecting online authentication," in *APF*, vol. 8319 of *Lecture Notes in Computer Science*, pp. 111–124, Springer, 2012.

[11] S. Brands, *Rethinking Public Key Infrastructures and Digital Certificates: Building in Privacy*. MIT Press, 2000.

[12] IBM Research Zurich Security team, "Specification of the identity mixer cryptographic library," tech. rep., IBM Research, April 2010.

[13] J. Camenisch and A. Lysyanskaya, "An efficient system for non-transferable anonymous credentials with optional anonymity revocation," in *EUROCRYPT*, vol. 2045 of *Lecture Notes in Computer Science*, pp. 93–118, Springer, 2001.

[14] P. Vullers and G. Alpár, "Efficient selective disclosure on smart cards using idemix," in *IDMAN*, vol. 396 of *IFIP Advances in Information and Communication Technology*, pp. 53–67, Springer, 2013.

[15] W. Lueks, G. Alpár, J. Hoepman, and P. Vullers, "Fast revocation of attribute-based credentials for both users and verifiers," in *SEC*, vol. 455 of *IFIP Advances in Information and Communication Technology*, pp. 463–478, Springer, 2015.

[16] R. L. Rivest, L. Adleman, and M. L. Dertouzos, "On data banks and privacy homomorphisms," in *Foundations of Secure Computation* (R. A. DeMillo, D. P. Dobkin, A. K. Jones, and R. J. Lipton, eds.), pp. 165–179, Academic Press.

[17] R. L. Rivest, A. Shamir, and L. M. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Commun. ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[18] T. E. Gamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," in *CRYPTO*, vol. 196 of *Lecture Notes in Computer Science*, pp. 10–18, Springer, 1984.

[19] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *EUROCRYPT*, vol. 1592 of *Lecture Notes in Computer Science*, pp. 223–238, Springer, 1999.

[20] C. Gentry, *A fully homomorphic encryption scheme*. PhD thesis, Stanford University, 2009.

[21] C. Gentry and S. Halevi, "Implementing gentry's fully-homomorphic encryption scheme," in *EUROCRYPT*, vol. 6632 of *Lecture Notes in Computer Science*, pp. 129–148, Springer, 2011.

[22] Z. Brakerski and V. Vaikuntanathan, "Efficient fully homomorphic encryption from (standard) LWE," in *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pp. 97–106.

[23] Z. Brakerski and V. Vaikuntanathan, "Fully homomorphic encryption from ring-lwe and security for key dependent messages," in *CRYPTO*, vol. 6841 of *Lecture Notes in Computer Science*, pp. 505–524, Springer, 2011.

[24] M. van Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, "Fully homomorphic encryption over the integers," in *EUROCRYPT*, vol. 6110 of *Lecture Notes in Computer Science*, pp. 24–43, Springer, 2010.

[25] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, "(leveled) fully homomorphic encryption without bootstrapping," 2012.

[26] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008. https://bitcoin.org/bitcoin.pdf.

[27] "Blockchain." https://en.bitcoin.it/wiki/Block_chain.

[28] E. Androulaki, G. Karame, M. Roeschlin, T. Scherer, and S. Capkun, "Evaluating user privacy in bitcoin," in *Financial Cryptography*, vol. 7859 of *Lecture Notes in Computer Science*, pp. 34–51, Springer, 2013.

[29] G. Zyskind, O. Nathan, and A. Pentland, "Enigma: Decentralized computation platform with guaranteed privacy," *CoRR*, vol. abs/1506.03471, 2015.

[30] A. E. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The blockchain model of cryptography and privacy-preserving smart contracts," *IACR Cryptology ePrint Archive*, vol. 2015, p. 675, 2015.

A Privacy-Preserving GWAS Computation with Homomorphic Encryption.

Chibuike Ugwuoke Zekeriya Erkin Reginald L. Lagendijk
Delft University of Technology
Department of Intelligent Systems, Cyber Security Group.
Delft, The Netherlands
C.I.Ugwuoke@tudelft.nl Z.Erkin@tudelft.nl R.L.Lagendijk@tudelft.nl

Abstract

The continuous decline in the cost of DNA sequencing has contributed both positive and negative feelings in the academia and research community. It has now become possible to harvest large amounts of genetic data, which researches believe their study will help improve preventive and personalised healthcare, better understanding of diseases and response to treatments. However, there are more information embedded in genes than are currently understood, just as a genomic data contains information of not just the owner, but relatives who might not subscribe to sharing them. Unrestricted access to genomic data can be privacy invasive, hence the urgent need to regulate access to them and develop protocols that would allow privacy-preserving techniques in both computations and analysis that involve these very sensitive data. In this work, we discuss how a careful combination of cryptographic primitives such as homomorphic encryption, can be used to privately implement common algorithms peculiar to genome-wide association studies (GWAS). This obviously comes at a cost, where we have to accommodate the trade-off between speed of computations and privacy.

1 Introduction

Biomedical research has long shown that human genome contain data from which information about their individual owners, and those related to them can be extracted [1, 2, 3, 4]. A lot of privacy-sensitive information are laced all over genomic data, which constitutes enormous worry for individuals whose data are available in electronic format [5, 6, 7]. The benefits of continuous research involving the genomic data are equally rife, these include: preventive and personalised healthcare, patient's response to treatment, predisposition to diseases, identification of new drug targets and perhaps a better understanding of cancer [1, 8, 4, 9, 10, 11, 12, 13, 14]. On the other hand, when genomic data is used for research or processed by medical personnels, they become exposed to possible misuse and even loss to unauthorised hands. In the face of this possibility, the risk of re-identifying individuals from an available genomic data calls for serious concern [5, 9, 15, 4, 3, 7], and has been recognised as a realistic threat. Other unwanted scenarios which could occur as direct consequence of leaking genomic data include: stigmatisation, discrimination, loss of insurance and even loss of employment opportunities for persons whose genomic data is public [16].

What is more worrisome about misuse of genomic data is the fact that the genome has longevity, when leaked, it can neither be revoked nor modified. So, it is obvious that this piece of data is highly sensitive and requires protection that should be adaptive to future security threats. Hence one can claim that any realistic solution should be one which, the security guarantees of the underlying primitives used for implementation

should withstand post quantum attacks. Therefore privacy protection techniques have been proposed as an adaptive solution by the cryptography community. The aim will be to allow productive research that utilise genomic data, while eliminating the privacy-risks inherent around these procedures.

Being that no standalone solution can best fit the challenge posed, it is considered that a good combination of *ethical, legal* and *technological* constraints can be employed, to properly manage the risks of privacy leaks that are otherwise possible within this research domain. Owing to this premise, our work seeks to contribute a technological solution to the underlying problem.

In the era of distributed computing, even the medical field has not been left out. It has been common for researchers and medical personnels to work without boundaries of country borders, albeit, via a virtual collaboration [17, 7, 2]. This means that more data can now be shared for research purposes and even diagnosis of diseases [6]. It also presents us with the possibility of allowing cloud services process medical data, even when they do not reside in the same country as the owners of the data. This need for collaboration, data sharing and cloud processing of genomic data further pushes for privacy-preserving secure computing protocols [2, 17].

Having a genomic dataset and controlling access to it is the main aim of this work. In a nutshell, this means that while these data is not available to the public, experts who need them for research are granted restricted access to only subsets relevant to their work [15]. Such access for processing data may include string searching and comparison, as well as GWAS computations.

Genome Wide Association Studies: As highlighted in [1, 18], the first ever human genome sequencing was achievable in 2001, after directly gulping a whooping US-$300 million from the initial budget of US-$3 billion. Fast-forward 6 years later, and the same feat is feasible for about US-$100,000. In 2006 [18], it was anticipated that in 2014, a further reduction to US-$1,000 was possible for sequencing the human genome. Recent literature [19, 11] have even suggested that a meagre US-$100, will be a reality in the very near future. If that be the case, one can deduce that amongst other possibilities, a direct consequences of affordable genomic data would be the torrential flow of genomic data *in silico*. It is obviously a good development for researchers, who would heavily rely on these data to improve on their research, refine and optimise diagnosis and many others positive possibilites. With a wealth of data in the form of genomic data lying at the disposal of researchers and medical personnels, learning and inferring from these data becomes an indisputable objective.

Without loss of generality in description, GWAS can simply be simplified to the activities presented above, it is about gathering genetic data, processing them and relying on them to investigate relationship (association) of genes to common known diseases. It will be possible to even detect unknown diseases and the effect of drugs on treatments. With GWAS researchers can now measure, analyse and predict previously unknown genetic influence on a person, this can help in early detection and prevention of certain diseases, as well as personalised healthcare. For useful gene-disease associations to be estimated, some computations become handy, and these will be discussed in subsection 2.1. Nonetheless, most of the computations can easily put the data owners at privacy-risk. It has led to the suggestion that protection of genomic data is a necessity, to address possible ethical, political, technological and privacy concerns. From the technological solution approach, we hope to address the privacy-threats using cryptographic primitives. Just to mention, with genomic data, data anonymization is not enough guarantee to avoid re-identification and also, conventional encryption might not offer much better protection against envisaged privacy-threats. These can simply be derived from the fact that the said data have longevity, their importance persists even after the demise of the data owner.

Related Works: Realising the privacy-sensitive nature of genomic data, researchers

have delved into search for privacy-preserving solutions, in the hope to protect privacy of owners while still being able to process and compute operations using these data. Some of these works are discussed here. Privacy-preserving GWAS spans across more possibilities than just GWAS-Computations. According to [15], other important categories include:

- Private string searching and comparison.
- Private release of aggregated data.
- Private read mapping.

of course, this list is not in itself exhaustive, but we will only consider works that directly address computations very peculiar to GWAS. As early as 1999 [20, 21, 22], some researchers had anticipated privacy risks involved with genomic data. So they proposed denominalization and de-identification as protection schemes, to preserve privacy. This did not stop re-identification attacks from being hugely successful, as discussed in [9]. Other authors [23] have subsequently recommended *Trusted Third Parties* and *Semitrusted Third Parties* but then, it is not always easy to completely trust a third party, who could still be susceptible to coercion, compulsion and even corruption to be compromised. More recently in [24], attempts were made to analyse genomic data while avoiding privacy-invasion of participants of the data. Summarily, they adopted differential privacy as a privacy-preserving technique, and documented to have obtained utility with their procedure. However, addition of noise using differential privacy is not a silver bullet to deflate possible re-identification. Especially when the published data can be augmented with other side information. But most importantly is the fact that differential privacy contains noise, which will evidently affect the utility, no matter the degree of noise. This is a huge trade-off, but it is only left for the geneticists and biostatisticians to decide if the noise only contributes a negligible disturbance to the final results.

While the last paper approach to resolving possible privacy breaches is via differential privacy, [25] chooses to adopt a different approach. The authors adopt homomorphic encryption as a tool to enable analysis of these privacy sensitive data. Homomorphic Encryption holds a lot of promises, and if its capabilities are optimally harnessed, can become a very productive primitive in guaranteeing privacy for processing genomic data. In this work, different scenarios are considered which include a setting that allows outsourcing encrypted genomic data to a cloud service. In the mentioned scenario, operations on the data by the cloud are still possible, without divulging the decryption keys but still hopeful of achieving utility.

Homomorphic Encryption was further relied on by some other team of researchers [26]. A shot was given to providing privacy guarantees on processing of genomic data, only that this time the focus was on homomorphic encryption scheme whose structures rely on RLWE (Ring Learning With Error). [26] documents an efficiency-improvement from existing implementation of GWAS using homomorphic encryption. They showed that χ^2 test for independence was achievable with improvement in both computation and communication time from existing implementations.

Subsequently, another team of researchers went further to demonstrate how much information can be extracted from computation of genomic data, even on the encrypted domain [27]. Basic genomic algorithms which are common to GWAS are shown to be implementable on encrypted genotype and phenotype data. Lauter et al. [27] report results that preserve utility of the original implementation (computation on unencrypted genomic data). Some of the algorithms demonstrated in their work include:

- Estimation Maximization (EM) algorithm for haplotyping.
- The $D^{'}$ and r^2-measures of linkage disequilibrium.

- Cochran-Armitage Test for Trend.

Also worth mentioning is the fact that this implementation relied on Homomorphic Encryption with assumption on RLWE.
Scenario and Assumptions: For the sake of this work, we will explicitly spell out the scenario in which our proposed protocol is targeted, and necessary assumptions. Our setting adopts the semi-honest security model, hence we assume that all parties will correctly follow the protocol by performing the right computations, but with a curiosity to observe the transitions of the protocol with a view to learning more details than they are statutorily allowed to learn. We assume that a researcher *Alice* is interested in a particular computation, say *Minor Allele Frequency (MAF)*. The data source or cloud *Bob*, who happens to have the computational powers not acquired by *Alice*, is trusted to perform all requests by performing the computation on encrypted data. The result of the computation (which however, is also encrypted), is returned to *Alice*.

2 Preliminaries

Up until here, we have established a clear direction to the challenge we hope to address. A genomic dataset is at our disposal and we intend to preserve privacy of data in the face of effective computations. So, we propose a protocol that encrypts all genomic data and outsources storage of these data to a semi-honest cloud service who possesses the computational requirements to run these expensive computations. It will be pertinent to have a mental picture of typical algorithms that will be deployed to perform computation, and how our cryptographic privacy enhancing technology optimally fits for a solution. Most of the algorithms are statistical operations that are often required by biostatisticians when trying to learn information from a dataset. And just like most statistical equations require simple arithmetic operation at the least, we show that our adopted primitive (homomorphic encryption), does provide us with the capabilities to perform simple *addition*, *multiplication*, and with a little more effort *division*.

2.1 GWAS Computation

Only a few statistical computations that are usually handy in GWAS are presented.

Minor Allele Frequency: Finding the ratio for which an allele of interest that is at a locus, occurs in a particular population of study is the allele frequency. *MAF* is therefore the allele frequency of the least common *allele*, which appears in that population. If we have a gene with two possible alleles say **A** and **S**, then in a monoploid gene setting, the allele frequency *f()* for **A** is simply computed as follow:

$$f(\mathbf{A}) = \frac{\sum_1^n \mathbf{AA}}{\sum_1^n \mathbf{AA} + \sum_1^m \mathbf{SS}} \tag{1}$$

where $N = n + m$ is the total population sample, and n and m are the counts of alleles **A** and **S** respectively. That was rather too easy, owing to the fact that we only have two possible genotypes, which are results of pure combination of possible alleles. What happens when we consider diploid gene settings? Using the same alleles at a particular locus, we consider the following expressions: **AA**, **AS** and **SS**. Just like we did above, we shall try to compute the frequency of the allele **A**. Let genotype distribution be as follows: $\mathbf{A} = 19, \mathbf{AS} = 21, \mathbf{SS} = 07$.

$$f(\mathbf{A}) = \frac{2 * \sum_1^n \mathbf{AA} + \sum_1^k \mathbf{AS}}{2(\sum_1^n \mathbf{AA} + \sum_1^k \mathbf{AS} + \sum_1^m \mathbf{SS})} \tag{2}$$

The total genotype count in this case is $N = n + k + m$, where n, k and m are counts for **AA**, **AS** and **SS** respectively. to compute the allele frequency of **A** using the values already

presented, we will have

$$f(\mathbf{A}) = \frac{2 * \mathbf{AA} + \mathbf{AS}}{2 * (\mathbf{AA} + \mathbf{AS} + \mathbf{SS})} = \frac{2 * 19 + 21}{2 * (19 + 21 + 07)} = \frac{59}{94} = 0.6277 \tag{3}$$

Since we only have two possible alleles in this population, the least common allele should be **S**, with MAF of $(1 - 0.6277) = 0.3723$

To calculate the genotype frequencies we have $\mathbf{AA} = \frac{n}{N}$, $\mathbf{AS} = \frac{k}{N}$, $\mathbf{SS} = \frac{m}{N}$

Linkage Disequilibrium: This is the non-random association of alleles at different loci. Unlike the single locus alleles considered previously, we will be considering two loci but mainly retaining the basic statistics we have developed thus far. The aim of this test is to suggest if SNPs at particular loci of interest behave or occur in such a manner that is not believed to be random. So we present two loci with the following alleles: **A**, **a** and **S**, **s**. When two genotype at different loci are independent of each other, Linkage Equilibrium is considered to have occurred. Simply put, this means that Linkage Disequilibrium happens when there is some degree of dependency between the two loci of interest. Leading to the Hardy-Weinberg Equilibrium (HWE), which is said to hold if allele frequencies are preserved in a population across generations, except otherwise altered by an external factor, including evolutional influences. To measure linkage disequilibrium, the following equations are used to compute $D^{'}$ and r^2.

$$D^{'} = \begin{cases} \frac{f(AS)f(as) - f(As)f(aS)}{min(f(A)f(s), f(a)f(S))} & \text{if } f(AS)f(aa) - f(As)f(aS) > 0 \\ \\ \frac{f(AS)f(as) - f(As)f(aS)}{min(f(A)f(S), f(a)f(s))} & \text{if } f(AS)f(aa) - f(As)f(aS) < 0 \end{cases} \tag{4}$$

$$r^2 = \frac{(f(AS)f(as) - f(As)f(aS))^2}{f(A)f(S)f(a)f(s)} \tag{5}$$

On the assumption that the allele frequencies can be obtained from encrypted genomic data, then it follows that the above computations can be computed.

2.2 Homomorphic Encryption

Homomorphic Encryption (HE) is a cryptographic primitive that allows for simple arithmetic operations over a ciphertext space. A HE scheme can either allow for simple addition, multiplication or even both. We have an additive scheme if it can only allow for addition operations and a *fully homomorphic encryption* (FHE) scheme if both addition and multiplication can be harnessed from the scheme. Give two messages m_1 and m_2, an encryption and decryption functions $Enc()$ and $Dec()$ respectively. We have that:

$$Enc(m_1) \oplus Enc(m_2) \rightarrow Enc(m_1 + m_2) : Dec(Enc(m_1 + m_2)) := (m_1 + m_2) \tag{6}$$

$$Enc(m_1) \otimes Enc(m_2) \rightarrow Enc(m_1 * m_2) : Dec(Enc(m_1 * m_2)) := (m_1 * m_2) \tag{7}$$

In 2009, [28] proposed an FHE scheme, which reduced its security to some well known difficult lattice problem. Further works were done to improve the original scheme, due to the complexity involved in implementation. Bringing about works like [29, 30, 31], which have been able to present a levelled homomorphic encryption scheme, that is capable of handling multiplication to a certain degree or depth, before the ciphertext becomes un-decryptable. The main idea is that for every operation, some noise is added to the ciphertext, and when this noise grows above a certain threshold, decryption of the ciphertext becomes a problem. While *addition* contributes small degree of noise, *multiplication* allows the noise to grow very fast. These schemes often reduce their security to lattice based problems like *shortest vector problem* (SVP) including *ring learning with error problems* (RLWE). Because the

multiplication function obtainable from these homomorphic encryptions are not arbitrary (as to control the noise growth), it is labelled *levelled* or *somewhat homomorphic encryption* (SHE). To show that the multiplication depth can only go as deep as the specified level, during parameter setup.

With a SHE scheme handy, and statistical algorithms available, we can then deploy this primitive to solve the arithmetic operations we identified earlier. It can be demonstrated that with SHE, these algorithms can be computed while preserving the utility and not trading privacy of the genomic data concerned.

3 Privacy Preserving χ^2 Statistic

In GWAS computation, X^2 is often computed and compared to the χ^2 distribution. A common test can be applied to know if the HWE holds in a given distribution. An example of a computation is presented below:

$$X^2 = \sum_{i=\{AA,AS,SS\}} \frac{(O_i - E_i)^2}{E_i} \tag{8}$$

O_i and E_i represent observed frequency allele and Expected frequency allele of the population. Since the frequency allele can easily be computed by simple addition and multiplication, and the required arithmetic operations are obtainable in our discussed Homomorphic Encryption. It can be concluded that the χ^2 statistics can be computed in a privacy-preserving manner, over encrypted datasets. Other computations such as the Cochran-Armitage Test for Trend can equally be computed using this procedure, and even meta-analysis of data from different experiments can be produced as well. For simplicity, we shall show how X^2 test statistic can be computed, borrowing the suggestions in [27, 26, 32], with a subtle modification. Every SNP representation is assumed to belong to a genotype classification. And for a single locus test, we produce 3 encryptions, $Enc(x)_{c,d} : x \in \{0, 1\}$, c and d are row and column indexes respectively. The rows depict the SNPs for participants, while the columns depict genotype (AA, AS, SS). Assuming that all loci representation correctly fall into a genotype class, then the summation of the row values $\sum_{c=1}^{N}$ will produce n, k and m, recall that $N = n + k + m$. It then becomes feasible to calculate the sum of the genotypes by simply adding the encrypted values for each column. This will require a constant cost of $3N$ numbers of additions using homomorphic encryption.

$$X^2 = \frac{(n - E_{AA})^2}{E_{AA}} + \frac{(k - E_{AS})^2}{E_{AS}} + \frac{(m - E_{SS})^2}{E_{SS}} = \frac{(n - E_{AA})^2 * E_b * E_c + (k - E_{AS})^2 * E_a * E_c + (n - E_{SS})^2 * E_a * E_c}{E_{AA} * E_{AS} * E_{SS}} \tag{9}$$

Again, to compute the X^2 test statistic, it becomes evident that this computation will require at least, $(3N + 5)$ *additions*, 14 *multiplications* and a single *division*. We deliberately ignore the computation of $E_{i=\{AA,AS,SS\}}$, since those can be easily precomputed and stored. But if we have a (2×2) or (2×3) contingency table as presented in [32], we can still show that these complex looking computations can be reduced to *additions*, *multiplications*, and a single *division*. Since our SHE scheme can perform *addition* and *multiplication* efficiently, we are left to show that a trivial non-cryptographically secure means can be used to efficiently carry out the division. We offer this trivial solution, with the knowledge that a cryptographically secure division will involve a multiparty computation, of which we do not wish to discuss, due to the complexities involve. The non-trivial solution would be as follows:

$$\frac{Enc(x)}{Enc(y)}, \quad r \leftarrow \mathbb{R}, \quad \frac{Enc(x) \otimes Enc(r)}{Enc(y) \otimes Enc(r)} \tag{10}$$

Both numerator and denominator are presented to the researcher, who can decrypt them and perform the division in clear. The test statistic is therefore obtained and compared to the appropriate $p-value$ that was chosen, with 1 degree of freedom. The obtained result will not lose utility, and yet achieves a privacy guarantee on the semi-honest settings. The cloud

to whom data processing is outsourced, does not know what values are encrypted, but can perform operations using only the ciphertext, and the researcher who queries the database for X^2 value can be sure to obtain a correct value.

Complexity: The complexity of the proposed protocol can only be as efficient as the HE scheme deployed to solve the problem. For instance, when computing allele frequencies, several additions and a few multiplications are required. Which means that the computational complexity can be bounded by the computational complexity of the underlying HE scheme. However, if an additive HE scheme is to be deployed, we envisage an extra cost associated with communication. This is because multiplication in additive schemes are often performed as a multi-party computation (MPC). For the simple case of computing X^2, we have a cost of $3N + 5$ additions, 14 multiplications and 1 division. Which will involve many rounds of communication for an additive homomorphic encryption scheme.

For future work, we strongly recommend adoption of SHE scheme over an additive HE scheme like *Paillier* .We will attempt to address the issue of division over encrypted domain. This should be an important addition to this work, and perhaps one can leverage on that to perform even faster computations of statistical GWAS algorithms.

4 Conclusion

With major enhancement of the described cryptographic primitives, we foresee further deployment of privacy enhancing techniques to create protocols for processing of genomic data. We believe that this is an achievable feat in the near future, as to prepare for the bloat in availability of genomic data *in silico*. This protocol should be able to preserve the utility of results as obtainable in unencrypted data scenario and better than anonymised data implementation. Though the performance values will be expensive as a result of the encrypted data and encoding needed to be done, we believe that with further attention paid to this area of research, performance optimization is very realistic.

References

[1] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[2] M. M. Baig, J. Li, J. Liu, H. Wang, and J. Wang, *Privacy protection for genomic data: current techniques and challenges*. Springer, 2010.

[3] S. Jha, L. Kruger, and V. Shmatikov, "Towards practical privacy for genomic computation," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 216–230, IEEE, 2008.

[4] F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls, "Privacy-preserving matching of dna profiles.," *IACR Cryptology ePrint Archive*, vol. 2008, p. 203, 2008.

[5] C. J. Willer, Y. Li, and G. R. Abecasis, "Metal: fast and efficient meta-analysis of genomewide association scans," *Bioinformatics*, vol. 26, no. 17, pp. 2190–2191, 2010.

[6] W. Xie, M. Kantarcioglu, W. S. Bush, D. Crawford, J. C. Denny, R. Heatherly, and B. A. Malin, "Securema: protecting participant privacy in genetic association meta-analysis," *Bioinformatics*, p. btu561, 2014.

[7] J. Wagner, J. N. Paulson, X.-S. Wang, B. Bhattacharjee, and H. C. Bravo, "Privacy-preserving microbiome analysis using secure computation," *bioRxiv*, p. 025999, 2015.

[8] C. Cao and J. Moult, "Gwas and drug targets," *BMC Genomics.*, p. 15(Suppl 4):S5, 2014.

[9] B. A. Malin, "An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future," *Journal of the American Medical Informatics Association*, vol. 12, no. 1, pp. 28–34, 2005.

[10] J. Zhang, K. Jiang, L. Lv, H. Wang, Z. Shen, Z. Gao, B. Wang, Y. Yang, Y. Ye, and S. Wang, "Use of genome-wide association studies for cancer research and drug repositioning," *PloS one*, vol. 10, no. 3, p. e0116477, 2015.

[11] D. L. Selwood, "Beyond the hundred dollar genome–drug discovery futures," *Chemical biology & drug design*, vol. 81, no. 1, pp. 1–4, 2013.

[12] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9362–9367, 2009.

[13] F. Liu, A. Arias-Vásquez, K. Sleegers, Y. S. Aulchenko, M. Kayser, P. Sanchez-Juan, B.-J. Feng, A. M. Bertoli-Avella, J. van Swieten, T. I. Axenovich, *et al.*, "A genomewide screen for late-onset alzheimer disease in a genetically isolated dutch population," *The American Journal of Human Genetics*, vol. 81, no. 1, pp. 17–31, 2007.

[14] C. C. Spencer, Z. Su, P. Donnelly, and J. Marchini, "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip," *PLoS Genet*, vol. 5, no. 5, p. e1000477, 2009.

[15] E. Ayday, M. Humbert, J. Fellay, M. Laren, P. Jack, J. Rougemont, J. L. Raisaro, A. Telenti, and J.-P. Hubaux, "Protecting personal genome privacy: Solutions from information security," tech. rep., EPFL, 2012.

[16] Z. Lin, A. B. Owen, and R. B. Altman, "Genomic research and human subject privacy," *SCIENCE-NEW YORK THEN WASHINGTON-.*, pp. 183–183, 2004.

[17] M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin, "A cryptographic approach to securely share and query genomic sequences," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 12, no. 5, pp. 606–617, 2008.

[18] G. Church, "The race for the $1000 genome," *Science*, vol. 311, 2006.

[19] D. McMorrow, "The $100 genome: Implications for the dod," tech. rep., DTIC Document, 2010.

[20] D. Gaudet, S. Arsenault, C. Bélanger, T. Hudson, P. Perron, M. Bernard, and P. Hamet, "Procedure to protect confidentiality of familial data in community genetics and genomic research," *Clinical genetics*, vol. 55, no. 4, pp. 259–264, 1999.

[21] L. Burnett, K. Barlow-Stewart, A. Proos, and H. Aizenberg, "The" genetrustee": a universal identification system that ensures privacy and confidentiality for human genetic databases.," *Journal of Law and Medicine*, vol. 10, no. 4, pp. 506–513, 2003.

[22] J. E. Wylie and G. P. Mineau, "Biomedical databases: protecting privacy and promoting research," *Trends in biotechnology*, vol. 21, no. 3, pp. 113–116, 2003.

[23] G. De Moor, B. Claerhout, F. De Meyer, *et al.*, "Privacy enhancing techniques the key to secure communication and management of clinical and genomic data," *Methods Archive*, vol. 42, no. 2, pp. 148–153, 2003.

[24] A. Johnson and V. Shmatikov, "Privacy-preserving data exploration in genome-wide association studies," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1079–1087, ACM, 2013.

[25] J. W. Bos, K. Lauter, and M. Naehrig, "Private predictive analysis on encrypted medical data," *Journal of biomedical informatics*, vol. 50, pp. 234–243, 2014.

[26] W. Lu, Y. Yamada, and J. Sakuma, "Efficient secure outsourcing of genome-wide association studies," in *Security and Privacy Workshops (SPW), 2015 IEEE*, pp. 3–6, IEEE, 2015.

[27] K. Lauter, A. López-Alt, and M. Naehrig, "Private computation on encrypted genomic data," in *Progress in Cryptology-LATINCRYPT 2014*, pp. 3–27, Springer, 2014.

[28] C. Gentry, *A fully homomorphic encryption scheme*. PhD thesis, Stanford University, 2009.

[29] Z. Brakerski and V. Vaikuntanathan, "Fully homomorphic encryption from ring-lwe and security for key dependent messages," in *Advances in Cryptology–CRYPTO 2011*, pp. 505–524, Springer, 2011.

[30] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, "(leveled) fully homomorphic encryption without bootstrapping," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 309–325, ACM, 2012.

[31] M. Yagisawa, "Fully homomorphic encryption without bootstrapping.," *IACR Cryptology ePrint Archive*, vol. 2015, p. 474, 2015.

[32] L. Kamm, D. Bogdanov, S. Laur, and J. Vilo, "A new way to protect privacy in large-scale genome-wide association studies," *Bioinformatics*, vol. 29, no. 7, pp. 886–893, 2013.

Security Analysis of the Authentication of Classical Messages by an Arbitrated Quantum Scheme

Helena Bruyninckx Dirk Van Heule
Royal Military Academy, Brussels, Belgium
Avenue de la Renaissance, 30, B-1000, Brussels
helena.bruyninckx@rma.ac.be dirk.van.heule@rma.ac.be

Abstract

In this paper, we consider the quantum authentication protocol among two distrustful parties from [1]. This protocol offers authentication of classical bit sequences which are encoded into quantum ciphertexts by using quantum encryption. The protocol can be implemented with the same technology as what is needed for quantum key distribution. Since the classical message is publicly available, the security of the protocol is analyzed with respect to an adversary who can perform a known plaintext attack. By using her knowledge of the plaintext, the adversary can perform an optimal measurement on the quantum states that constitute the ciphertext.

1 Introduction

Quantum cryptography exploits quantum mechanical effects in order to perform cryptographic tasks. Information is typically encoded in quantum states, called quantum bits or qubits, and the security is mainly derived by exploiting the counter-intuitive properties of those quantum states. Quantum Key Distribution is the most well-known example of a quantum protocol, but researchers are also working on other topics in order to enlarge the application domain of quantum cryptography.

The subject of this paper will be a two-party quantum scheme for the authentication of classical messages. The scheme is based on symmetric key principles, and since the participants do not trust each other, a semi-trusted arbiter is present during the first phase in order to generate the non-repudiation token. The scheme can be implemented with the same technology as what is needed for quantum key distribution and uses photon polarization to encode the classical information. Since the scheme is symmetrical, the intended receiver can correctly measure the qubits and thus retrieve the classical information without errors. An adversary however needs to gain information about unknown qubits, which means that she needs to have a certain strategy to measure the qubits. By using her knowledge of the classical plaintext, which is publicly available, the adversary can perform an optimal measurement on the quantum states. In this paper, we will provide a first theoretical security analysis by analyzing the information gathered by an adversary.

Paper outline. The paper is organized as follows. In section 2, we introduce the terminology and concepts that are needed for this paper. The restrictions and assumptions for the security analysis is discussed in section 3. A brief description of the quantum authentication scheme with arbitration from [1] is given in section 4. We focus on the proposed construction and present its analysis in section 5. Finally, we draw our conclusions and discuss future work.

2 Preliminaries

2.1 Classical cryptography

One-way hash functions. Informally, a function f is called a *one-way function* if it is unfeasible to invert f. A *one-way hash function* is a one-way function $h : \{0,1\}^* \to \{0,1\}^l$ that takes as input a string of arbitrary length to produce an output string, called the *tag*, of fixed length, l.

Authentication. Authentication allows a receiver to verify that the transmitted message arrives without modification. This is usually performed by sending a message together with a tag t, where the tag is calculated from the message. Verification is done by recalculating the tag from the received message and comparing it with the received tag. The algorithm to generate the tag that will be considered in this paper, is based on *universal hashing*.

Universal hashing. We denote $\mathcal{M}$ the set of classical messages, $\mathcal{T}$ the set of tags and $\mathcal{K}$ the set of shared keys. Let $\mathcal{H} = \{h_k : \mathcal{M} \to \mathcal{T}\}_{k \in \mathcal{K}}$ be a *family of hash functions*. Each family is indexed by a n-bit key k and consists of 2^n hash functions mapping a message $m \in \mathcal{M}$ into a l-bit hash output. From this family, a hash function is chosen uniformly at random by means of the shared key $k \in \mathcal{K}$. An authentication scheme with key recycling offers information-theoretic security, if the hash function is selected from an $\epsilon-$*almost XOR universal*$_2$ ($\epsilon-$AXU$_2$) *family of hash functions* [3]. Alice appends $h_{k_1}(m) \oplus r$ to her message m, where k_1 is used for all messages and r is a one-time pad (OTP). Note that a key k chosen uniformly at random (i.e., $k \in_R \{0,1\}^n$) can efficiently select a hash function from the total family at random.

2.2 Quantum tools

Single and multiple quantum bit systems. Quantum bits, or *qubits*, can be seen as the quantum equivalent of classical bits and are described as two-dimensional normalized vectors in a complex Hilbert space: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad \alpha, \beta \in \mathbb{C} : |\alpha|^2 + |\beta|^2 = 1$. The conjugate transpose is denoted as $\langle\psi|$. The set $\{|0\rangle, |1\rangle\}$ denotes the orthonormal basis of the two-dimensional Hilbert space $\mathcal{H}_2$, and is also called the standard basis. The pair $\{|+\rangle, |-\rangle\}$ denotes the diagonal or Hadamard basis, where $|\pm\rangle = (|0\rangle \pm |-\rangle)/\sqrt{2}$. The standard and diagonal bases are two most commonly used *conjugate bases* and are referred to as the set $\{+, \times\}$.

Quantum systems consisting of two or more qubits are *composite systems*. The joint state of this system is given by the tensor product $|\psi_1\rangle \otimes \ldots \otimes |\psi_n\rangle$ and the set $\mathcal{S} = \mathcal{H}_2^{\otimes n}$ denotes all $n-$qubit quantum states. Suppose a quantum system is in one state $|\psi_i\rangle$ and that there are n possible pure states ($i = 1, \ldots, n$), where each state has probability p_i, $p_i \geq 0, \sum_{i=1}^n p_i = 1$. The entire system $\{|\psi_i\rangle, p_i\}$ has *density operator* ρ given by a positive semi-definite density matrix, $\rho = \sum_{i=1}^n p_i |\psi_i\rangle\langle\psi_i|$, where p_i denotes the eigenvalues of ρ. In a *completely mixed state*, the probability for the system to be in each state is identical. Therefore, considering a state space of n dimensions, we have $\rho = \frac{1}{n} I_n = \frac{1}{n} \sum_{i=1}^n |i\rangle\langle i|$.

Transformations. Performing transformations on quantum systems is captured in the description of a linear and unitary operator, U. An important single qubit transformation is the *Hadamard transformation*, which is used to create superposition states. Its matrix representation is given by

$$H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Measurements. We consider two special types of measurement. *Projective measurements* describe measurement outcomes consisting of mutually exclusive possibilities. Measuring a single qubit in the standard basis means applying the measurement described by $P_0 = |0\rangle\langle 0|$ and $P_1 = |1\rangle\langle 1|$. The index indicates which measurement outcome is considered. If only the probabilities of the different outcomes are of interest, then a *Positive Operator Value Measurement* (POVM) is useful.

Von Neumann entropy. The quantum analogue for determining the amount of entropy in a quantum state is the *Von Neumann entropy*. The Von Neumann entropy of a quantum state with density operator ρ is $S(\rho) = H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$, where H denotes the classical entropy function.

Conjugate coding. Wiesner developed the idea to encode classical information into quantum states, more specifically into conjugate bases [4]. This primitive is called *conjugate coding* and encodes each classical bit into a single quantum state. A classical message $m \in \{0,1\}^n$ can be encoded into a quantum string by applying this principle. The two participants agree on the set of conjugate bases, e.g. $\{+, \times\}$, and they share a classical key $k \in \{0,1\}^n$. If $k_i = 0$, m_i is encoded with the standard basis; otherwise, m_i is encoded with the Hadamard basis. Therefore, if $\theta \in \{+, \times\}^n$, we write $|m\rangle_\theta = |m_1\rangle_{\theta_1} \otimes \ldots \otimes |m_n\rangle_{\theta_n}$. The encryption is denoted as $|m\rangle = E_k(m)$ and an example of the encoding of classical bits by conjugate coding is illustrated in Table 1.

Table 1: Example of Conjugate Coding

classical bit, m_i	1	0	0	1	0	1	1	0	1
classical key, k_i	0	0	1	1	0	1	0	1	1
basis choice	+	+	×	×	+	×	+	×	×
quantum encoding	$\|1\rangle$	$\|0\rangle$	$\|+\rangle$	$\|-\rangle$	$\|0\rangle$	$\|-\rangle$	$\|1\rangle$	$\|+\rangle$	$\|-\rangle$

3 Restrictions

When analyzing the security of a cryptographic scheme, it is very important to take the following aspects into account: (1) the challenge for the adversary, (2) the power the adversary has, and (3) the assumptions on the cryptographic scheme. By considering that there is no restriction on adversarial resources such as computing power and memory size, we speak about *unconditional* (or, *information-theoretic*) security.

This unconditional security is extremely hard to obtain and therefore, we impose more realistic restrictions on the adversary. We consider that the adversary's quantum memory size is limited, meaning that she can only store a limited amount of qubits. The adversary needs thus to measure the intercepted qubits, consequently destroying the information. On the other hand, transmission and measurement of qubits is within reach of current technology. Thus, the honest parties, following the protocol, do not need quantum memory.

Other assumptions for the authentication scheme are related to a common source with initially shared randomness and the assumption that the scheme is realized with ideal components. Noise on the quantum channel can be solved with quantum error correcting codes.

4 Review of the quantum authentication scheme

The arbitrated quantum authentication scheme from [1] consists of different phases. We assume that Alice wants to send a message m to the receiver Bob.

During the **initialization phase**, all the necessary keys are prepared and distributed. Moreover, the participants agree on a one-way, collision resistant hash function and on a family of $\epsilon-\mathrm{AXU}_2$ hash functions. These choices are publicly known and remain unchanged during the complete execution of the scheme.

During the **setup phase**, a non-repudiation token for Alice is created by interaction between Alice and Trent. Alice sends her message m and other information to be authenticated together with a one-time ticket to the arbiter, i.e., $|\phi\rangle = E_{K_{k_2}}\left(K_A^i||H_{k_1}\left(m_p, K_A^i\right)\right)$. The data to be authenticated is $m_p = (ID_A, ID_B, m, i)$ and is publicly known (i is a counter). K_A^i is Alice's one-time ticket, calculated by Alice from the publicly known hash function selected during the initialization phase. ID_A and ID_B denote the identity of Alice and Bob, respectively and $||$ denotes concatenation. The arbiter verifies the received information and computes a non-repudiation token, R_T, over the received data which he sends to Alice.

During the **authentication phase**, Alice will create her authenticated message which Bob can verify to be authentic. Trent does not interact in this stage. In this phase, Alice and Bob use their shared key K_{AB}, unknown to Trent.

Each shared key consists of two parts, e.g. the shared key between Alice and Trent $K_{AT} = k_1||k_2$. The first part of the key will select uniformly at random a hash function from the family of $\epsilon-\mathrm{AXU}_2$ hash functions. This means that the participants can mutually agree on different hash functions, only known to two participants each time. This hash function is included in the scheme in order to detect manipulations by an adversary.

The second part of the shared key is used to perform quantum encryption (conjugate coding) on the classical messages. Ideally, this key must be renewed for each message that needs to be encrypted. This requirement is impractical and therefore, a pseudo random number generator (PRNG) is included in the scheme. The seed for this PRNG consists of the second part of the shared secret key, i.e., k_2. Figure 1 gives a graphical representation of the different components of the encryption part for the first message sent by Alice to Trent during the setup phase.

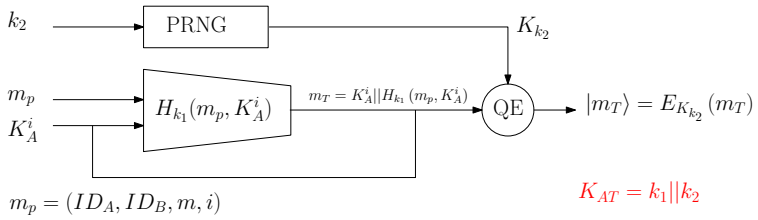


Figure 1: Construction for the encryption and authentication of a publicly known classical message m_p, by means of a shared key $K_{AT} = k_1||k_2$. The authentication tag of the message $m_p||K_A^i$ is computed by the hash function H_{k_1}, resulting in a n_1-bit tag. Then, encrypting the message $m_T \in \{0,1\}^{n_2}$ is performed by the quantum cipher with secret-key $K_{k_2} \in \{0,1\}^{n_2}$.

5 Analysis

We will analyze the security of this construction depicted in Figure 1. An external opponent (which could also be Bob) is considered. The goal of the adversary can consist of (partially) breaking the quantum cipher, and thus obtaining some information on the shared key. Moreover, since the hash key H_{k_1} is reused for several messages and the scheme is classified as an authentication scheme satisfying the non-repudiation property, her ultimate goal is to perform a *substitution attack*, that is to produce a valid message-authentication pair on a previously unseen message.

In [5], the authors introduce quantum ciphers with perfect security. Even if an adversary knows the plaintext and applies an optimal measurement on the quantum messages, her Shannon uncertainty on the key is almost maximal. In the construction from Figure 1, the encryption key is generated by a cryptographic PRNG. The quality of this generator determines the security of the scheme. In order to perform an attack, the adversary needs to know some information on the sequence of output bits produced by the generator.

The analysis of the scheme constitutes thus of several parts: (1) demonstrate that the principle of conjugate coding is a quantum cipher respecting the properties from [5, 6]; (2) demonstrate that attacking the PRNG isn't possible, due to the limited information an adversary can retrieve on the generated sequence; (3) include the fact that several messages are exchanged using the same hash function.

5.1 Conjugate coding is a quantum cipher with perfect security

Conjugate coding (see Subsection 2.2) can be considered as a quantum cipher to encrypt classical messages using classical keys and produces quantum states. The encryption and decription processes are modeled by unitary operations on the plaintext. Following [5], we give a formal definition of the conjugate coding cipher, denoted as C_n-cipher:

Definition 1. The C_n-cipher is an $(n,n)-$quantum cipher. Given a classical message $m \in \{0,1\}^n$ and a classical key $k \in \{0,1\}^n$, it outputs the following $n-$qubit state as

ciphertext:

$$|\phi\rangle = H^{k_1} \otimes H^{k_2} \otimes \ldots \otimes H^{k_n} |m_1 m_2 \ldots m_n\rangle ,$$

where H is the Hadamard transform.

This technique deals with each (qu)bit separately and is described as follows. Alice transforms the classical message she wants to transmit into an $n-$bit quantum state $|\psi\rangle$, consisting of $|0\rangle$ and $|1\rangle$ states, and described by ρ. Then, she applies to each qubit of this state the Hadamard transformation, only if the corresponding bit of the shared key equals 1. Afterwards, she sends the result $|\phi\rangle$ to Bob. Knowing the key k, Bob can decrypt the received qubits by reversing the Hadamard transformation, retrieving ρ.

The data hiding and key hiding properties of a general quantum cipher are satisfied. They guarantee that an adversary cannot obtain any information about the key and the message when an arbitrary ciphertext is seen [5]. Therefore, we require that Eve, who does not know the key, always obtains the same density matrix ρ_E by monitoring the channel. This property is captured in the scenario of a *Private Quantum Channel,* formalized as follows [6].

Definition 2. Let $\mathcal{S} \subseteq \mathcal{H}_{2^n}$ be a set of pure $n-$qubit states, $\mathcal{E} = \left\{\sqrt{p_i}U_i | 1 \leq i \leq n\right\}$ be a superoperator where each U_i is a unitary mapping on $\mathcal{H}_{2^n}$, $\sum_{i=1}^{n} p_i = 1$, and ρ_E be an $n-$bit density operator. Then $[\mathcal{S}, \mathcal{E}, \rho_E]$ is called a Private Quantum Channel (PQC) if and only if for all $|\psi\rangle \in \mathcal{S}$ we have

$$\mathcal{E}\left(|\psi\rangle\langle\psi|\right) = \sum_{i=1}^{n} p_i U_i \left(|\psi\rangle\langle\psi|\right) U_i^{\dagger} = \rho_E,$$

where $U_i^{\dagger}$ is the complex-transpose of U_i. $\mathcal{E}$ denotes the superoperator which applies U_i with probability p_i to its argument.

The C_n-cipher, with key $k \in \{0,1\}^n$, applies a Hadamard transform to each qubit individually if $k_i = 1$, otherwise, the qubit remains unchanged (denoted by applying the identity operator I_n). If $m_i = 0$, a uniform mixture of $|0\rangle$ and $|1\rangle$ is sent across the quantum channel, expressed by $\rho_0 = \frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}|+\rangle\langle +|$. Similarly, if $m_i = 1$, we obtain $\rho_1 = \frac{1}{2}|1\rangle\langle 1| + \frac{1}{2}|-\rangle\langle -|$. We use $\overline{H^k}$ to denote the $n-$bit unitary transformation $H^{k_1} \otimes H^{k_2} \otimes \ldots \otimes H^{k_n}$.

Theorem 3. *If* $\mathcal{H}_2 = \{\cos\theta|0\rangle + \sin\theta|1\rangle \,|\, \theta \in \{0, \pi/2\}\}$, $\mathcal{S} = \mathcal{H}_2^{\otimes n}$, $\mathcal{E} = \left\{\frac{1}{\sqrt{2^n}}\overline{H^k} | k \in \{0,1\}^n\right\}$ *and* $\rho_E = \frac{1}{2^n} I_{2^n}$, *then* $[\mathcal{S}, \mathcal{E}, \rho_E]$ *is a Private Quantum Channel.*

Proof. To each qubit $|\psi_i\rangle$, we apply the Hadamard transformation H or the identity operator, each with probability 1/2. This action puts each qubit in a completely mixed state. The superoperator $\mathcal{E}$ applies this strategy to each of the n qubits individually, hence $\mathcal{E}\left(|\psi\rangle\langle\psi|\right) = \frac{1}{2^n} I_{2^n}$ for every $|\psi\rangle \in \mathcal{H}_2^{\otimes n}$. □

We now calculate the Von Neumann entropy of the quantum state $|\phi\rangle$, i.e., $S(\rho)$. For the quantum state $|\psi\rangle$, which is the tensor product of n qubits, we have that $\mathcal{E}\left(|\psi\rangle\langle\psi|\right) = \frac{1}{2^n} I_{2^n}$. Therefore, we can calculate the Von Neumann entropy as follows:

$$S(\rho) = S\left(\frac{1}{2^n} I_{2^n}\right) = S\left(\frac{1}{2} I_2 \otimes \ldots \otimes \frac{1}{2} I_2\right) = nS\left(\frac{1}{2} I_2\right) = n\left(\frac{1}{2} + \frac{1}{2}\right) = n.$$

5.2 Including the authentication tag

In the construction from Figure 1, the message K_A^i and the authentication tag $H_{k_1}(m_p, K_A^i)$ are encrypted together by the C_n-cipher. The authentication tag was constructed by an $\epsilon-\mathrm{AXU}_2$ family of hash functions:

Definition 4. ([3]) A family of hash function $\mathcal{H} = \{h_k : \mathcal{M} \to \mathcal{T}\}_{k\in\mathcal{K}}$ for $\mathcal{T} = \{0,1\}^{n_1}$ is said to be $\epsilon-$almost XOR universal$_2$ if for k chosen uniformly at random and all distinct $m_1, m_2 \in \mathcal{M}$ and all $t \in \mathcal{T}$,

$$\Pr_k[h_k(m_1) \oplus h_k(m_2) = t] \leq \epsilon.$$

From this definition, we can derive some statements about $\epsilon-\mathrm{AXU}_2$ hash functions, and we only note the one that is useful for our discussion. The number of $\epsilon-\mathrm{AXU}_2$ hash functions taken from $\mathcal{H}$ that satisfies $h_k(m_1) \oplus h_k(m_2) = t$ is exactly $|\mathcal{H}| \,/\, |\mathcal{T}| = 2^{n-n_1}$, where $k \in \{0,1\}^n$.

5.3 Attack on PRNG

A pseudo-random generator is a deterministic function $G_n : \{0,1\}^k \to \{0,1\}^n$ that takes as input a seed of size k and produces a pseudo-random sequence of length $n(k)$, i.e., a polynomial in k. We require a cryptographic pseudo-random generator to withstand any statistical test that tries to distinguish its output from a truly random sequence (i.e., *distinguishing attacks*). In order to perform such attacks, the adversary needs to know a subsequence of generated bits from the generator. If the adversary can only access a limited number e of generated bits from the pseudo-random generator G_n, then there exists constructions for generators providing provable cryptographic security, even if the adversary has infinite computational resources. In [7], the authors show that there exist such generators if $e \leq k/\log_2 n$. If we use those constructions as suitable generators for the construction of Figure 1, we need to calculate the probability that an adversary can determine at most e bits from its output sequence.

Eve has access to the public message m_p and the quantum encoded version of the message m_T. By using her knowledge of m_p, she can perform an optimal measurement, modeled by a POVM. Without the hash function $H_{k_1}(\cdot)$, the quantum encryption from the scheme ensures that the probability to correctly guess a bit from the encryption key is 1/4. This result is based on the knowledge an adversary has on the public message m_p, combined with her measurement of the quantum bits. By measuring each quantum bit (by selecting at random a measuring basis from $\{+, \times\}$), the adversary obtains either the classical result 0 or 1, resulting in a classical message m_E. By comparing m_E with the public message m_p, she can determine the correct key bit at those instances where her outcome differs from the public message. This happens with probability 1/4 for each bit. Therefore, the probability to guess e bits correctly is $(1/4)^e$. Next, we include the hash function $H_{k_1}(\cdot)$ which takes as input the public message and the one-time ticket K_A^i, unknown to Eve. In this case, the probability that Eve correctly guesses e bits $= 2^{n-n_1}(1/4)^e$.

To summarize, by selecting an $\epsilon-\mathrm{AXU}_2$ hash function and a PRNG generator from e.g. [7], we demonstrated that Eve can perform an attack on the construction with

probability $2^{n-n_1}(1/4)^e$, where $e = k/log_2 n_2$ and Alice and Bob share a secret-key $K_{AT} = k_1||k_2$ of size $k + n_2$;

6 Conclusion and future work

In this paper, we analyzed the security of the construction used in an arbitrated quantum authentication scheme on which an adversary can perform a known-message attack. We demonstrated that the quantum encryption from the scheme is a quantum cipher with perfect security. Then, we explained that an adversary can retrieve only a limited amount of information on the keystream. This is insufficient to mount an attack when a suitable generator is used. In future work, we will consider the effect of reusing the same hash function and give a mathematical proof of the security of the complete scheme.

References

[1] H. Bruyninckx, and D. Van Heule, "Arbitrated Secure Authentication realized by using quantum principles," Communications (ICC), 2015 IEEE International Conference on, London, June 2015, pp. 7420-7425.

[2] N. Asokan, G. Tsudik, and M. Waidner, "Server-supported signatures," Journal of Computer Security, vol. 5, n°1, pp. 91-108, Jan 1997.

[3] Rogaway, "Bucket hashing and its application to fast message authentication," Journal of Cryptology, vol. 12, n°2, pp. 91–115, 1999.

[4] S. Wiesner. "Conjugate coding," SIGACT News, 15 (1): 78-88, 1983.

[5] I. Damgård, T. Pedersen, and L. Salvail, "On the Key-Uncertainty of Quantum Ciphers and the Computational Security of One-way Quantum Transmission," in Advances in Cryptology - EUROCRYPT 2004, Springer-Verlag, 2004, pp. 91-108.

[6] A. Ambainis, M. Mosca, A. Tapp and R. deWolf, "Private Quantum Channels," Proceedings of the 41st Annual Symposium on Foundations of Computer Science, 2000, pp. 547–553.

[7] U. Maurer and J. Massey, "Local Randomness in Pseudorandom Sequences," Journal of Cryptology, vol. 4, pp. 135–149, 1991.

Fighting asymmetry with asymmetry in Reverse Fuzzy Extractors

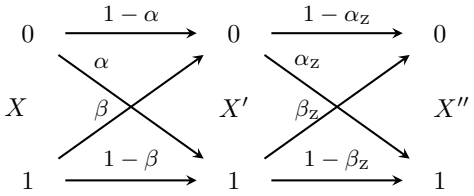
A. Schaller, T. Stanko, B. Škorić, S. Katzenbeisser

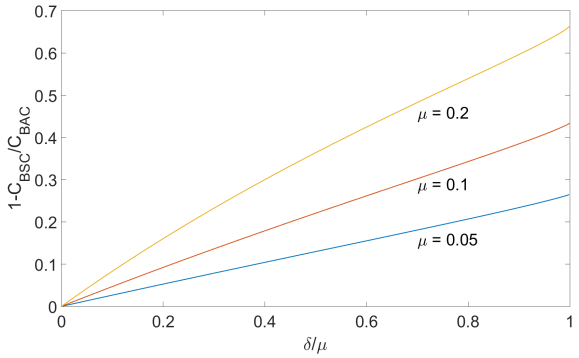
A Fuzzy Extractor is a cryptographic primitive that extracts a secret from a noisy source in a noise-tolerant way. The best known examples is the Code Offset Method, which makes use of a binary linear error-correcting code (ECC). The 'Reverse Fuzzy Extractor' is a variant in which the prover outsources the syndrome decoding step to the verifier. This allows for lightweight implementation of the prover device, but causes key leakage if the noise is data-dependent and privacy problems if the source has a drift. In this work we quantify both of these problems and propose solutions.

We briefly review the Reverse Fuzzy Extractor [1]. Measurement of the source at enrollment yields $X \in \{0,1\}^n$. The prover stores helper data $W = \mathtt{Syn}\, X$, where $\mathtt{Syn}$ denotes the ECC syndrome. The verifier stores the key $K = \mathtt{KDF}(X)$, where $\mathtt{KDF}$ is a key derivation function. At reconstruction time, the prover measures X'. He computes $\Sigma = W \oplus \mathtt{Syn}\, X' = \mathtt{Syn}(X \oplus X')$ and sends Σ to the verifier. The verifier computes the error pattern $E \in \{0,1\}^n$, $\mathtt{Syn}\, E = \Sigma$, and sends E. The prover reconstructs $\hat{X} = X' \oplus E$ and $\hat{K} = \mathtt{KDF}(\hat{X})$.

We model X as independent bits with bias p; data-dependent noise as a Binary Asymmetric Channel (BAC) for each position in X independently, with probability α for $0 \to 1$ flips and β for $1 \to 0$. An attacker observes k different error patterns from the same device. We define $T_i \in \{0, \ldots, k\}$ as the number of patterns containing '1' in location i. The leakage $I(X; T_1, \ldots, T_n)$ equals $n\mathsf{H}(T_i) - n\mathsf{H}(T_i|X_i)$, where the probabilities $T_i|X_i$ are binomial: $\Pr[T_i = t|X_i = 0] = \binom{k}{t}\alpha^t(1-\alpha)^{k-t}$ and $\Pr[T_i = t|X_i = 1] = \binom{k}{t}\beta^t(1-\beta)^{k-t}$. The leakage becomes worse with increasing k.

As a countermeasure we propose to concatenate precisely tuned synthetic Z-channel noise to the BAC. The prover adds the synthetic noise N to X' and only then computes the syndrome. After receiving E from the verifier he removes N and then computes $\hat{X}$. The effect of our solution is twofold: (i) the combined channel (BAC+Z) is symmetric, and hence the leakage is eliminated; (ii) the amount of noise has increased, which reduces the key length. We quantify the 2nd effect by looking at the channel capacity reduction w.r.t. the original BAC. See figure below. The μ, δ are defined as $\alpha = \mu - \delta$, $\beta = \mu + \delta$. BSC stands for Binary Symmetric Channel. If $\alpha \geq \beta$ then $\alpha_z = 0$, $\beta_z = 2|\delta|/(1 + 2|\delta|)$; if $\alpha \leq \beta$ then $\beta_z = 0$, $\alpha_z = 2\delta/(1 + 2\delta)$. The resulting BSC noise parameter is $(\mu + |\delta|)/(1 + 2|\delta|)$.





Drift in X causes a constant part in $\mathtt{Syn}\, X$ which makes a device recognizable (but it does not cause serious key leakage). A simple countermeasure [2] is to keep track of the drift and to compensate it before computing the syndrome.

[1] A. Herrewege, S. Katzenbeisser, R. Maes, R. Peeters, A.-R. Sadeghi, I. Verbauwhede, and C. Wachsmann, *Reverse Fuzzy Extractors: enabling lightweight mutual authentication for PUF-enabled RFIDs*, Financial Cryptography 2012, LNCS 7397, pp. 374–389.

[2] A. Schaller, B. Škorić, S. Katzenbeisser, *On the systematic drift of Physically Unclonable Functions due to aging*, TRUSTED 2015, pp. 15–20.

[3] `https://eprint.iacr.org/2014/741`

Linear Cryptanalysis of Reduced-Round Speck

Tomer Ashur Daniël Bodden
KU Leuven and iMinds
Dept. ESAT, Group COSIC
Address Kasteelpark Arenberg 10 bus 2452, B-3001 Leuven-Heverlee, Belgium
tomer.ashur-@-esat.kuleuven.be dbodden-@-esat.kuleuven.be

Abstract

Since DES became the first cryptographic standard, block ciphers have been a popular construction in cryptology. Speck is a recent block cipher developed by the NSA in 2013. It belongs to the cipher family known as ARX. ARX constructions are popular because of their efficiency in software. The security of the cipher is derived from using modular addition, bitwise rotation and xor. In this paper we employ linear cryptanalysis for variants of Speck with block sizes of 32, 48, 64, 96, and 128 bits. We illustrate that linear approximations with high bias exist in variants of Speck.

1 Introduction

A recent surge of new block cipher designs has urged the need for cryptanalysts to scrutinize their security. Lightweight ciphers are designed for resource-constrained devices. Designing ciphers for these devices means that one has to make design trade-offs keeping in mind the limitation that such an environment presents, such as limited memory, restricted computational and energy resources, etc.

A popular construction for block ciphers is Addition, Rotation and XOR, abbreviated as ARX. In this construction a block is split into 2 or more words, which are then added, XORed and rotated by the round function. The popularity of ARX construction stems from its good performance in software. Confusion is achieved by using modular addition in the round function, whereas diffusion is achieved by using cyclic rotation and xor.

In this paper we analyse the security of Speck, a block cipher that has been published by the NSA in 2013 [2]. Speck is a lightweight block cipher designed to achieve good performance in software. The block consists of two words, each 16/24/32/48, or 64 bits, which are processed a number of times by the round function. At the end of the last round a ciphertext is obtained. Speck comes in different variants for which different security and performance levels are provided; see Table 1.

Since the publication of the cipher in 2013, several papers have analyzed its security [1, 4, 7]. The best published attacks on Speck are differential cryptanalytic attacks described in [7]. In this paper we analyse the resistance of all variants of Speck against linear cryptanalysis.

Linear cryptanalysis is a known plaintext attack developed in 1993 by Matsui [9]. When using this method the adversary searches for possible correlations between bits of the input and bits of the output. Once such a correlation is found, the known plaintexts and ciphertexts can be used to recover bits of the secret key. In our analysis we find linear approximations covering 7, 8, 11, 10, and 11 rounds for versions of Speck with block size 32, 48, 64, 96, and 128, respectively. These approximations can be further extended to attack more rounds using methods such as those presented in [8].

This paper is structured as follow: In section 2 we describe Speck. In section 3 we discuss shortly the related work. In section 4 we explain the automatic search method that has been used to obtain linear approximations for Speck. In section 5

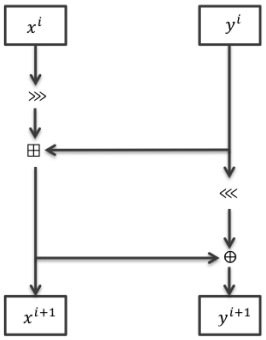


Figure 1: Speck round function

we present the approximations we found. In section 6 we explain how to obtain a linear distinguisher from the linear trails presented in section 5. Finally, in section 7 we conclude this paper.

2 A Brief Description of Speck

The cipher comes in several versions sharing the same Feistel-structure and round function. The different parameters of all Speck versions are presented in Table 1.

Table 1: Variants of the Speck-family [2]

Block Size n	Key Size	Word Size	α	β	Rounds
32	64	16	7	2	22
48	72	24	8	3	22
	96				23
64	96	32	8	3	26
	128				27
96	96	48	8	3	28
	144				29
128	128	64	8	3	32
	192				33
	256				34

The round function uses three operations:

- ◇ Modular addition, $\boxplus$.
- ◇ Left and right circular bit-shifts, by α and β.
- ◇ Bitwise XOR, $\oplus$.

The output of the round function is:
$x^{i+1} = (x^i \ggg \alpha) \boxplus (y^i)$
$y^{i+1} = ((x^i \ggg \alpha) \boxplus (y^i)) \oplus (y^i \lll \beta))$
The output words x^{i+1} and y^{i+1} are the input words for the next round, and the output of the last round is the ciphertext. In Figure 1 the round function of Speck is shown.

Table 2: Previous attacks concerning Speck

Variant n/key size	Rounds Attacked/ Total Rounds	Time	Data	Memory	Source
32/64	11/22	$2^{46.7}$	$2^{30.1}$	$2^{37.1}$	[1]
	14/22	2^{63}	2^{31}	2^{22}	[7]
48/72	12/22	2^{43}	2^{43}	na	[4]
	14/22	2^{65}	2^{41}	2^{22}	[7]
48/96	12/23	2^{43}	2^{43}	na	[4]
	15/23	2^{89}	2^{41}	2^{22}	[7]
64/96	16/26	2^{63}	2^{63}	na	[4]
	18/26	2^{93}	2^{61}	2^{22}	[7]
64/128	16/27	2^{63}	2^{63}	na	[4]
	19/27	2^{125}	2^{61}	2^{22}	[7]
96/96	15/28	$2^{89.1}$	2^{89}	2^{48}	[1]
	16/28	2^{85}	2^{85}	2^{22}	[7]
96/144	16/29	$2^{135.9}$	$2^{90.9}$	$2^{94.5}$	[1]
	17/29	2^{133}	2^{85}	2^{22}	[7]
128/128	16/32	2^{116}	2^{116}	2^{64}	[1]
	17/32	2^{113}	2^{113}	2^{22}	[7]
128/192	18/33	$2^{182.7}$	$2^{126.9}$	$2^{121.9}$	[1]
	18/33	2^{177}	2^{113}	2^{22}	[7]
128/256	18/34	$2^{182.7}$	$2^{126.9}$	$2^{121.9}$	[1]
	19/34	2^{241}	2^{113}	2^{22}	[7]

3 Related Work

Since the publication of Speck there has been a fair amount of analyzing done on the security and performance of the cipher. The best attacks to date are of the family of differential cryptanalysis [1, 4, 7]. Differential cryptanalysis has been developed by Biham and Shamir in 1990 [3]. When using differential cryptanalysis the adversary investigates how differences in the input affect the output, trying to discover non-random behavior. Several specialized techniques of differential cryptanalysis have been used to analyse Speck, such as the boomerang and the rectangle attacks [4, 7].

Results of previous attacks concerning Speck are presented in Table 2. From these results we gather that these methods are successful in attacking a large number of rounds in the chosen plaintext model. Looking at the design of Speck we might expect that linear cryptanalysis be effective as well. This paper complements previous work by evaluating the resistance of Speck against linear cryptanalysis.

3.1 Linear Cryptanalysis

Linear cryptanalysis [9] is a powerful cryptanalytic tool with regards to cryptanalysis of block ciphers. When using linear cryptanalysis, an adversary tries to find a linear expression that approximates a non-linear function with a probability different than $\frac{1}{2}$. When a good approximation, consisting of a relation between the plaintext and ciphertext, is found, the adversary gains information about the secret key. The approximation has the form:

$$P_i \oplus \ldots \oplus P_j \oplus C_k \ldots \oplus C_l = K_m \ldots \oplus K_n \tag{1}$$

with $P_i \dots P_j$ being plaintext bits, $C_k \dots C_l$ ciphertext bits and $K_m \dots K_n$ key bits. The approximation holds with some probability p, and its quality is usually measured by the bias which is defined as $\epsilon = |p - \frac{1}{2}|$.

Knowing the probability of the linear approximation, the adversary can determine the number of required known plaintexts, i.e., the data complexity ϵ^{-2}.

Once a good approximation is found the adversary can retrieve one bit of the key. To combine linear approximations we use the Piling Up Lemma [9]. This will tell us the bias of the combined approximation, which is the amount the probability deviates from $\frac{1}{2}$. For two approximations, with ϵ_1 and ϵ_2 as their respective biases, combining them gives an overall bias of:

$$\epsilon = 2 \cdot \epsilon_1 \epsilon_2. \tag{2}$$

This can be generalized for σ approximations:

$$\epsilon = 2^{\sigma-1} \cdot \prod_{i=1}^{\sigma} \epsilon_i. \tag{3}$$

The combined approximations can be used to form a linear distinguisher ζ for the cipher. This ζ can be used to detect non-random behavior in supposed to be random signals. Meaning that ζ can be used to detect if a certain cipher is being used. When a good ζ is found, the input bits are masked by $\{\lambda_{x^1}; \lambda_{y^1}\}$, and the output bits are masked using $\{\lambda_{x^r}; \lambda_{y^r}\}$, resulting in:

$$T = \#\left\{\lambda_{x^1} \cdot x_1 \oplus \lambda_{y^1} \cdot y_1 \oplus \lambda_{x^r} \cdot x_r \oplus \lambda_{y^r} \cdot y_r = 0\right\} \tag{4}$$

with x_1 and y_1 being the plaintexts, x_r and y_r the ciphertexts, r the length of ζ, T a counter and $\cdot$ the dot product. This should be repeated for at least ϵ^{-2} messages. If T is around:

$$\epsilon^{-2} \cdot (\frac{1}{2} \pm \epsilon) \tag{5}$$

the cipher can be distinguished from a random permutation.

3.2 Properties of Modular Addition

The modular addition operation, which is used as the nonlinear operation of Speck, consists of an xor and a carry. A chain of carries means using the previous carry in the current operation. This makes the behavior of modular addition nonlinear. To approximate the behavior of Speck, we have to deal with modular addition in such a way that accounts for this nonlinear behavior. The property of modular addition that we use is exploring the correlation between two neighboring bits. Suppose we have 3 words x, y and z with $z = x \boxplus y$. According to Cho and Pieperzyk [5], each single $z_{(i)}$ bit written as function of $x_{(i)}$,$\cdots$,$x_{(0)}$ and $y_{(i)}$,$\cdots$,$y_{(0)}$ bits can be expressed as:

$$z_{(i)} = x_{(i)} \oplus y_{(i)} \oplus x_{(i-1)} \oplus y_{(i-1)} \oplus \sum_{j=0}^{i-2} x_{(j)} y_{(j)} \prod_{k=j+1}^{i-1} x_{(k)} \oplus y_{(k)}, \quad i = 1, \dots, n-1 \tag{6}$$

with x_i and y_i each representing 1 bit each. The carry is represented by $x_{i-1} \oplus y_{i-1} \oplus \sum_{j=0}^{i-2} x_j y_j \prod_{k=j+1}^{i-1} x_k \oplus y_k$ which we refer to as $R_i(x, y)$. The parity of two consecutive bits can be approximated as:

$$z_{(i)} \oplus z_{(i-1)} = x_{(i)} \oplus y_{(i)} \oplus x_{(i-1)} \oplus y_{(i-1)}, \quad p = \frac{3}{4} \tag{7}$$

In this expression, we use a property of modular addition mentioned in another paper by Cho and Pieperzyk [6] that removes the carry chain from Equation 6. Meaning that if a mask λ contains only two consecutive bits, one can write:

$$P\left[\lambda \cdot (x \boxplus y) = \lambda \cdot (x \oplus y)\right] = \frac{3}{4}. \tag{8}$$

This expression means that using a mask λ to mask the bits we want to throw away and keep the bits we are interested in (e.g two consecutive bits), we linearize the left side of the expression by replacing it with the right-side. This approximation holds with probability $\frac{3}{4}$ and has a bias $\frac{1}{4}$. For an ARX construction, Equation 8 remains valid as long as the following two conditions are avoided:

1. Bitwise rotation moves a pair of approximated bits to the MSB (most significant bit position) and LSB (least significant bit position) hence not adhering to the Cho and Pieperzyk framework; or
2. Xor creates a single bit hence not adhering to the Cho and Pieperzyk framework.

4 Automated Search for Linear Approximations in ARX Constructions

In order to automate the search for a good linear approximation, we have designed an ARX toolbox. The toolbox is implemented using a parallel programming language called OpenCL. The configuration on which the computation was run has been a 40 core Intel Xeon machine with a clock speed of 3.10GHz, exclusively using CPU's. The time it took to compute all pairs of consecutive bits for the 32-bit input variant of Speck was less than 0.01 seconds to less than a week for the largest version of Speck.

We have analyzed in our work all possible pairs of consecutive bits over the length of a word (e.g. starting with one pair of two consecutive bits up to $\frac{n}{2}$ pairs of two consecutive bits). We have done this analysis in three ways, forward direction, backward direction and combining forward and backward together. We have iterated over all pairs of consecutive bits until one of the two stop conditions was met. The first stop condition is when a mask containing non-consecutive bits is encountered in the round function entering into the modular addition. The other condition upon which the computation will be stopped is when the counter exceeds the maximum bias. In order to find approximations with good bias we restrict the allowable counter to $T \leq \frac{n}{2} + 1$. The bias for one pair of consecutive bits is $1 - P[R_i(x, y) = 0] = \frac{1}{4}$. Generalized to ℓ pairs of consecutive bits the bias can be calculated using the Piling Up Lemma:

$$2^{\ell-1} \cdot (1 - P[R_i(x, y) = 0])^{\ell} \tag{9}$$

with values for $\ell = 1, \ldots, \frac{n}{2}$. The bias is calculated with the value obtained by the hamming weight of the mask passing the modular addition in the round function divided by 2, plus the bias of the previous rounds, calculated on the approximated modular addition $\lambda \cdot (x \oplus y)$.

5 Linear Approximation of the Round Function of Speck

Linear cryptanalysis relies on collecting a large amount of input and output pairs in order to verify whether the approximation has been satisfied or not. In this work we

started with fixing one mask λ_{x^1} with one pair of consecutive bits (0x3, 0x6, ...) and keeping the other mask λ_{y^1} zero, checking how the masks evolve both in the forward and backward direction. After receiving promising results for one pair we extended the computation to all possible pairs of two consecutive bits given a block size.

An overview of the obtained results from our experiments is shown in Table 3. These results show that the distinguisher can cover up to 11-rounds for different versions of Speck. More details of linear cryptanalysis on Speck are given in Table 4.

A careful note to the reader when reviewing the results in Table 4 is that for 96-bit version of Speck we retrieved 10-rounds, whereas for the 64 and 128-bit versions we retrieved 11-rounds each. The reason for the different r between these versions is stop condition 1. For each version of Speck, the length of a word and mask vary according to n. Meaning that bitwise rotation operates over a different length (i.e., $\frac{n}{2}$).

We present in Table 5 the full linear trail in the forward and in the backward direction for the 32/64 variant. The information found in this table gives an idea why the computation broke in that particular round. Using this information in future research could extend the distinguisher further.

Table 3: Distinguisher length for different versions of Speck

Variant n	Distinguished rounds / Total rounds	Bias	Time Complexity	Data Complexity
32	7/22	2^{-14}	2^{28}	2^{28}
48	8/22	2^{-22}	2^{44}	2^{44}
64	11/26	2^{-32}	2^{64}	2^{64}
96	10/28	2^{-47}	2^{92}	2^{92}
128	11/32	2^{-63}	2^{144}	2^{144}

Table 4: Combined results for different versions of Speck

Variant n	Distinguished rounds / Total rounds	Bias ϵ	λ_{x^1}	λ_{y^1}	Number of rounds backward	Number of rounds forward
32	7/22	2^{-14}	0x0006	0x0000	3	4
48	8/22	2^{-22}	0x000003	0x000000	5	3
64	11/26	2^{-32}	0x00000003	0x00000000	8	3
96	10/28	2^{-47}	0x000000000000	0x01800000000c	5	5
128	11/32	2^{-63}	0x0000003000000003	0x0000000000000000	8	3

Table 5: Linear trail for Speck32/64

Direction	Cost	λ_{x^i}	λ_{y^i}	$\lambda_{x^{i+1}}$	$\lambda_{y^{i+1}}$	Reasons stopped
				0x0300	0xe0c7	broke consecutive for 0x300 $\oplus$ 0xe0c7
	1	0x0300	0xe0c7	0x8301	0x8307	
	2	0x8301	0x8307	0x0300	0x0006	
Backward	1	0x0300	0x0006	0x0006	0x0000	
Forward	1	0x0006	0x0000	0x3c00	0x3000	
	2	0x3c00	0x3000	0xc198	0xc1e0	
	3	0xc198	0xc1e0	0xf00c	0xc18f	
	3	0xf00c	0xc18f	0x785f	0x61bf	
		0x785f	0x61bf			broke consecutive for 0x785f $\ggg \alpha$

6 A Linear Distinguisher for Speck

Using the results from this paper we can build a linear distinguisher for each version of Speck. We present here a distinguisher for the 32-bit version. To build the linear distinguisher we use Equation 4 and fill in with information from Table 4, giving:

$$2^{28} \cdot (\frac{1}{2} \pm 2^{-14}) \tag{10}$$

using this distinguisher we can distinguish Speck32/64 from a random permutation. In Algorithm 1 a pseudo code of the distinguishing attack is shown.

Data: 2^{28} messages, input masks $\{\lambda_{x^1}; \lambda_{y^1}\}$, output masks $\{\lambda_{x^r}; \lambda_{y^r}\}$
set $T = 0$
for 2^{28} *messages* **do**
  if $\{\lambda_{x^1} \cdot x \oplus \lambda_{y^1} \cdot y \oplus \lambda_{x^r} \cdot x \oplus \lambda_{y^r} \cdot y = 0\}$ **then**
    $T = T + 1$
  end
end
if $T \approx 2^{28} \cdot (\frac{1}{2} \pm 2^{-14})$ **then**
  return 1
else
  return 0
end

Algorithm 1: A pseudo-code of the distinguishing attack

7 Conclusion

In this paper we investigated the linear behavior of Speck in the known plaintext model. We explained the theory behind our method, presented linear approximations for each version of Speck and gave one example a linear distinguisher for Speck32/64.

The analysis in this paper tested the strength of the Speck round function and demonstrated that the cipher offers sufficient resistance against linear cryptanalysis. Future work may find better linear approximations. Proven that linear cryptanalysis can achieve a meaningful distinguisher, we believe that future research on ARX constructions may deliver good linear approximations for these ciphers. Our tool can be used to test the resistance of such constructions performing the computation with the latest parallel computing techniques.

Acknowledgments. The authors wish to Thank Vincent Rijmen for his support, Mathy Vanhoef for reviewing an earlier version of this paper, and R. Tzach for some thought provoking discussions.

This work was partially supported by the Research Fund KU Leuven, OT/13/071. The first author was also supported by Google Europe Scholarship for Students with Disabilities.

References

[1] Abed, F., List, E., Wenzel, J., Lucks, S.: Differential cryptanalysis of round-reduced simon and speck. Preproceedings of Fast Software Encryption (FSE 2014)(2014, to appear) (2014)

[2] Beaulieu, R., Shors, D., Smith, J., Treatman-Clark, S., Weeks, B., Wingers, L.: The simon and speck families of lightweight block ciphers. IACR Cryptology ePrint Archive (2013) 404

[3] Biham, E., Shamir, A.: Differential cryptanalysis of the data encryption standard. Volume 28. Springer-Verlag New York (1993)

[4] Biryukov, A., Roy, A., Velichkov, V.: Differential analysis of block ciphers simon and speck. In: International Workshop on Fast Software Encryption-FSE. (2014)

[5] Cho, J.Y., Pieprzyk, J.: Algebraic attacks on sober-t32 and sober-t16 without stuttering. In: Fast Software Encryption, Springer (2004) 49–64

[6] Cho, J.Y., Pieprzyk, J.: Multiple modular additions and crossword puzzle attack on nlsv2. In: Information Security. Springer (2007) 230–248

[7] Dinur, I.: Improved differential cryptanalysis of round-reduced speck. In: Selected Areas in Cryptography–SAC 2014. Springer (2014) 147–164

[8] Knudsen, L.R., Mathiassen, J.E.: A chosen-plaintext linear attack on des. In: Fast Software Encryption, Springer (2001) 262–272

[9] Matsui, M.: Linear cryptanalysis method for des cipher. In: Advances in Cryptology—EUROCRYPT'93, Springer (1994) 386–397

An Efficient Privacy-Preserving Comparison Protocol in Smart Metering Systems

Majid Nateghizad Zekeriya Erkin Reginald L. Lagendijk
Delft University of Technology
Department of Intelligent Systems, Cyber Security Group
Delft, The Netherlands
M.Nateghizad@tudelft.nl Z.Erkin@tudelft.nl R.L.Lagendijk@tudelft.nl

Abstract

In smart grids, providing power consumption statistics to the customers and generating recommendations for managing electrical devices are considered to be effective methods that can help to reduce energy consumption. Unfortunately, providing power consumption statistics and generating recommendations rely on highly privacy-sensitive smart meter consumption data. From past experience, we see that it is essential to find scientific solutions that enable the utility providers to provide such services for their customers without damaging customers' privacy. One effective approach relies on cryptography, where sensitive data is only given in the encrypted form to the utility provider and is processed under encryption without leaking content. The proposed solutions using this approach are very effective for privacy protection but very expensive in terms of computation. In this paper, we focus on an essential operation for designing a privacy-preserving recommender system for smart grids, namely comparison, that takes two encrypted values and outputs which one is greater than the other one. We improve the state-of-the-art comparison protocol based on Homomorphic Encryption in terms of computation by 56% by introducing algorithmic changes and data packing. As the smart meters are very limited devices, the overall improvement achieved is promising for the future deployment of such cryptographic protocols for enabling privacy enhanced services in smart grids.

1 Introduction

Smart grids, as the next generation of power grid, are utilizing both communication technologies and information processing to monitor and manage power grids to enhance reliability, efficiency, and sustainability of power generation. One of the advantages of smart grids compared to traditional power grids is the ability to observe the power consumption of households in very short time intervals in the order of seconds to minutes. As a result of the fine-coarse data reporting, it is possible to provide power consumption statistics to the consumers, which might help to reduce the overall consumption by changing customer behavior, as pointed out in several works [1]. For example, Honebein *et al.* [2] defined people as the only true smart part of a smart grid; therefore monitoring, understanding, and promoting the end-users' roles from passive to active is considered as a fundamental action in smart grids. To this end, there are already several utility companies providing their customers devices and smart phone applications to monitor their real time consumption. Furthermore, one of the goals of the utility providers, balancing the supply and the demand, also known as demand response (DR), can be achieved more effectively if the utility provider can also provide statistics about the power usage in the surrounding area and generate personalized recommendations, for example to manage electrical devices like electric cars, heating systems, and ovens in the household [3].

Providing statistics on power consumption and generating personalized recommendations to inform customers are heavily dependent on the smart meter consumption readings. Unfortunately, these readings are highly privacy-sensitive [4]. The utility provider can use the readings from the smart meters for other purposes, misuse them or even transfer them to other entities without the consent of the customers. As seen in many cases, privacy is considered to be a big challenge for using smart meters to the fullest extent, e.g. enabling personalized services such as generating recommendations.

In this paper, we assume that the utility provider generates statistics and recommendations for the customers so that the customers can adjust the electrical devices for the most cost-effective and environmentally friendly manner. To achieve this, we rely on cryptography, which provides us tools to create Privacy by Design algorithms. For instance, there are already a number of studies for computing bills and aggregating data [5, 6]. The main idea in this research line is to provide only the encrypted power consumption to the utility provider and enable processing the encrypted data without decrypting any sensitive information. This way, the utility provider cannot access to the content but at the same time can perform the algorithms required for the service. Unfortunately, the cryptographic algorithms for this purpose are expensive in terms of computation, which mostly require smart meters to be involved in the computation [7, 8, 9]. Since the smart meters are very limited devices, improving the efficiency of the cryptographic algorithms is a challenge.

We address the efficiency problem of a fundamental operation, namely comparison, which is required to design any recommender system. In our setting, the encrypted consumption readings are collected from the customers by an aggregator and the utility provider has the decryption key. For privacy reasons, the aggregator cannot transfer the data directly to the utility provider but can co-operate with the utility provider to generate recommendations. One important step in the system is to compare values, which are only available in the encrypted form. More precisely, the aggregator has two encrypted values, and it needs to know which one is greater than the other one without revealing their contents to anyone including itself.

There are numerous comparison protocols designed for comparing encrypted values [7, 9]. In this paper, we improve the state-of-the-art comparison protocol that relies on homomorphic encryption in terms of run-time by 56% by introducing algorithmic changes. Furthermore, we also reduce decryption cost of the protocol by deploying data packing [10]. Together, these improvements increase the overall efficiency of the comparison protocol with encrypted inputs, bringing smart meters one step closer to run privacy-preserving cryptographic protocols based on homomorphic encryption.

2 Preliminaries

In this section, we describe the application setting, the security assumptions, and the cryptographic tools used in this work.

Application Setting

In our application setting, we define three roles: 1) smart meters installed at the households, 2) a data aggregator, and 3) a utility provider. Smart meters measure, encrypt, and send consumers' power consumption to the data aggregator, which collects and analyzes encrypted power consumption. Then, the utility provider generates recommendations for its customers by running a cryptographic protocol with the data aggregator. The output of the cryptographic protocol, which depends on the purpose of the recommender system, is in the encrypted form, thus it is not available neither to the data aggregator nor to the utility provider. The output is then revealed to the customer by using another a secure decryption protocol.

Security Model

The proposed protocol in this work is built on the semi-honest adversarial model, where the data aggregator, and the utility provider are honest in the sense that they faithfully follow the designed protocol but will try to infer information from the protocol execution transcript. This assumption is realistic since companies are expected to properly perform required services mentioned in the service level agreement, when engaging in a collaboration. We assume that the utility provider is the only party holding the private keys, while the smart meters and the data aggregator have the public keys for the encryption schemes. We assume that neither party colludes.

Homomorphic Encryption

In this work, we rely on two additively homomorphic cryptosystems, Paillier [11] and DGK (Damgård, Geislet and Krøigaard) [7]. An additively homomorphic encryption scheme preserves certain structure that can be exploited to process ciphertexts without decryption. Given $\mathcal{E}_{pk}(m_1)$ and $\mathcal{E}_{pk}(m_2)$, a new ciphertext whose decryption yields the sum of the plaintext messages m_1 and m_2 can be obtained by performing a certain operation over the ciphertexts: $\mathcal{D}_{sk}(\mathcal{E}_{pk}(m_1)) \otimes (\mathcal{E}_{pk}(m_2)) = m_1 + m_2$.

Consequently, exponentiation of any ciphertext with a public value yields the encrypted product of the original plaintext and the exponent: $\mathcal{D}_{sk}(\mathcal{E}_{pk}(m)^e) = e \cdot m$.

In the rest of the paper, we denote the ciphertext of a message m by $[m]$ for the Paillier cryptosystem and $[\![m]\!]$ for the DGK.

3 Secure Comparison Protocol with Secret Inputs

In this section, we describe the state-of-the-art secure comparison protocol (SCP), which takes two encrypted inputs and outputs the greater one in the encrypted form. SCP based on the DGK construction introduced in [12] is one of the widely-used comparison protocols due to its efficiency. The DGK comparison protocol is a sub-protocol in the SCP, where each party possesses a secret but plaintext value. The sub-protocol also uses the DGK cryptosystem for efficiency reasons.

The comparison protocol in [12] is modified and used by Erkin *et al.* in [8], and Veugen proposed an improved DGK comparison protocol (IDCP) in [9]. In the following, we describe the SCP construction.

For the sake of simplicity, we use the names Alice and Bob as the data aggregator and the utility provider, respectively. We assume that Bob has the secret key sk and Alice has access to two encrypted values, $[a]$ and $[b]$, and wants to know if $a < b$.

Initially, Alice computes $[z] = [2^\ell + a - b] = [2^\ell] \cdot [a] \cdot [b]^{-1}$, and then obtains the result of comparison as follows:

$$[z_\ell] = [2^{-\ell} \cdot (z - (z \bmod 2^\ell))] = ([z] \cdot [z \bmod 2^\ell]^{-1})^{2^{-\ell}}\,, \tag{1}$$

where $[z_\ell]$ is the most significant bit of $[z]$ and the result of comparison. If $z_\ell = 1$ then we have $a > b$, and otherwise $a < b$. A more efficient method of computing $[z_\ell]$ is based on the IDCP, where we can compute $z_\ell = \lfloor z/2^\ell \rfloor$ and $[a < b] = [1 - z_\ell] = [1] \cdot [z_\ell]^{-1}$, but we still need to compute $[z \bmod 2^\ell]$. A more detailed explanation regarding computation of $[z_\ell]$ is provided in the following sections.

Computing $[z \bmod 2^\ell]$

Notice that Alice has access only to $[z]$, and interaction with Bob, who has the private key, is needed to compute modulo reduction, $[z \bmod 2^\ell]$. However, Alice cannot

give $[z]$ directly to Bob since this value reveals information on the difference of a and b. Therefore, Alice masks $[z]$ using a random value as follows:

$$[d] = [z + r] = [z] \cdot [r] \ , \tag{2}$$

where r is a $(\kappa + \ell)$-bit uniformly random number and κ is a security parameter. After masking, Alice sends $[d]$ to Bob to perform modulo reduction, where Bob first decrypts $[d]$, then computes $\acute{d} = d \bmod 2^\ell$ and sends $[\acute{d}]$ and $[d/2^\ell]$ back to Alice. Subsequently, to obtain $[z \mod 2^\ell]$, Alice computes $[\tilde{z} \mod 2^\ell] = [\acute{d} - r \mod 2^\ell] = [\acute{d}] \cdot [r \mod 2^\ell]^{-1}$.

Note that $z \bmod 2^\ell = \tilde{z} \bmod 2^\ell$ if $\acute{d} > r \bmod 2^\ell$. When $\acute{d} < r \bmod 2^\ell$, an underflow occurs, and Alice has to add 2^ℓ to $[\tilde{z}]$ to make the value positive again. Therefore, Alice needs to determine whether $\acute{d} > r \bmod 2^\ell$ or not. This is achieved by computing an encrypted value, $[\lambda]$, which shows the relation between $\acute{d}$ and $r \mod 2^\ell$. Then, Alice can perform following computation to obtain $[z \mod 2^\ell]$:

$$[z \mod 2^\ell] = [\tilde{z} + \lambda 2^\ell] = [\tilde{z}] \cdot [\lambda]^{2^\ell} \ . \tag{3}$$

Alice can obtain $[z_\ell]$ by using Equation 1. $[z_\ell]$ can be computed more efficiently as follow:

$$[z_\ell] = [\Psi(z)] = [\Psi(d)] \cdot [\Psi(r)]^{-1} \cdot [\lambda]^{-1} \tag{4}$$

where $\Psi(x) = \lfloor x/2^\ell \rfloor$. For computing $[\lambda]$, we run a secure comparison protocol with private inputs as described in the following section.

Computing $[\lambda]$

This protocol outputs an encrypted bit, which shows whether $\acute{d} > \hat{r} = r \bmod 2^\ell$ or not. However, different than the original problem of comparing encrypted a and b, in this protocol Alice and Bob possess $\hat{r}$ and $\acute{d}$ in plaintext, respectively. Based on this setting, the IDCP for computing $[\lambda]$ securely works as follows:

1. Bob sends a bitwise encryption of his input, $[\![\acute{d}_0]\!], ..., [\![\acute{d}_{\ell-1}]\!]$, to Alice.

2. Alice chooses uniformly random bit δ, where $\delta \in \{0, 1\}$. Then she computes $s = 1 - 2 \cdot \delta$ and $[\![c_i]\!]$ as follows,

$$\begin{aligned} [\![c_i]\!] &= [\![\acute{d}_i - \hat{r}_i + s + 3 \sum_{j=i+1}^{\ell-1} \acute{d}_j \oplus \hat{r}_j]\!] \\ &= [\![\acute{d}_i]\!] \cdot [\![\hat{r}_i]\!]^{-1} \cdot [\![s]\!] \cdot \left(\prod_{j=i+1}^{\ell-1} [\![\acute{d}_j \oplus \hat{r}_j]\!] \right)^3 , \end{aligned} \tag{5}$$

   where $[\![\acute{d}_j \oplus \hat{r}_j]\!] = [\![\acute{d}_j]\!] \cdot [\![\hat{r}_j]\!] \cdot [\![\acute{d}_j]\!]^{-2 \cdot \hat{r}_j}$, and $i = 0, ..., \ell - 1$.

3. Alice blinds each $[\![c_i]\!]$ with a uniformly random $h_i \in_R \mathbb{Z}_u^*$ such that

$$[\![e_i]\!] = [\![c_i \cdot h_i]\!] = [\![c_i]\!]^{h_i} \ , \tag{6}$$

   then permutes $[\![e_i]\!]$ and sends them to Bob. Note that if $c_t = 0$, where $t \in \{0, ..., \ell - 1\}$ then $e_t = 0$ as well.

4. Bob checks whether there is a zero among $[\![e_i]\!]$ values. If none of the $[\![e_i]\!]$ values are encrypted zero then he sets $\tilde{\lambda} = 0$, otherwise $\tilde{\lambda} = 1$. Then he encrypts $\tilde{\lambda}$ and sends $[\tilde{\lambda}]$ to Alice.

5. Alice corrects $[\tilde{\lambda}]$ to obtain $[\lambda]$ as follows:

$$[\lambda] = \begin{cases} [\tilde{\lambda}] & \text{if } s = 1 \\ [1] \cdot [\tilde{\lambda}]^{-1} & \text{if } s = -1 \end{cases}$$

After obtaining $[\lambda]$, Alice computes $[z \mod 2^{\ell}]$ and $[z_{\ell}]$ based on Equations 3 and 1 respectively.

4 Efficient Privacy-Preserving Comparison Protocol (EPPCP)

In this section, we describe a new version of the original SCP based on the DGK construction, which is significantly more efficient in terms of run-time cost.

Proposed Comparison Protocol

Complexity analysis and experimental results reveal that the XOR operation in computing $[\![c_i]\!]$, in Equation 5, has a significant impact on the overall efficiency of the DGK comparison protocol for the following two reasons:

1. Computing XOR is computationally expensive, since $[\![\hat{r} \oplus \acute{d}]\!] = [\![\hat{r}]\!] \cdot [\![\acute{d}]\!] \cdot [\![\acute{d}]\!]^{-2 \cdot \hat{r}}$. Veugen [9] proposed a more efficient technique of computing XOR, where $[\![\hat{r} \oplus \acute{d}]\!] = [\![\acute{d}]\!]$ when $\hat{r} = 0$; otherwise, $[\![\hat{r} \oplus \acute{d}]\!] = [\![1]\!] \cdot [\![\acute{d}]\!]^{-1}$ (Recall that Alice and Bob have access to values $\hat{r}$ and $\acute{d}$, respectively and Alice is computing XOR). Thus, if $\hat{r}$ equals to 1, one multiplication and one exponentiation with negative exponent should be computed over DGK ciphertexts, which affects the performance of DGK comparison protocol significantly.

2. Since the equation that involves XOR is computed during the protocol with encrypted inputs, it is not possible to introduce pre-computation for $[\![c_i]\!]$ to obtain a more efficient protocol.

Table 1 shows that computing $[\![c_i]\!]$ constitutes 70% of the overall run-time of the IDCP for Alice.

Based on these two facts, we propose a more efficient way of computing $[\![c_i]\!]$, which does not rely on the original XOR computation. The value $[\![c_i]\!]$ can be re-written as follows:

$$[\![c_i]\!] = [\![\acute{d}_i - \hat{r}_i + s + \sum_{j=i+1}^{\ell-1} (\acute{d}_j \cdot 2^j \ - \ \hat{r}_j \cdot 2^j)]\!] \,. \tag{7}$$

Alice computes Equation 7 in three steps:

1. Bob computes $[\![t_i]\!] = [\![\acute{d}_i + \sum_{j=i+1}^{\ell-1} \acute{d}_j \cdot 2^j]\!]$, and sends $[\![t_i]\!]$ to Alice,

2. Alice computes $[\![v_i]\!] = [\![s - \hat{r}_i - \sum_{j=i+1}^{\ell-1} \hat{r}_j \cdot 2^j]\!]$, and

3. Alice computes $[\![c_i]\!]$ as follows,

$$[\![c_i]\!] = [\![t_i + v_i]\!] = [\![t_i]\!] \cdot [\![v_i]\!] \ . \tag{8}$$

Note that Alice can pre-compute $[\![v_i]\!]$ and factor '3' is not needed in the computation of $[\![c_i]\!]$. After computing all $[\![c_i]\!]$ values, Alice masks each $[\![c_i]\!]$ and sends masked values to Bob, where he checks if any of the given masked $[\![c_i]\!]$ is zero, then generates $[\tilde{\lambda}]$, and sends it to Alice. She corrects $[\tilde{\lambda}]$ based on value s to obtain $[\lambda]$, computes Equation 3, and 1 to obtain $[z_\ell]$ as in the original protocol. Note that we compare $2\acute{d}$ and $2\hat{r}$ instead of $\acute{d}$ and $\hat{r}$ respectively for technical reasons explained in the following section.

Table 1: Run-time performance for several steps of the IDCP.

Function	Time (second)	Overal computation (%)
	Alice	
Computing $[\![c_i]\!]$	15	70
$[\![e_i]\!] \leftarrow$ Masking $[\![c_i]\!]$	3.15	15
Other	3.15	15
	Bob	
DGK zero-check	27.3	38
Paillier decryption	44.4	62
Total	93	

4.1 Data Packing

According to Table 1, Paillier decryption of $[d]$ (Equation 2) dominates more than 62% of the comparison protocol execution time at Bob side. We decrease the run-time of Paillier decryption by employing data packing similar to [10, **?**]. The main idea behind data packing is to efficiently use the message space of the Paillier cryptosystem that is much larger than the values to be compared.

Assume that z and r are ℓ and $\ell+\kappa$-bit integers, respectively. Then, $[d] = [z+r]$ is a $(\ell+\kappa+1)$-bit integer. Let the message space of the Paillier cryptosystem be $\eta = pq$, then Alice packs $\rho = \lfloor(\ell+\kappa+1)/\eta\rfloor$ into one Paillier message as follows:

$$[\hat{d}] = \sum_{j=0}^{\rho-1} [d]_j \cdot (2^{\ell+\kappa+1})^j \ , \tag{9}$$

and sends $[\hat{d}]$ to Bob. Then, Bob computes $\mathcal{D}_{sk}([\hat{d}])$, unpacks ρ different values and performs modulo reduction on each unpacked value.

Employing the data packing technique not only reduces the number of very expensive Paillier decryption to be performed, but also decreases the number of encrypted messages to be transmitted.

5 Performance Analysis

In this section, we analyze the number of operations over ciphertexts, since they are computationally expensive compared to operations on the plaintext and dominate the protocol execution run-time, and provide experimental results for run-time performance. For this purpose, we implemented the EPPCP using C++ and SeComLib [13] library, on a Linux machine running Ubuntu 14.04 LTS, with 64-bit microprocessor and 8 GB of RAM. The experiments are repeated for $10,000$ comparisons. Table 2 provides more information about parameters and their corresponding values in our implementation.

Table 2: Parameters and their values used in the implementation.

Parameter	Symbol	Value
Bit size of inputs	ℓ	25 bits
Security parameter	κ	40 bits
Paillier message space	η	2048 bits
DGK message space	n	32 bits
Paillier/DGK key size	n	2048 bits
Number of $[d]$ packed into one Paillier ciphertext	ρ	31

Table 3 shows the computational complexity of the original DGK comparison protocol, the IDCP, and the EPPCP. Note that the number of multiplications and exponentiations are regarding the computation of $[\![c_i]\!]$.

Table 3: Computational complexity of original DGK [12, 8], the IDCP and the EPPCP.

Function	Original DGK	IDCP	EPPCP
Encryption	$1_{Paillier} + \ell_{DGK}$	$1_{Paillier} + \ell_{DGK}$	$1_{Paillier} + \ell_{DGK}$
Decryption	$1_{Paillier}$	$1_{Paillier}$	$(\frac{1}{\rho})_{Paillier}$
DGK zero-check	ℓ	ℓ	ℓ
Multiplication	$\ell(\ell+2)$	$\sim \frac{\ell(\ell+11)}{4}$	ℓ
Exponentiation(+)	ℓ	ℓ	0
Exponentiation(−)	$\frac{\ell(\ell+1)}{2}$	$\sim \frac{\ell(\ell+3)}{4}$	0

According to Table 4, running EPPCP $10,000$ times takes 41 seconds, where it takes 93 seconds for the IDCP. Table 4 also shows that pre-computation phase takes more time in EPPCP as a result of the new method of computing $[\![c_i]\!]$, which allows performing more initial computations before run-time.

References

[1] Verbong, G.P., Beemsterboer, S., Sengers, F.: Smart grids or smart users? involving users in developing a low carbon electricity economy. Energy Policy **52**, 117–125 (2013)

Table 4: Overall performance of the IDCP and the EPPCP.

Protocol	Run-time (second)	Pre-computation (second)
IDCP	93	7.4
EPPCP	41.4	13.8
Improvement	+56%	−87%

[2] Honebein, P.C., Cammarano, R.F., Boice, C.: Building a social roadmap for the smart grid. The Electricity Journal **24**(4), 78–85 (2011)

[3] Giordano, V., Gangale, F., Fulli, G., Jiménez, M.S., Onyeji, I., Colta, A., Papaioannou, I., Mengolini, A., Alecu, C., Ojala, T., *et al.*: Smart Grid Projects in Europe: Lessons Learned and Current Developments. Publications Office of the European Union, Luxembourg (2011)

[4] Liu, J., Xiao, Y., Li, S., Liang, W., Chen, C.: Cyber security and privacy issues in smart grids. IEEE Communications Surveys and Tutorials **14**(4), 981–997 (2012)

[5] Birman, K., Jelasity, M., Kleinberg, R., Tremel, E.: Building a Secure and Privacy-Preserving Smart Grid. ACM SIGOPS Operating Systems Review **49**(1), 131–136 (2015)

[6] Erkin, Z., Tsudik, G.: Private Computation of Spatial and Temporal Power Consumption with Smart Meters. In: Applied Cryptography and Network Security - 10th International Conference, Singapore. Proceedings, 561–577 (2012). Springer

[7] Damgard, I., Geisler, M., Kroigard, M.: A correction to 'efficient and secure comparison for on-line auctions'. International Journal of Applied Cryptography **1**(4), 323–324 (2009)

[8] Erkin, Z., Franz, M., Guajardo, J., Katzenbeisser, S., Lagendijk, R.L., Toft, T.: Privacy-Preserving Face Recognition. In: Privacy Enhancing Technologies, 9th International Symposium, Seattle, USA. Proceedings, pp. 235–253 (2009). Springer

[9] Veugen, T.: Improving the DGK comparison protocol. In: IEEE International Workshop on Information Forensics and Security, Costa Adeje, Tenerife, Spain, pp. 49–54 (2012). IEEE

[10] Troncoso-Pastoriza, J.R., Katzenbeisser, S., Celik, M., Lemma, A.: A secure multidimensional point inclusion protocol. In: Proceedings of the 9th workshop on Multimedia & Security, Dallas, Texas, USA, pp. 109–120 (2007). ACM

[11] Paillier, P.: Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. In: Advances in Cryptology - EUROCRYPT, International Conference on the Theory and Application of Cryptographic Techniques, Prague, Czech Republic. Proceeding, pp. 223–238 (1999). Springer

[12] Damgård, I., Geisler, M., Krøigaard, M.: Efficient and Secure Comparison for On-Line Auctions. In: Information Security and Privacy, 12th Australasian Conference, Townsville, Australia. Proceedings, pp. 416–430 (2007). Springer

[13] Cyber Security Group: SeComLib Secure Computation Library. http://cybersecurity.tudelft.nl Accessed 2013

Security Analysis of the Drone Communication Protocol: Fuzzing the MAVLink protocol

Karel Domin Eduard Marin Iraklis Symeonidis
KU Leuven
ESAT-COSIC and iMinds
Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium
karel.domin@student.kuleuven.be {first.surname}@esat.kuleuven.be

Abstract

The MAVLink protocol, used for bidirectional communication between a drone and a ground control station, will soon become a worldwide standard. The protocol has been the subject of research many times before. Through this paper, we introduce the method of fuzzing as a complementing technique to the other research, to find vulnerabilities that have not been found before by different techniques. The goal is to identify possible vulnerabilities in the protocol implementation in order to make it more secure.

1 Introduction

Currently, drones are used to support critical services such as forest fire and illegal hunting detection, search and rescue operations or to deliver medical supplies. For this purpose, they are often equipped with a navigation system (GPS), a camera and an audio interface. Furthermore, they have a radio that enables wireless communication with the Ground Control Station (GCS) or a remote control. Besides the clear benefits of using drones in all these services, they can also pose important security and privacy threats. The wireless communication channel opens up the door for several types of remote attacks. For example, adversaries could attempt to obtain sensitive data by eavesdropping the wireless medium, send malicious commands to the drone, or alter its software.

Previous research has focused on analysing the wireless communication protocols of commercial drones [3], and exploiting the lack of security measures in the communication channel [4, 5]. However, we are not aware of any security analysis of the drone's software. In this paper, we tackle this problem, and carry out a software security analysis of the MAVLink protocol, which is expected to become a world-wide standard within the DroneCode project [2]. More specifically, we investigate potential design or implementation protocol flaws using fuzzing techniques. The goal is to inject invalid or semi-invalid data to produce an unexpected software behaviour. We briefly formulate three different research questions that we would like to explore more in detail in the rest of this paper. This includes: (i) *how can we identify software security flaws in the MAVLink framework?*, (ii) *what are the consequences of exploiting these security flaws?* and (iii) *can we provide countermeasures to mitigate such issues?*.

2 Related Work

Most of the research conducted in the past years resulted in attacks against the security of drones. When a vulnerability was found, the vulnerability was exploited and the drone could be hijacked or could crash. Academic research had the goal of reproducing a certain attack, make a theoretical background or proof and try to come up with countermeasures to assure the security of the drone. A large portion of the

conducted research is about GPS spoofing. A GPS spoofing attack attempts to mislead the drone's GPS receiver by broadcasting fake GPS signals while pretending to be a legitimate GPS signal sent by a satellite. This attack can trick any device using GPS signals into changing its trajectory or make the device believe that it is at another location [11]. Other research focuses on the lack of security mechanisms like authentication and encryption. There is lot of bidirectional communication between the drone and the GCS. Many different types of communication channels can be used for this like WiFi, Bluetooth or Radio Channels. The major problem with these communication channels is that they are used without any form of encryption or that they are used together with weak encryption that can be cracked. Some research about the vulnerabilities of MAVLink includes adding encryption to the protocol [12, 5, 4, 13], but this has not yet been implemented in the MAVLink protocol. We want to look at the vulnerability analysis from another perspective. We want to search the space of software vulnerabilities of the MAVLink protocol. As far as we know, the fuzz-testing method has not yet been applied for analysis of the MAVLink protocol.

3 Background information

3.1 Fuzzing

Fuzzing is a technique for finding vulnerabilities and bugs in software programs and protocols by injecting malformed or semi-malformed data. The injected data may include minimum / maximum values and invalid, unexpected or random data. Subsequently the system can be observed to find any kind of unexpected behavior, e.g. if the program crashes. There are three main types of fuzzing variants that can be distinguished: Plain Fuzzing, Protocol Fuzzing and State-based Fuzzing [15, 16]. `Plain Fuzzing` is the most simple way of testing. The input data is usually made by changing some parts of correct input that has been recorded. It provides very little assurance on code coverage because it does not go very deep into the protocol [14]. In `Protocol Fuzzing`, the input is generated based on the protocol specifications like packet format and dependencies between field. This is called `Smart generation` and is able to create semi-valid input. The opposite, `dumb generation`, is the corruption of data packets without awareness of the data structure. Protocol fuzzers typically generate test cases with minimum values, maximum values [14, 17]. `State-based Fuzzing` is a fuzzing technique that does not try to find errors and vulnerabilities by changing the content of the packets, but instead attempts to fuzz the state-machine of the software [14]. The most common method is to start with a dumb and basic fuzzer and then increase the amount of intelligence when necessary to create a smarter fuzzer [18]. Depending on the availability of source code, we can also distinguish between `Black-box Fuzzing` and `White-box Fuzzing`. The actual techniques used for fuzzing are typically a combination of black-box or white-box fuzzing with dumb or smart fuzzing. Unlike black-box fuzzing, white-box fuzzing requires a greater testing effort, however it provides a better test coverage [19].

3.2 MAVLink

The Micro Air Vehicle Communication Protocol (MAVLink Protocol) is a point-to-point communication protocol that allows two entities to exchange information. It is used for bidirectional communications between the drone and the GCS. MAVLink is a part of the DroneCode project, governed by the Linux Foundation [20]. A MAVLink message is sent bytewise over the communication channel, followed by a checksum for error correction. If the checksum does not match, then it means that the message is

corrupted and will be discarded. Figure. 1 shows the structure of a MAVLink message. We will now give a brief description of the fields included in the message:

- ◇ *Magic*: indicates the beginning of a new messages.
- ◇ *Length*: indicates the length of the payload field.
- ◇ *Sequence number*: indicates the sequence number of the packet.
- ◇ *System ID*: ID of the sending system.
- ◇ *Component ID*: ID of the sending component.
- ◇ *Message ID*: ID of the message in the payload.
- ◇ *Payload*: payload of the packet, which contains the parameters of the message.
- ◇ *CRC*: checksum for validation.

MAVLink messages are handled by the `handleMessage(msg)` function. This function has a switch statement, handling the different message IDs [21].

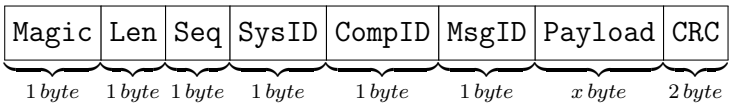


Figure 1: MAVLink packet structure

3.3 Fuzzing Methodology

For our experiments we built a fuzzer capable of creating custom MAVLink messages. Several strategies are applied to construct the messages that are sent to the drone. Initially, we started with a random dumb fuzzing to observe how the software handles invalid messages. We then made a smarter fuzzer which takes into account the message format, and constructs semi-valid messages. The techniques for constructing the payload of the messages are different for every test case.

4 Methodology

4.1 Lab Setup

Our laboratory setup employs the Software In The Loop environment (SITL) [7], which provides simulators for the ArduCopter, ArduPlane and ArduRover. We use the drone simulator for the ArduCopter [1]. The simulator is run on a Linux virtual machine and the fuzzer on a host machine. The host system is running OS X El Capitan (8gb RAM, 2,4 GHz Intel Core i5) and the virtual machine for the virtual drone is running Ubuntu 14.04 TLS 64-bit. The communication between both machines is via a TCP connection.

Configuration:

Name	Specification	Function	IP-address
Host System	Mac OSX	Host	192.168.56.1
System 2	VM2: Ubuntu	Virtual Drone	192.168.56.102

4.1.1 Case Studies

Our fuzzer, which is implemented in Python, is capable of constructing valid MAVLink messages. We now discuss how every field in the packet is constructed in our fuzzer. The `Magic` is a fixed value and is set to `"fe"`, whereas the *Length* field is set to the size of the payload field. In every transmitted message the `Seq` is increased by one, and it is reset to zero if it reaches the value of 255. The `SysID` and `CompID` are kept fixed, i.e. `"ff"` and `"00"`, respectively. The `MsgId` is a value in the range of `0-255`. The `payload` contains the parameters that are used internally, (e.g. the height of the drone), and is generated based on different strategies, which we will discuss more in detail for each experiment. The `CRC` value is generated using a CRC-16 function; its polynomial generator is 0x1021, the initial value is `FFFF`, the input data bytes are reversed, and the CRC result is reversed before the final XOR operation. The CRC parameters were obtained by looking at the available documentation and testing the generator on [23]. The input to the generation function is as follows: *Length+seq+SysID+CompID+MsgID+Payload+Seed*.
The seed is a x25 checksum generated over the message name, followed by the type and name of each field. This seed is used to capture changes in the XML describing the message definitions. This results in messages being rejected by the recipient if they do not have the same XML structure.

There are some properties that a fuzzer needs to have. A fuzzer must be able to record the test cases for reproduction. Therefore, every constructed message is written to a file before it is sent to the virtual drone. Another property is the ability of transmitting the test cases to the system under test. Since we are using SITL with a TCP connection therefore, the fuzzer initiates a three-way handshake with the virtual drone and a connection is established via sockets. The generated messages can now be sent to the drone over this socket. To observe the behaviour of the drone, we can use different approaches. A simple observation is to check whether or not the connection with the drone is still alive. To further investigate its behaviour, the virtual drone can be run inside gdb [22].

To start the virtual drone, the command in Listing 1 is used. This starts the Arducopter simulation for a quadcopter, at a certain location with all of the memory erased and faster operation. We define the following test cases.

Listing 1: Startup command

```
./arducopter.elf --home -35,149,584,270 --model quad
--speedup 100 --wipe
```

`Test Case 1` We establish a connection to the drone and start to send completely random data including numbers, letters and characters. The actual structure of a MAVLink message is ignored at the moment. The length of the data sent to the virtual drone ranges from 1 to 1000 characters. We do this to test how the software handles incorrect data.

`Test Case 2` For every message ID, we create a message with payloads of length ranging from the minimum length (i.e. 1 byte) to the maximum length (i.e. 255 bytes). The payload consists of completely random combinations of hexadecimal values. We repeat this test several times. This test is used to give an indication of how semi-valid messages are handled. This is important since these messages can go deeper into the software.

`Test Case 3` We also test the system's behaviour when messages without any payload are sent. For every message id, a new message (with no payload) will be constructed with the length set to zero. This test aims to find vulnerabilities that do not depend on the payload.

`Test Case 4` We construct payloads consisting entirely out of the minimum value. This test investigates how the implementation handles the minimum value `"00"`. This is done for payloads with a length ranging from 1 to 255 bytes.

`Test Case 5` following the previous test case, we do the same for the maximum value `"ff"`.

`Test Case 6` Within this test case, we do not send the messages byte per byte. An entire message with all the necessary fields is included in one TCP packet. We incrementally increase the length of the random payload from 1 to 255 bytes extending the length of the entire message.

`Test Case 7` Within this test case, we try to identify vulnerabilities depending on the value of the length field and the actual value of the payload. In a first run, we constructed messages with up to 5 bytes of payload and set the value of the length field to the length of the payload minus one. In the second run, we did the same, except that the value of the length field was set to the length of the payload plus one.

5 Results Obtained and Discussion

Resulting from the listed test cases, we were able to identify a few security flaws. Particularly, from the sixth test case, where the payload increased randomly, the fuzzing script was able to crash the virtual drone. The error caused by the fuzzing script can be seen in Listing. 2.

Listing 2: Flaoting Point Exception

```
ERROR: Floating point exception - aborting
Aborted (core dumped)
```

To investigate the cause of the exceptions we used the gdb debugger and the core dump of the memory when the kernel crash occurred. From an analysis we identified that errors corresponds to three specific functions. However, further investigation needs to be performed to identify the exact cause of the exceptions.

The next step of our work is to complete the entire range of the test cases aiming to gain more results identifying error flaws of the MAVLink software implementation. Moreover, we aim to further investigate the causes of the identified software flaws. However, we have to stress that fuzzing all possible test cases is a resource demanding operation. For instance, there is a limitation concerning to the memory usage. With the current setup, it is not possible to try all possible permutations of payloads, for all possible message and payload lengths. Currently, we are looking to further improve the fuzzing scripts aiming to make the fuzzing operations more memory efficient.

6 Conclusion

This work aims to identify software vulnerabilities, by using the technique of fuzzing. Currently we focused our research on the MAVLink protocol. MAVLink is used as a communication protocol between a drone and a ground control station. The protocol is actively developed by the community and aims to become one of the drone communication standards. Our aim is to contribute to the identification of the security flaws of the protocol and to help the development community to mitigate these flaws. At the same time, we want to proof the suitability of fuzzing techniques for discovering vulnerabilities in the implementation of the protocol. Currently, we have identified software vulnerabilities that we are further investigating.

Many fuzzing platforms do already exist and can be used for the software analysis of the MAVLink protocol. Our next step is to further research and extend our fuzzing scripts aiming to generate more complex fuzzing scenarios.

7 Acknowledgements

This work was supported in part by the Research Council KU Leuven (C16/15/058).

References

[1] ArduCopter, https://www.dronecode.org/, 24 03 2016.

[2] DroneCode, http://ardupilot.org/copter/index.html, 24 03 2016.

[3] B. Hond, *Fuzzing the GSM Protocol*, Radboud University Nijmegen, Netherlands, 2011.

[4] J. A. Marty, *Vulnerability Analysis of the MAVLink Protocol for Command and control of Unmanned Aircraft* Air Force Institute of Technology, USA, 2014.

[5] N. Butcher, A. Stewart and Dr. S. Biaz , *Securing the MAVLink Communication Protocol for Unmanned Aircraft Systems*, Appalachian State University, Auburn University, USA, 2013.

[6] Sulley Manual, http://www.fuzzing.org/wp-content/SulleyManual.pdf, 24 03 2016.

[7] Software In The Loop, http://ardupilot.org/dev/docs/sitl-simulator-software-in-the-loop.html, 24 03 2016.

[8] AR drone that infects other drones with virus wins dronegames, http://spectrum.ieee.org/automaton/robotics/diy/ar-drone-that-infects-other-drones-with-virus-wins-dronegames, 03 05 2016.

[9] SkyJack, https://github.com/samyk/skyjack, 03 05 2016.

[10] Maldrone, http://garage4hackers.com/entry.php?b=3105, 03 05 2016.

[11] GPS Spoofing, https://capec.mitre.org/data/definitions/628.html, 03 05 2016.

[12] Thomas M. DuBuisson, Galois, Inc.1 , *SMACCMPilot Secure MAVLink Communications*, Galois, Inc.1, 2013.

[13] MAVLink 2.0 packet signing proposal, https://docs.google.com/document/d/1ETle6qQRcaNWAmpG2wz0oOpFKSF_bcTmYMQvtTGI8ns, 03 05 2016.

[14] B. Hond, *Fuzzing the GSM Protocol*, Radboud University Nijmegen, 2011.

[15] A. Takanen, C. Miller, J. DeMott *Fuzzing for Software Security Testing and Quality Assurance*, ARTECH HOUSE, INC., 2008.

[16] A. Greene, M. Sutton, P. Amini *Fuzzing Brute Force Vulnerability Discovery*, Addison-Wesley, 2007.

[17] OWASP Fuzzing, https://www.owasp.org/index.php/Fuzzing, 03 05 2016.

[18] 15 Minute Guide to Fuzzing, https://www.mwrinfosecurity.com/our-thinking/15-minute-guide-to-fuzzing/, 03 05 2016.

[19] J.eystadt, *Automated Penetration Testing with White-Box Fuzzing*, Microsoft Corporation, 2008.

[20] MAVLink Protocol, `http://qgroundcontrol.org/mavlink/start`, 03 05 2016.

[21] S. Balasubramanian, MAVLink Tutorial for Absolute Dummies (part-I), `http://dev.ardupilot.com/wp-content/uploads/sites/6/2015/05/MAVLINK_FOR_DUMMIESPart1_v.1.1.pdf`, 03 05 2016.

[22] GDB: The GNU Project Debugger , `https://www.gnu.org/software/gdb/`, 03 05 2016.

[23] CRC Generator, `http://www.zorc.breitbandkatze.de/crc.html`, 03 05 2016.

Parallel optimization on the Entropic Cone

Benoît Legat Raphaël Jungers
Université catholique de Louvain
ICTEAM
4 Av. G. Lemaître
1348 Louvain-la-Neuve, Belgium
benoit.legat@student.uclouvain.be raphael.jungers@uclouvain.be

Abstract

We introduce a parallelizable algorithm for approximate optimization on the entropic cone. We also present the toolbox `EntropicCone.jl`. Its aim is to improve the computational repoducibility of the recent progress on the approximation of the entropic cone and to make them easily accessible for its many applications. These applications include the capacity region of multi-source network coding, converse theorems for multi-terminal problems of information theory, bounds on the information ratios in secret sharing schemes and conditional independence among subvectors of a random vector.

1 The problem setting

Given n random variables, we can compute the entropy of any of the 2^n subsets of these n variables. The set $\mathcal{E}_n \triangleq \mathbb{R}^{2^n-1}$ of vectors indexed by the nonempty subsets of $[n] \triangleq \{1, \ldots, n\}$ is called the *entropy space*.The entropy vector of a set of n random variables is the entropy vector h such that h_I is the entropy of the set $\{ X_i \mid i \in I \}$.

We denote the set of vectors of $\mathbb{R}^{2^n-1}$ that are entropic as:

$$\mathcal{H}_n \triangleq \{ h \in \mathbb{R}^{2^n-1} \mid \exists X_1, \ldots, X_n, \forall \emptyset \neq S \subseteq [n], h_S = H_b(\{ X_i \mid i \in S \}) \}.$$

It is known that the set $\mathcal{H}_n$ is not a cone for $n \geq 3$ but its closure $\operatorname{cl} \mathcal{H}_n$ is a convex cone [20]. The difference between $\mathcal{H}_n$ and $\operatorname{cl} \mathcal{H}_n$ is only on the boundary of $\operatorname{cl} \mathcal{H}_n$. More precisely, it has been shown that the relative interior of $\operatorname{cl} \mathcal{H}_n$ is contained in $\mathcal{H}_n$ [15].

For $n \leq 3$, $\operatorname{cl} \mathcal{H}_n$ is equal to the *polymatroid cone* $\mathcal{P}_n$. This is the set of entropy vectors h that are

- $\diamond$ *nonnegative*: $h_I \geq 0$ for any $I \subseteq [n]$,
- $\diamond$ *nondecreasing*: $h_I \leq h_J$ for any $I \subseteq J \subseteq [n]$ and
- $\diamond$ *submodular*: $h_J + h_K \geq h_{J \cup K} + h_{J \cap K}$ for any $J, K \subseteq [n]$.

These three sets of conditions are linear inequalites on the entropy vector h. Since $\mathcal{P}_n$ is defined by a finite subset of linear inequalities, it is a polyhedral cone.

For $n \geq 4$, $\operatorname{cl} \mathcal{H}_n$ is a strict subset $\mathcal{P}_n$. Moreover, $\operatorname{cl} \mathcal{H}_n$ is not polyhedral [14] and not even semialgebraic* [17].

The entropic cone has a varietie of applications including the capacity region of multi-source network coding [1], converse theorems for multi-terminal problems of information theory [19], bounds on the information ratios in secret sharing schemes [2]

*A set is *semialgebraic* if it is the projection of an algebraic set. A set is *algebraic* if it can be defined by finitely many polynomial inequalities.

and conditional independence among subvectors of a random vector [18]. This motivates the research on filling the gap between $\mathcal{P}_n$ and $\operatorname{cl}\mathcal{H}_n$.

In Section 2, we review the methods used to find tighter approximations of $\mathcal{H}_n$ than $\mathcal{P}_n$. The current methods are linear and generate polyhedral outer approximations. However, as one can anticipate, the number of facets of a polyhedral approximation that would be "everywhere close" to $\mathcal{H}_n$ would have a sizeable amount of facets since it is high-dimensional and not semialgebraic. In practical applications, one is often looking at the simpler problem of solving an optimization problem involving the entropic cone so we are only looking for an outer approximation that is "close" to $\mathcal{H}_n$ "near" the optimum. In Section 3, we show a parallelizable algorithm for this problem.

2 Generating Non-Shannon Inequalities

The inequalities that are valid for $\mathcal{P}_n$ are called *Shannon inequalities* and those that are valid for $\mathcal{H}_n$ but not for $\mathcal{P}_n$ are called *non-Shannon inequalities*.

The current approach in generating an outer bound for $\mathcal{H}_n$ is to generate non-Shannon inequalities and to intersect $\mathcal{P}_n$ with the halfspaces they define.

There are currently four known methods for generating non-Shannon inequalities. The first one, which is the most commonly used, was introduced by Zhang and Yeung in order to generate the first non-Shannon inequality [21]; its description and analysis can be found in [13]. The three other methods are respectively described in [12], [15] and [6].

The first two methods are equivalent [10] and it is still unknown whether the third and four methods can generate inequalities that cannot be generated by the first two ones. It is also unknown whether these four methods can generate all non-Shannon inequalities. In this paper, we will only use the first method which is described in Section 2.1.

2.1 Entropic Cone and adhesivity

We define the *inner-adhesivity* and *self-adhesivity* operators respectively as

$$\mathsf{ia}_{J,K|I}(h) = \{\, g \in \mathcal{E}_{n'} \mid \forall I \subseteq L \subseteq J \cup K, g_L = h_{L\cap J} + h_{L\cap K} - h_I \,\}, \qquad I = J \cap K,$$
$$\mathsf{sa}_{J|I}(h) = \{\, g \in \mathcal{E}_{n'} \mid \forall I \subseteq L \subseteq J' \cup K, g_L = h_{L\cap J'} + h_{L\cap K} - h_I \,\}, \quad I \subseteq J.$$

where for inner-adhesivity, $n' = |J \cup K|$ and for self-adhesivity $n' = n + |J \setminus I|$, $J' = [n]$ and $K = ([n'] \setminus J') \cup I$.

Definition 1. We say that a family of sets $\mathcal{S}_n \subseteq \mathcal{E}_n$ is *inner-adhesive* if for any n, $x \in \mathcal{S}_n$ and $J, K \subseteq [n]$, there exists $y \in \mathcal{S}_{|J\cup K|}$ such that $y \in \mathsf{ia}_{J,K|J\cap K}(x)$ and we say that it is *self-adhesive* if for any n, $x \in \mathcal{S}_n$ and $I \subseteq J \subseteq [n]$, there exists $y \in \mathcal{S}_{n+|J\setminus I|}$ such that $y \in \mathsf{sa}_{J|I}(x)$.

The following theorem gives the relation between adhesivity and the entropic cone.

Theorem 1. The entropic cone $\mathcal{H}_n$ is inner-adhesive and self-adhesive.

Proof. Consider $h \in \mathcal{H}_n$ and n jointly distributed random variables X_i of joint probability mass function p such that h is their entropy vector. Let p_I denote the probability mass function of the marginal distribution of I. For inner-adhesivity, the entropy vector $g \in \mathcal{H}_{|\mathcal{S}\cup\mathcal{T}|}$ of the probability function

$$\frac{p_J(x_J)p_K(x_K)}{p_I(x_I)} = p_{J|I}(x_J|x_I)p_{K|I}(x_K|x_I)p_I(x_I)$$

belongs to $\mathsf{ia}_{J,K|I}(h)$ and for self-adhesivity, the entropy vector $g \in \mathcal{H}_{n+|J\setminus I|}$ of the probability function

$$\frac{p_{J'}(x_{J'})p_J(x_K)}{p_I(x_I)} = p_{J'|I}(x_{J'}|x_I)p_{J|I}(x_K|x_I)p_I(x_I),$$

where $J' = [n]$, belongs to $\mathsf{sa}_{J|I}(h)$. □

Let $\mathsf{ia}^{-1}_{J,K|I}(\mathcal{S})$ (resp. $\mathsf{sa}^{-1}_{J|I}(\mathcal{S})$) be the set of vectors x such that there exists $y \in \mathcal{S}$ such that $y \in \mathsf{ia}_{J,K|I}(x)$ (resp. $y \in \mathsf{sa}_{J|I}(\mathcal{S})(x)$). By Theorem 1, given an integer n_0 and a sequence $(J_1, I_1), \ldots, (J_m, I_m)$ such that $I_i \subseteq J_i \subseteq [n_i]$ and $n_i = n_{i-1} + |J_i \setminus I_i|$ for $i = 1, \ldots, m$, the set[†]

$$(\mathsf{sa}^{-1}_{J_1|I_1} \circ \cdots \circ \mathsf{sa}^{-1}_{J_m|I_m})(\mathcal{P}_{n_m})$$

provides an outer approximations of $\mathcal{H}_{n_0}$. Therefore an outer approximation of $\mathcal{H}_{n_0}$ can be obtained by projecting the polyhedral cone of dimension $\sum_{i=0}^{m} 2^{n_i} - 1$ given by

$$\{\, (h_0, h_1, \ldots, h_m) \in \mathcal{P}_{n_0} \times \mathcal{P}_{n_1} \times \cdots \times \mathcal{P}_{n_m} | h_i \in \mathsf{sa}_{J|I}(h_{i-1}), i = 1, \ldots, m \,\}. \quad (1)$$

on the first $2^{n_0} - 1$ variables.

This method was used to generate hundreds of non-Shannon inequalities in [8]. Using Benson's algorithm [3] to compute the projection of (1), even more non-Shannon inequalities were uncovered in [7, 11].

As mentioned earlier, $\mathrm{cl}\,\mathcal{H}_n$ is strictly included in $\mathcal{P}_n$ for $n \geq 4$. As the dimension of $\mathcal{H}_n$ is exponential in n, the methods are usually benchmarked using $\mathcal{H}_4$. An important numerical quantity, related with the geometric properties of $\mathcal{H}_4$ is the *Ingleton score* defined as

$$\mathbb{I}^* \triangleq \inf_{0 \neq h \in \mathcal{H}_4} \mathbb{I}_{ij}(h)$$

where $\mathbb{I}_{ij}(h) = \langle \square_{ij}, h \rangle / h_{[n]}$ [8, Definition 3] and $\square_{ij}$ is the Ingleton dual entropy vector [9]. The current best lower bound on $\mathbb{I}^*$ is equal to -0.15789 [8]. Upper bounds on $\mathbb{I}^*$ can be obtained from exhibiting four jointly distributed variables for which the entropy vector has low Ingleton score. The current best upper bound on $\mathbb{I}^*$ is equal to -0.09243 [16].

3 Parallelizable optimization on the Entropic Cone

In this section, we show how to decompose polyhedra such as described by (1) to solve optimization problems on it in an efficient and parallelizable manner using ideas from *Stochastic Programming* [5]. In stochastic programming, large scale linear programs are decomposed into smaller linear programs linked together by a markov chain. The linear program at each state u of the markov chain is:

$$\begin{aligned} Q(x, u) = \text{minimize } & c^T y + \mathcal{Q}_u(y) \\ \text{s.t. } & W_u y = h_u - T_u x, \\ & x \geq 0 \end{aligned}$$

where $\mathcal{Q}_u(y)$ is the sum of $Q(x, v)$ for each state v accessible from u weighted by the probability to go from state u to state v. When the program is infeasible for some x,

[†]Of course this also works if we include inner-adhesive operations in the sequence, we have only included self-adhesivity in the sequence to keep simple notation.

$Q(x,u) = \infty$. At the initial state of the markov chain, there is no term $-T_u x$ and the solution at this stage is the solution of the original large scale linear program. Note that if v is accessible from different states u, u', the number of variables at u and u' must match for $Q(\cdot, v)$ to be well-defined.

For the adhesive operations, we propose to define a state for each dimension n and adhesive operation. That is, for every n, J, K, we have a state for the operation $\mathsf{ia}_{J,K|J\cap K}$ and for every n, J, I with $I \subseteq J$, we have a state for the operation $\mathsf{sa}_{J|I}$. The linear program at the initial state is

$$\begin{aligned} &\text{minimize } c^T h + \mathcal{Q}_0(h) \\ &\text{s.t. } h \in \mathcal{P}_{n_0}, \end{aligned}$$

the linear program for each state representing an inner-adhesivity is

$$\begin{aligned} Q(h, (n, \mathsf{ia}_{J,K|J\cap K})) = \text{minimize } & \mathcal{Q}_{(n,\mathsf{ia}_{J,K|J\cap K})}(g) \\ \text{s.t. } & g \in \mathsf{ia}_{J,K|J\cap K}(h) \\ & g \in \mathcal{P}_{|J\cup K|} \end{aligned}$$

and the linear program for each state representing an self-adhesivity is

$$\begin{aligned} Q(h, (n, \mathsf{sa}_{J|I})) = \text{minimize } & \mathcal{Q}_{(n,\mathsf{sa}_{J|I})}(g) \\ \text{s.t. } & g \in \mathsf{sa}_{J|I}(h) \\ & g \in \mathcal{P}_{n+|J\setminus I|}. \end{aligned}$$

A state (n, a) is accessible from a state u if the dimension[‡] of the linear program at u is $2^n - 1$. Since there is no objective in states other than the initial state, the probability assigned for each transition does not matter. The only relevant information in $Q(h, (n, \mathsf{a}))$ is whether it is infinite or zero.

In stochastic programming, the domain of $Q(\cdot, (n, \mathsf{a}))$ is approximated by a polyhedron by starting with the approximation $\mathbb{R}^n$ and adding *feasibility cuts*. The feasibility cuts are computed as follows: the linear program is solved for some h, if it is infeasible, an unbounded ray of the dual linear program is computed and used to generate a feasibility cut.

These observations lead to Algorithm 1 for approximate optimization on the entropic cone that is inspired from the *Stochastic Dual Dynamic Programming* algorithm used in stochastic programming. Note that the initial node can have a nonlinear objective function and additionnal nonlinear constraints. This algorithm is easily parallelizable as the linear programs of the different states only need to communicate cuts and optimal solutions.

We developed a new toolbox `EntropicCone.jl` in Julia [4] for working with the entropic cone. This algorithm is one of the features implemented in the toolbox. We tested Algorithm 1 to find lower bounds for the Ingleton score and obtained the best known lower bound -0.15789 in under a minute.

4 Conclusion

Searching for non-Shannon inequalities to provide tighter outer approximations of the entropic cone may seem computationally demanding due to the use of a projection algorithm in high-dimensional space. However, if we restrict ourself to the optimization

[‡]the number of variables of the linear program

Algorithm 1 Approximate minimization of $c(h)$ subject to $x \in \mathcal{H}_n \cap \mathbb{F}$ for parameters $n_{\max}, K, m, \rho$.

Given a maximal value $n_{\max}$ for n, generate all states such that the dimension of the linear program is at most $2^{n_{\max}} - 1$
for $k = 1, 2, \ldots, K$ **do**
Pick a set P of ρ random paths of length m starting at the initial state.
Solve the optimization program

$$\begin{aligned} &\text{minimize } c(h) \\ &\quad \text{s.t. } h \in \mathcal{P}_{n_0} \cap \mathbb{F} \cap \operatorname{dom}(\mathcal{Q}_0), \end{aligned}$$

where the domain of $\mathcal{Q}_0(h)$ is approximated by the feasibility cuts.
if the program is infeasible **then**
return Infeasible
end if
for $i = 1, 2, \ldots, m$ **do**
for all $p \in P$ **do**
Solve $Q(h, (n_{p,i}, \mathsf{a}_{p,i}))$ where h is the value of the optimal solution of the previous state in the path.
if the program is infeasible **then**
Add a feasibility cut for $\operatorname{dom}(Q(\cdot, (n_{p,i}, \mathsf{a}_{p,i})))$
Remove p from P
end if
end for
end for
end for

of a (possibly nonlinear) objective on the entropic cone (possibly under additionnal constraints), the algorithms used in stochastic programming can be used to provide a parallelizable algorithm that can provide bounds on the objective.This method is able to obtain current best lower bound on the Ingleton score in under a minute.

References

[1] Riccardo Bassoli, Hugo Marques, Jose Rodriguez, Kenneth W Shum, and Rahim Tafazolli. Network coding theory: A survey. *Communications Surveys & Tutorials, IEEE*, 15(4):1950–1978, 2013.

[2] Amos Beimel. Secret-sharing schemes: a survey. In *Coding and cryptology*, pages 11–46. Springer, 2011.

[3] Harold P Benson. An outer approximation algorithm for generating all efficient extreme points in the outcome set of a multiple objective linear programming problem. *Journal of Global Optimization*, 13(1):1–24, 1998.

[4] Jeff Bezanson, Stefan Karpinski, Viral B Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

[5] John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[6] Laszlo Csirmaz. Book inequalities. *Information Theory, IEEE Transactions on*, 60(11):6811–6818, 2014.

[7] László Csirmaz. Using multiobjective optimization to map the entropy region. *Computational Optimization and Applications*, 63(1):45–67, 2016.

[8] Randall Dougherty, Chris Freiling, and Kenneth Zeger. Non-Shannon information inequalities in four random variables. *arXiv preprint arXiv:1104.3602*, 2011.

[9] AW Ingleton. Conditions for representability and transversality of matroids. In *Théorie des Matroïdes*, pages 62–66. Springer, 1971.

[10] Tarik Kaced. Equivalence of two proof techniques for non-Shannon-type inequalities. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 236–240. IEEE, 2013.

[11] Andreas Löhne and Benjamin Weißing. The vector linear program solver Bensolve–notes on theoretical background. *arXiv preprint arXiv:1510.04823*, 2015.

[12] Konstantin Makarychev, Yury Makarychev, Andrei Romashchenko, and Nikolai Vereshchagin. A new class of non-Shannon-type inequalities for entropies. *Communications in Information and Systems*, 2(2):147–166, 2002.

[13] František Matúš. Adhesivity of polymatroids. *Discrete Mathematics*, 307(21):2464–2477, 2007.

[14] František Matúš. Infinitely many information inequalities. In *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*, pages 41–44. IEEE, 2007.

[15] František Matúš. Two constructions on limits of entropy functions. *Information Theory, IEEE Transactions on*, 53(1):320–330, 2007.

[16] F. Matúš and L. Csirmaz. Entropy region and convolution. *ArXiv e-prints*, October 2013.

[17] Carolina Mejia and J Andres Montoya. The almost-entropic regions are not semi-algebraic. *arXiv preprint arXiv:1510.02658*, 2015.

[18] Milan StudenÃ¡. *Probabilistic conditional independence structures*. Springer Science & Business Media, 2006.

[19] Raymond W Yeung. *A first course in information theory*. Springer Science & Business Media, 2012.

[20] Zhen Zhang and Raymond W Yeung. A non-Shannon-type conditional inequality of information quantities. *Information Theory, IEEE Transactions on*, 43(6):1982–1986, 1997.

[21] Zhen Zhang and Raymond W Yeung. On characterization of entropy function via information inequalities. *Information Theory, IEEE Transactions on*, 44(4):1440–1452, 1998.

Compute-and-forward on the Multiple-access Channel with Distributed CSIT

Shokoufeh Mardani, Jasper Goseling

Stochastic Operations Research, University of Twente, The Netherlands

{s.mardanikorani, j.goseling}@utwente.nl

Abstract

This paper considers the two-user block-fading multiple access channel in which each user knows its own channel gain, but not the channel gain of the other user (distributed CSIT). The receiver has complete channel state information. We consider the use of lattice codes and a compute-and-forward strategy in which multiple linear combinations of the messages are decoded in order to recover the messages themselves. Users choose their rate and scale their lattice as a function of their channel gain. It is shown that under certain constraints on the channel gains this strategy is sum-rate optimal if no outage is tolerated. Finally, the strategy is extended to the case that outage is allowed.

1 Introduction

Lattice codes have received great interest after Erez and Zamir showed that they can achieve the capacity of the additive white Gaussian Noise (AWGN) point-to-point channel [1]. It was, for instance, shown that lattice codes can achieve the capacity of the Gaussian Multiple-access channel (MAC) [2] which is commonly used to model uplink communication in wireless networks. In wireless communications there is signal fading owing to multipath transmission of signal, shadowing or inherent variability of the channel, which makes the communication less reliable. In [2], the full capacity region of the multiple-access channel in a fading scenario was achieved by using lattice codes. More precisely, in [2] it is assumed that both transmitters and the receiver have access to the full channel state information (full CSI).

In this paper we consider the case of distributed channel state information at the transmitters. More precisely, we consider a slow-fading two-user symmetric Gaussian multiple access channel in which each user knows its own channel gain, but not the channel gain of the other user (distributed CSIT). The receiver has complete channel state information. For this channel we consider the adaptive-rate capacity, in which users adapt their rate and power in each block and no outage is tolerated, i.e. the employed rate-pairs of transmitter in each block should be inside the instantiated MAC capacity region. The adaptive-rate capacity of this channel is given in [3]. In [4, 5] a simple distributed strategy is presented for choosing power and rate. It is shown that this strategy achieves the adaptive sum-rate capacity for the case that the channel state statistics are symmetric. Moreover, a rate-splitting strategy is presented in [4] that enables successive decoding at the receiver, providing an alternative to the joint typicality decoding given in [3].

The contribution of the current paper is to use lattice codes in a compute-and-forward framework [6]. Two linear combinations of the messages are decoded in order to recover the messages themselves. By this we provide another practical encoding and decoding mechanism for the multiple-access channel with distributed CSIT. Similar to the strategies that are presented in [2] we provide scaling of the lattices at the encoders to deal with the varying channel gains. The difference of our strategy compared to [2] is that in our case the rate and lattice scaling can be a function of the channel state of one user only. Our main result is that under certain constraints on the channel

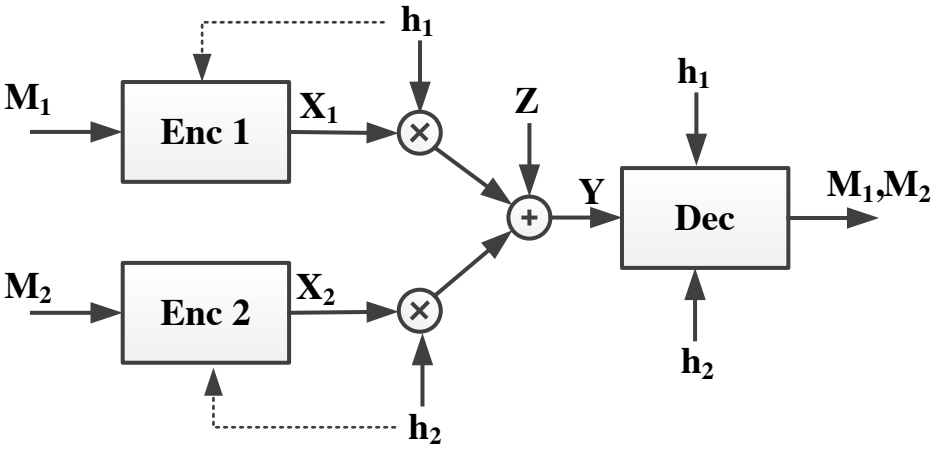


Figure 1: Channel Model.

gains this strategy is adaptive sum-rate optimal in the sense that we achieve the same rates as the adaptive sum-rate optimal scheme of [4]. In addition to our main result we provide an extension of our strategy by allowing for outage in some rounds. We demonstrate that this improves that expected sum rate.

The channel model that we study in this paper is of relevance in a random-access scenario, in which users only sporadically have messages to transmit. In this case we cannot assume complete channel state information at all transmitters. In [7] the use of compute-and-forward for random access is considered for the case of unit channel gains. In this work we move away from the assumption of unit channel gains. In [8] this assumption is also leveraged, but a different strategy is used to deal with the varying channel gains. Each user that has a sufficiently good channel inverts the channel, effectively providing equal channel gains from the receiver's perspective. This strategy, though simple and effective, does lead to outage if the channel quality of a user is not good enough. It is, moreover, not known how far this strategy is away from optimality.

The remainder of this paper is organized as follows. In Section 2 we present the channel model, problem statement and notation. In Section 3 we review some of the background on lattices and present a result from [2] that will be used later. In Section 4 we provide our main result an achievable strategy for the adaptive rate, outage free, scenario. In Section 5 we extend our strategy to the setting that allow for outage. Finally, in Section 6 we provide conclusions and an outlook on future work.

2 Model and Problem Statement

We consider the 2-user symmetric Gaussian MAC

$$y = h_1 x_1 + h_2 x_2 + z, \tag{1}$$

where $x_i \in R$ is the transmitted message by user i, $h_i \in R$ is the fading coefficient of the channel from user i to the receiver, and z is additive white Gaussian noise with unit variance. We consider symmetrical users (i.e. they have same power constraints and channel fading distributions). The channel fading coefficients are i.i.d and remain

fixed for a sufficiently large blocks over which the codewords last. We assume the fading states have Rayleigh distribution and each transmitter has access to own fading coefficient and can use this information to adapt its rate. All fading coefficients are known at the receiver.The model is illustrated in Figure 1.

In the proposed coding scheme, we represent the messages as M_1 and M_2. User i has a message $M_i \in \{1, \ldots, 2^{nR_i}\}$ to transmit and maps it to the channel input $x_{i,1}, \ldots, x_{i,n}$, satisfying a fixed power constraint P. We will refer to the *throughput* of a strategy as the expected sum-rate, namely,

$$T = E[R_1(h_1)] + E[R_2(h_2)], \tag{2}$$

where averages are over all values of h_1 and h_2, respectively. Note that it is possible to adapt the power over different blocks. We do not analyze such power control strategies explicitly, but instead rely on results from [4] that indicate that our proposed strategy is optimal if the right power control strategy is used.

In this work we consider MAC with individual CSI which each transmitter has instantaneous access to its own fading and choose its rate based on this information. This is possible since the users have channel state information available. Therefore, in our rate-strategy, the user i employs a code-book of rate $R_i(h_i)$ according to the individual fading state h_i. Let $C_{MAC}\left(\widehat{\mathbf{h}}\right)$ denotes the instantaneous capacity region of a Gaussian MAC with fixed channel gains $\hat{\mathbf{h}}$

$$C_{MAC}\left(\hat{\mathbf{h}}\right) = \left\{ \widehat{R} : \forall S \subseteq \{1,2\}, \sum_{i \in S} R_i \leq \frac{1}{2} \log\left(1 + \sum_{i \in S} |h_i|^2 P\right) \right\}. \tag{3}$$

A strategy under individual CSI, is called outage-free if

$$\forall \hat{\mathbf{h}} \in \mathrm{H}, (R_1(h_1), R_2(h_2)) \in C_{MAC}\left(\hat{\mathbf{h}}\right), \tag{4}$$

where H is the fading space.

The goal of the present paper is to provide an encoding and decoding mechanism that is achieving the maximum achievable throughput that is outage free and that is based on lattice codes.

Throughout this paper, vectors are denoted by bold letters.

3 Preliminaries

Lattice code are a discrete subgroup of R^n and can be written as a linear transformation of integer vector as

$$\Lambda = \{\lambda = G\mathbf{x} : \mathbf{x} \in Z^n\}, \tag{5}$$

where λ is a lattice point of Λ. More details about lattice and lattice code can be found in [1]. Similar to [2] we will use a nested lattice code combined with the compute-and-forward technique of Nazer and Gastpar [6] to decode two linearly independent equations of the messages as $a_1M_1 + a_2M_2$ (first equation) and $b_1M_1 + b_2M_2$ (second equation). These equations are decoded sequentially, enabling successive cancellation techniques after decoding the first equation. The messages M_1 and M_2 can be obtained from the linear equations that are decoded.

The proposed scheme in [2] allows users to use full CSI for controlling transmitted rate by scaling second moment of the coarse lattice. In general, not the whole dominant face of the capacity region can be achieved. By defining $A = \frac{h_1 h_2 P}{\sqrt{1+h_1^2 P + h_2^2 P}}$, if $A < 3/4$

any points on the capacity region can not be achieved via this strategy, if $3/4 \leq A < 1$ parts of the capacity region can be achieved, and if $A \geq 1$ the full dominant face are achieved [2].

Assume β_1, β_2 are positive numbers, we can construct lattices $\Lambda_k^s \subseteq \Lambda$ for $k = 1, 2$ which both lattices are simultaneously good with second moment, i.e.

$$\sigma^2(\Lambda_k) = \beta_k^2. \tag{6}$$

We collect both βs into one vector $\boldsymbol{\beta}$. The following theorem gives achievable message rate-pairs for the 2-user Gaussian MAC.

Theorem 1 ([2])**.** *Consider 2-user multiple access channel in (1) with full-CSI. The following rate pair is achievable*

$$R_k = \begin{cases} r_k(\mathbf{a}, \boldsymbol{\beta}), & \text{if } b_k = 0, \\ r_k(\mathbf{b}|\mathbf{a}, \boldsymbol{\beta}), & \text{if } a_k = 0, \\ \min\{r_k(\mathbf{a}, \boldsymbol{\beta}), r_k(\mathbf{b}|\mathbf{a}, \boldsymbol{\beta})\}, & \text{otherwise.} \end{cases} \tag{7}$$

where

$$r_k(\mathbf{a}, \boldsymbol{\beta}) = \frac{1}{2}\log\frac{\beta_k^2(1 + h_1^2P + h_2^2P)}{K(\mathbf{a}, \boldsymbol{\beta})}, \tag{8}$$

$$r_k(\mathbf{b}|\mathbf{a}, \boldsymbol{\beta}) = \frac{1}{2}\log\frac{\beta_k^2 K(\mathbf{a}, \boldsymbol{\beta})}{\beta_1^2\beta_2^2(a_2b_1 - a_1b_2)^2}, \tag{9}$$

$$K(\mathbf{a}, \boldsymbol{\beta}) = \sum_k a_k^2\beta_k^2 + P(a_1\beta_1h_2 - a_2\beta_2h_1)^2. \tag{10}$$

4 Proposed Strategy

The *main contribution* of this work consists of a proposed means of scaling $\beta_1 = 1$ and β_2 as a function of h_2 and the selection of rates as following

$$R_1 = \frac{1}{4}\log_2\left(1 + 2h_1^2P\right), \quad R_2 = \frac{1}{4}\log_2\left(1 + 2h_2^2P\right). \tag{11}$$

This choice of the rates correspond to the midpoint strategy that is presented in [4]. Therefore, we know from [4] that these rates are sum-rate optimal.

In our strategy we use $\mathbf{a} = (1, 1)$ and $\mathbf{b} = (0, 1)$. For our choice of the rates, it follows from Theorem 1 that these equations can be decoded if $\beta_1 = 1$ and $\beta_2^{'} \leq \beta_2 \leq \beta_2^{''}$, or equivalently $A \geq 3/4$, where

$$\beta_2^{'}(h_1, h_2) = \frac{2h_1h_2P + S - \sqrt{SD}}{2(1 + h_1^2P)}, \quad \beta_2^{''}(h_1, h_2) = \frac{2h_1h_2P + S + \sqrt{SD}}{2(1 + h_1^2P)}, \tag{12}$$

$$S = \sqrt{1 + h_1^2P + h_2^2P}, \quad D = 4h_1h_2P - 3S. \tag{13}$$

The following theorem provides the value of β_2 that is used by our strategy.

Theorem 2. *Let β^* be a positive real variable and $0 < u < v$ be arbitrary constants. If $u \leq h_1 \leq v$ and for every values of h_1,*

$$\sqrt{1+2h_2^2P} \leq K\left(\mathbf{a}, \beta^*\right) \leq \frac{1+h_1^2P+h_2^2P}{\sqrt{1+2h_1^2P}} \tag{14}$$

and

$$\sup_{u\leq h_1\leq v} \beta_2'(h_1,h_2) \leq \beta^* \leq \inf_{u\leq h_1\leq v} \beta_2''(h_1,h_2), \tag{15}$$

where

$$\beta^* = \beta_2(h_2) = \frac{h_2^2P + \sqrt{(1+h_2^2P)\left(\sqrt{1+2h_2^2P}-1\right) - h_2^2P}}{1+h_2^2P}$$

and

$$K\left(a, \beta^*\right) = 1 + \beta^{*2} + P(h_2 - \beta^* h_1)^2,$$

then Users 1 and 2 can send at rate-pair (11) and receiver can reliably decode M_1 and M_2. Moreover, this rate pair is adaptive sum-rate optimal.

Proof. We use lattice code combined with a compute-and-forward technique such as [2] and set equation coefficient vectors to be $\mathbf{a} = (1, 1)$ and $\mathbf{b} = (0, 1)$. User 2 choose β^* in $\left[\beta', \beta''\right]$ for all values of h_1, since β_2' and β_2'' are decreasing function over h_1, this condition will be reduced to $\sup_{u\leq h_1\leq v} \beta_2'(h_1,h_2) \leq \beta^* \leq \inf_{u\leq h_1\leq v} \beta_2''(h_1,h_2)$. Hence the achievable rate points via lattice code for each values of h_1 is

$$R_1 = r_1\left(\mathbf{a}, \beta^*\right), \quad R_2 = r_2\left(\mathbf{b}|\mathbf{a}, \beta^*\right). \tag{16}$$

For reliable communication, the rate-point (11) should be inside the instantiated achievable rate region given by (16) which makes error probability arbitrarily small, or equivalently

$$r_1\left(a, \beta^*\right) \geq \frac{1}{4}\log\left(1+2h_1^2P\right), \quad r_2\left(b|a, \beta^*\right) \geq \frac{1}{4}\log\left(1+2h_2^2P\right), \tag{17}$$

which leads to the following constraint

$$\sqrt{1+2h_2^2P} \leq K\left(\mathbf{a}, \beta^*\right) \leq \frac{1+h_1^2P+h_2^2P}{\sqrt{1+2h_1^2P}}. \tag{18}$$

□

Figure 2 illustrates this result. It can be seen that the achievable rate pairs $(r_1(a;\beta_2); r_2(b|a;\beta_2))$ lies on the dominant area of the capacity region. Figure 3, shows the acceptable region in terms of h_1^2P and h_2^2P for $u = 1.2$ and $v = 8$.

5 Proposed Strategy with Outage

The proposed strategy in the previous section is throughput-optimal without outage, but in general tolerating outage leads to higher performance in term of expected throughput [4]. We can modify the constraints given by (14) and (15), in the case that a user has a high channel gain. Since based on the fading distribution (Rayleigh) the probability of having a large channel coefficient is small, the transmitter will take

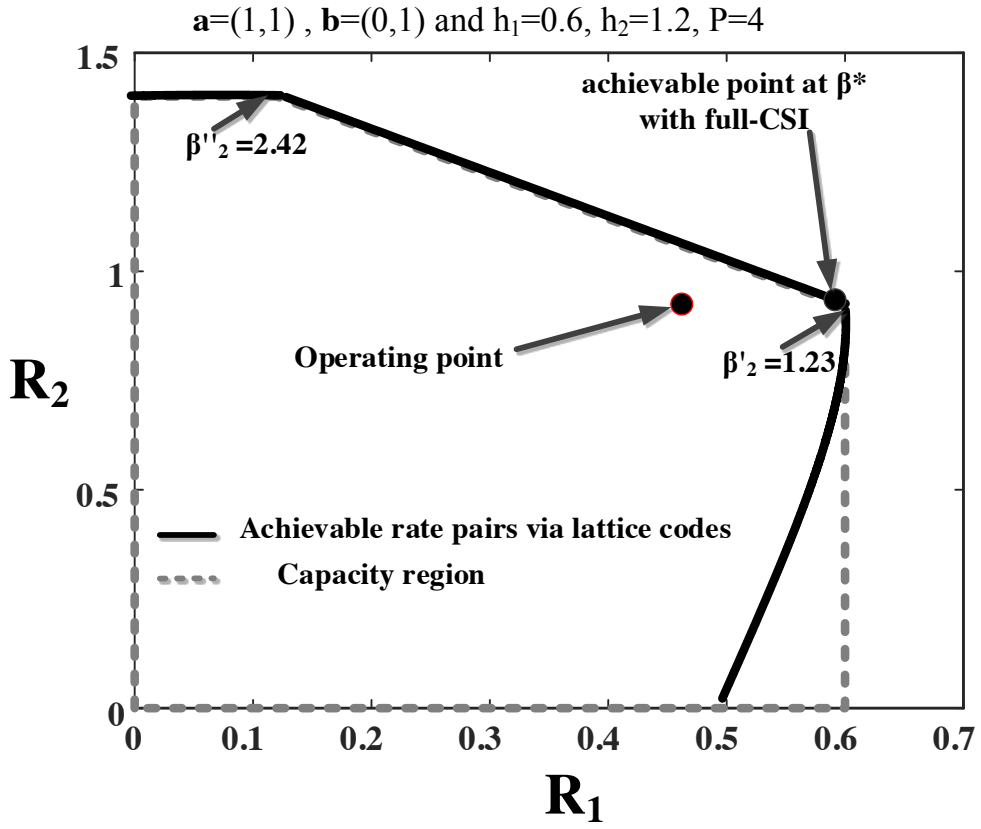


Figure 2: The pentagon dashed plot shows the capacity region of instantiated MAC and the solid plot shows the achievable rate pairs by choosing $\mathbf{a} = (\mathbf{1}, \mathbf{1})$ and $\mathbf{b} = (\mathbf{0}, \mathbf{1})$ by varying $\beta_2 \in \left[\beta_2', \beta_2''\right]$ [2]. User 1 and 2 send their messages at rates corresponding to (11). The plot shows that the operating rate point that is achievable with our scheme as well as the rate point that would be achievable with full CSI (16).

a risk, assume that the other user has a small channel gain, and sends at higher rate. In our modified strategy both users choose their rates as follows [4]

$$R\left(h\right) = \begin{cases} \frac{1}{4}\log\left(1+2h^2P\right), & \text{for } h \leq h_t, \\ \frac{1}{2}\log\left(1+h^2P+{h_t}^2P\right) - \frac{1}{4}\log\left(1+2{h_t}^2P\right), & \text{otherwise,} \end{cases} \tag{19}$$

by choosing $\beta_1 = 1$ and β_2 as

$$\beta_2 = \beta_t^*\left(h_2, h_t\right) = \frac{h_2 h_t P \pm \sqrt{{h_2}^2{h_t}^2P - \left(1+h_t^2P\right)\left(1+h_2^2P - \frac{1+{h_2}^2P+{h_t}^2P}{\sqrt{1+2{h_t}^2P}}\right)}}{1+h_t^2P}. \tag{20}$$

In this case we give more freedom for choosing β_2 as

$$\beta_2\left(h_2\right) = \begin{cases} \beta^*, & \text{for } h_2 \leq h_t, \\ \beta_t^*, & \text{otherwise .} \end{cases} \tag{21}$$

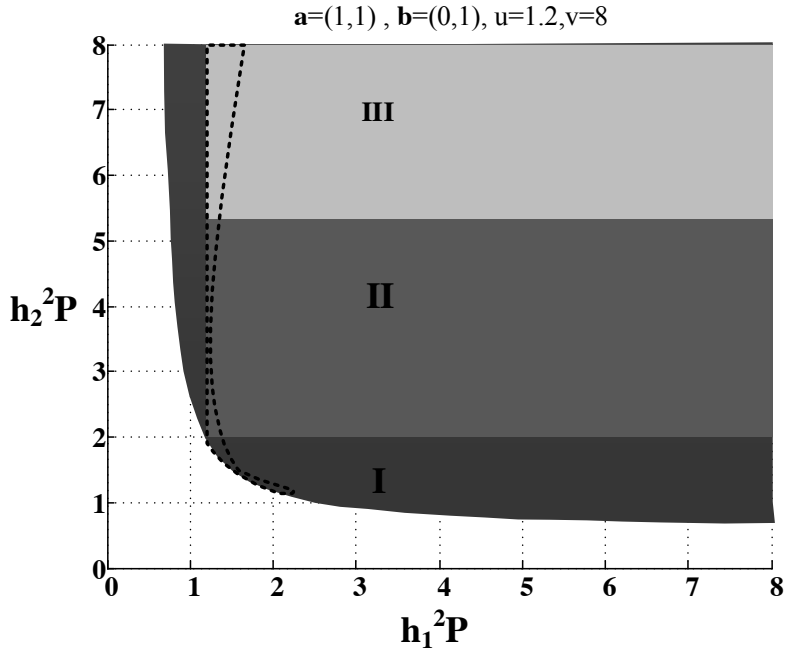


Figure 3: The valid region for the proposed strategy. dashed region is the valid region for (14), the region I corresponds to $A \geq \frac{3}{4}$. Region II corresponds to case that $\beta^* \leq \inf_{u \leq h_1 \leq v} \beta_{2''}(h_1, h_2)$ is satisfied and region III corresponds to $\beta^* \geq \sup_{u \leq h_1 \leq v} \beta_2'(h_1, h_2)$.

6 Conclusion

In this work, we have shown that lattice code can achieve the sum capacity of MAC under individual CSIT and identical fading statics across users when (14) and (15) are satisfied. We analyzed this strategy for two-user MAC, and in this strategy the sum capacity can be achieved with a single-user decoder without time sharing or rate splitting. For having larger throughput we consider the case that outage is permitted and find the appropriate choice of scaling coefficient. Figure 3 shows the achievability of our scheme for different values of received SNR. Where receiver can reliably decode both messages at the optimal rate (11).

References

[1] U. Erez and R. Zamir, "Achieving 1/2 log (1+SNR) on the AWGN channel with lattice encoding and decoding," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2293–2314, 2004.

[2] J. Zhu and M. Gastpar, "Gaussian (dirty) multiple access channels: A compute-and-forward perspective," in *2014 IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 2949–2953.

[3] C. S. Hwang, M. Malkin, A. E. Gamal, and J. M. Cioffi, "Multiple-access channels with distributed channel state information," in *2007 IEEE International Symposium on Information Theory (ISIT)*, June 2007, pp. 1561–1565.

[4] Y. Deshpande, S. R. B. Pillai, and B. K. Dey, "On the sum capacity of multiaccess block-fading channels with individual side information," in *2011 IEEE Information Theory Workshop (ITW)*, Oct 2011, pp. 588–592.

[5] S. Sreekumar, B. K. Dey, and S. R. B. Pillai, "Distributed rate adaptation and power control in fading multiple access channels," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5504–5524, 2015.

[6] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6463–6486, 2011.

[7] J. Goseling, M. Gastpar, and J. H. Weber, "Random access with physical-layer network coding," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 3670–3681, 2015.

[8] J. Goseling, C. Stefanovic, and P. Popovski, "Sign-compute-resolve for tree splitting random access," *Preprint, arXiv:1602.02612*, 2016.

Autoregressive Moving Average Graph Filter Design

Jiani Liu Elvin Isufi Geert Leus
Delft University of Technology
Faculty of EEMCS
2826 CD Delft, The Netherlands
j.liu-1@tudelft.nl e.isufi-1@tudelft.nl g.j.t.leus@tudelft.nl

Abstract

To accurately match a finite-impulse response (FIR) graph filter to a desired response, high filter orders are generally required leading to a high implementation cost. Autoregressive moving average (ARMA) graph filters can alleviate this problem but their design is more challenging. In this paper, we focus on ARMA graph filter design for a known graph. The fundamental aim of our ARMA design is to create a good match to the desired response but with less coefficients than a FIR filter. Our design methods are inspired by Prony's method but using proper modifications to fit the design to the graph context. Compared with FIR graph filters, our ARMA graph filters show better results for the same number of coefficients.

1 Introduction

Graph signal processing (GSP) extends classical digital signal processing to signals connected with the topology of a graph [1], [2]. More and more concepts and tools from classical signal processing are transferred to the field of GSP, including the uncertainty principle [3], graph wavelets [4], graph signal classification [5], graph signal recovery [6],[7], and graph signal sampling [8].

Depending on the definition of the graph frequency and the graph Fourier transform (GFT) [1], [2], many different kinds of graph filters have been designed as linear operators acting upon a graph signal. Similar to traditional digital signal processing, graph filters amplify or attenuate the graph signal at different graph frequencies. Graph filters have been used for many signal processing applications [9]. Traditionally, graph filters are expressed as a polynomial in the so-called graph shift matrix (e.g., the adjacency matrix, the graph Laplacian, or any of their modifications), resulting in a so-called finite-impulse response (FIR) graph filter [10], [12]. However, to accurately match some given filter specifications, FIR filters require high filter orders leading to a high implementation cost.

An alternative filter approach is the so-called infinite impulse response (IIR) graph filter [13]. Compared with FIR graph filters, characterized by a polynomial frequency response, IIR graph filters have a rational polynomial response, which brings more flexibility to their design. As such, IIR graph filters have the potential to fit complicated filter specifications with small degrees. As a particular instance of IIR graph filters, autoregressive moving average (ARMA) graph filters have been introduced as an extension of the potential kernel [14]. Such ARMA graph filters can be built as parallel and periodic concatenations of the potential kernel. ARMA graph filters were introduced as universal filters that do not depend on the particular topology of the graph, although in this paper we will design them for a particular graph in mind. Note that ARMA graph filters are capable of not only shaping the graph signal in the graph

frequency domain, but also in the regular temporal frequency domain, in case the graph signal is time varying [15]. And although our proposed designs can be used in that context, we will only focus on the graph domain in this paper.

Although ARMA graph filters show great promises, their design is challenging. This paper tries to tackle this issue for known graphs, which means that we only have to match the ARMA graph filter to the desired response in a set of known frequencies. For unknown graphs, we would have to match it over a continuous range of frequencies, which is much more complicated and is left for future work. We will present two design procedures, both inspired by Prony's method. The first design transforms the problem from the graph frequency domain to the coefficient domain (as done in the classical Prony's method). However, in contrast to the regular time domain, in the graph domain, this transformation is generally not optimal and does not lead to a good solution. Our second design shows that staying in the graph frequency domain preserves the optimality of the solution and also simplifies the problem. We conclude the paper by showing some simulation results for our two ARMA graph filter design procedures based on a known graph. They illustrate that in some cases, the FIR filter can be outperformed for the same number of filter coefficients.

2 FIR Graph Filter

Consider an undirected graph $G = (V, E)$, where V is a set of N nodes and E is the set of edges. We indicate with $\mathbf{M}$ the symmetric graph shift matrix, which could be the adjacency matrix, the graph Laplacian or any of their modifications. An eigenvalue decomposition of $\mathbf{M}$ leads to $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U}^T$ is the graph Fourier transform (GFT) matrix and $\mathbf{\Lambda}$ is a diagonal matrix with on the diagonal, the graph frequencies λ_n.

A graph filter $\mathbf{G}$ is a linear operator that acts upon a graph signal $\mathbf{x}$, leading to the output $\mathbf{y} = \mathbf{G}\mathbf{x}$. A finite impulse response (FIR) graph filter of order K, or FIR(K) in short, can be expressed as a K-th order polynomial in $\mathbf{M}$:

$$\mathbf{G} = g(\mathbf{M}) = \sum\nolimits_{k=0}^{K} g_k \mathbf{M}^k, \tag{1}$$

where g_k are the FIR filter coefficients. This means that $\mathbf{G}$ is diagonizable by the GFT matrix $\mathbf{U}^T$, i.e., $\mathbf{U}^T\mathbf{G}\mathbf{U} = g(\mathbf{\Lambda})$, and the filter indeed reshapes the spectrum of the input:

$$\hat{\mathbf{y}} = \mathbf{U}^T\mathbf{y} = \mathbf{U}^T\mathbf{G}\mathbf{U}\mathbf{U}^T\mathbf{x} = g(\mathbf{\Lambda})\hat{\mathbf{x}},$$

where $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ represent that GFT of the input and output signal, respectively. The frequency response of the filter at frequency λ_n, $n = 1, 2, \ldots, N$, is thus given by

$$\hat{g}_n = g(\lambda_n) = \sum_{k=0}^{K} g_k \lambda_n^k. \tag{2}$$

Define now the $N \times (K+1)$ Vandermonde matrix $\mathbf{\Psi}_{K+1}$ as the $N \times (K+1)$ matrix with entries $[\mathbf{\Psi}]_{n,k} = \lambda_n^{k-1}$. Then stacking the filter coefficients h_k in the $(K+1) \times 1$ vector $\mathbf{g}$ and the frequency response $\hat{g}_n$ in the $N \times 1$ vector $\hat{\mathbf{g}}$, we have

$$\hat{\mathbf{g}} = \mathbf{\Psi}_{K+1}\mathbf{g}. \tag{3}$$

From [11] we know that if the graph is known and if the graph frequencies are distinct, the filter coefficients can be computed as

$$\mathbf{g} = \mathbf{\Psi}_{K+1}^{\dagger} \hat{\mathbf{g}}, \tag{4}$$

where $\mathbf{A}^{\dagger}$ is the pseudo-inverse of the matrix $\mathbf{A}$.

It is well known (and clear from (3) and (4)) that for a known graph of N nodes with N distinct graph frequencies, a finite impulse response (FIR) graph filter of order $K = N - 1$ can exactly represent any desired response. In that case, the Vandermonde matrix $\mathbf{\Psi}_{K+1} = \mathbf{\Psi}_N$ is a square invertible matrix. However, for large (i.e., more than 100 nodes) graphs, which are regularly encountered in real applications (e.g., a temperature prediction system containing hundreds of cities), the computation of the FIR filter coefficients could be numerically infeasible because of the ill-conditioning of the related system matrix $\mathbf{\Psi}_N$. This issue is usually resolved by reducing the order of the FIR filter below the number of nodes of the graph ($K < N - 1$), at the cost of a reduced accuracy. But for large graphs, this still leads to large FIR filter orders which are costly to implement. In this work, we introduce another approach to resolve these problems of FIR filters.

3 ARMA Graph Filter

In order to achieve a better accuracy with a smaller order filter, we apply an autoregressive moving average (ARMA) filter to the signal living on the known graph G. Such ARMA graph filters can be obtained by running a dynamic diffusion process on the graph signal, as shown in [14]. Note that for an ARMA filter, it has been shown [14] that working with a translated version of the normalized graph Laplacian leads to the best stability conditions, but we will make abstraction of this here and simply work with a general shift matrix $\mathbf{M}$. From now on, we also assume that all graph frequencies λ_n are distinct.

For an ARMA(P,Q) graph filter, the graph frequency response at frequency λ_n, $n = 1, 2, \ldots, N$, can be written as

$$\hat{g}_n = g(\lambda_n) = \frac{\sum_{q=0}^{Q} b_q \lambda_n^q}{1 + \sum_{p=1}^{P} a_p \lambda_n^p}. \tag{5}$$

For future use, we also define $\mathbf{a} = [1, a_1, \ldots, a_p]^T$ and $\mathbf{b} = [b_0, b_1, \ldots, b_q]^T$ as the ARMA filter coefficients. Note that because the graph and thus the graph frequencies are known, we can always write $\hat{g}_n$ for all $n = 1, 2, \ldots, N$ as an $(N-1)$th order polynomial in λ_n with fixed coefficients:

$$\hat{g}_n = \sum_{k=0}^{N-1} g_k \lambda_n^k, \tag{6}$$

or in other words, the ARMA(P,Q) graph filter can always be written as an FIR($N-1$) graph filter if the graph is known. As before, the FIR filter coefficients g_k can be stacked in the $N \times 1$ vector $\mathbf{g}$ and the frequency response $\hat{g}_n$ in the $N \times 1$ vector $\hat{\mathbf{g}}$. The relation between $\hat{\mathbf{g}}$ and $\mathbf{g}$ is then given by $\hat{\mathbf{g}} = \mathbf{\Psi}_N \mathbf{g}$, where $\mathbf{\Psi}_N$ is defined as before but with K replaced by $N - 1$.

Now assume that we are given a prescribed frequency response $\hat{h}_n$, which can again be related to a set of N FIR filter coefficients h_k through

$$\hat{h}_n = h(\lambda_n) = \sum_{k=0}^{N-1} h_k \mu_n^k. \tag{7}$$

As before, we respectively stack h_k and $\hat{h}_n$ into $\mathbf{h}$ and $\hat{\mathbf{h}}$, which are related by $\hat{\mathbf{h}} = \mathbf{\Psi}_N \mathbf{h}$. The problem statement in this work is then to find the ARMA coefficients $\mathbf{a}$ and $\mathbf{b}$, resulting in a frequency response $\hat{\mathbf{g}}$ that best represents the desired frequency response $\hat{\mathbf{h}}$.

4 ARMA Filter Design

Optimally, we would try to minimize the error between $\hat{\mathbf{g}}$, which is parameterized by $\mathbf{a}$ and $\mathbf{b}$, and $\hat{\mathbf{h}}$, or in other words, we would try to solve

$$\min_{\mathbf{a},\mathbf{b}} \sum\nolimits_{n=1}^{N} \left| \hat{h}_n - \frac{\sum_{q=0}^{Q} b_q \mu_n^q}{1 + \sum_{p=1}^{P} a_p \mu_n^p} \right|^2 . \tag{8}$$

But since that is a difficult problem to solve, as in Prony's method, we choose to solve the related (yet not equivalent) problem

$$\min_{\mathbf{a},\mathbf{b}} \sum\nolimits_{n=1}^{N} \left| \hat{h}_n \left(1 + \sum\nolimits_{p=1}^{P} a_p \mu_n^p\right) - \sum\nolimits_{q=0}^{Q} b_q \mu_n^q \right|^2 . \tag{9}$$

Following a similar idea as in Prony's method, we use (7) to expand $\hat{h}_n$ in (9), resulting in

$$\min_{\mathbf{a},\mathbf{b}} \sum\nolimits_{n=1}^{N} \left| \left(\sum\nolimits_{k=0}^{N-1} h_k \mu_n^k\right) \left(1 + \sum\nolimits_{p=1}^{P} a_p \mu_n^p\right) - \sum\nolimits_{q=0}^{Q} b_q \mu_n^q \right|^2 .$$

Using some simple algebra, this can be transformed into the following least squares (LS) problem, writing in matrix form as

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{\Psi}_{N+P} \mathbf{H} \mathbf{a} - \mathbf{\Psi}_{Q+1} \mathbf{b}\|^2, \tag{10}$$

where $\mathbf{H}$ is the $(N+P) \times (P+1)$ Toeplitz matrix expressed by

$$\mathbf{H} = \begin{bmatrix} h_0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ h_{N-1} & \dots & h_0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & h_{N-1} \end{bmatrix} .$$

Note that the two terms in this LS problem are nothing more than the frequency response of the concatenation of the filter $\mathbf{h}$ and $\mathbf{a}$ on the left and the frequency response of the filter $\mathbf{b}$ on the right.

Defining $\hat{\mathbf{a}} = \mathbf{\Psi}_{P+1} \mathbf{a}$ and $\hat{\mathbf{b}} = \mathbf{\Psi}_{Q+1} \mathbf{b}$ as the frequency responses of the graph filters $\mathbf{a}$ and $\mathbf{b}$, we can thus also write (10) as

$$\min_{\mathbf{a},\mathbf{b}} \|\hat{\mathbf{h}} \circ \hat{\mathbf{a}} - \hat{\mathbf{b}}\|^2, \tag{11}$$

where $\circ$ represents the element-wise Hadamard product. Again, following Prony's approach, we now try to transform the problem (10) from the frequency domain to the

filter coefficient domain. This can be done by observing that the matrices $\mathbf{\Psi}_{N+P}$ and $\mathbf{\Psi}_{Q+1}$ can both be written as a function of the frequency transformation matrix $\mathbf{\Psi}_N$ as

$$\mathbf{\Psi}_{N+P} = \mathbf{\Psi}_N[\mathbf{I}_N, \mathbf{T}], \mathbf{\Psi}_{Q+1} = \mathbf{\Psi}_N \begin{bmatrix} \mathbf{I}_{Q+1} \\ \mathbf{0}_{(N-Q-1)\times(Q+1)} \end{bmatrix},$$

where $\mathbf{T}$ is some $N \times P$ transformation matrix that can be computed from the first equation since both $\mathbf{\Psi}_N$ and $\mathbf{\Psi}_{N+P}$ are known. So defining

$$\bar{\mathbf{H}} = [\mathbf{I}_N, \mathbf{T}]\mathbf{H}, \bar{\mathbf{b}} = \begin{bmatrix} \mathbf{I}_{Q+1} \\ \mathbf{0}_{(N-Q-1)\times(Q+1)} \end{bmatrix} \mathbf{b} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0}_{(N-Q-1)\times 1} \end{bmatrix},$$

we can rewrite (10) as

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{\Psi}_N \bar{\mathbf{H}}\mathbf{a} - \mathbf{\Psi}_N \bar{\mathbf{b}}\|^2, \tag{12}$$

Multiplying both terms with $\mathbf{\Psi}_N^{-1}$ (note that this corresponds to a loss of optimality), we finally transform the problem from the frequency domain to the filter coefficient domain:

$$\min_{\mathbf{a},\mathbf{b}} \|\bar{\mathbf{H}}\mathbf{a} - \bar{\mathbf{b}}\|^2. \tag{13}$$

Since $\bar{\mathbf{b}}$ has $N-Q-1$ zeros at the bottom, we can use the bottom $N-Q-1$ equations of (13) to solve for $\mathbf{a}$. The vector $\mathbf{b}$ can then be estimated using the top $Q+1$ equations of (13). Alternatively, we can use (12) (or (10)) after plugging in the estimate for $\mathbf{a}$.

In the time domain, this works well since $\mathbf{\Psi}_N$ then simply is the discrete Fourier transform (DFT) matrix, which is well-conditioned and even unitary up to a scale. In the general graph domain, however, $\mathbf{\Psi}_N$ can be very badly conditioned, especially for large graphs. So computing the desired filter coefficients h_k (and thus the matrix $\mathbf{H}$) as well as the transformation matrix $\mathbf{T}$, is numerically infeasible. Moreover, since $\mathbf{\Psi}_N$ is generally far from a scaled unitary matrix, the problems (12) and (13) are far from equivalent. Hence, we need to tackle this problem in a different way.

The basic idea is not to expand $\hat{h}_n$ in (9), and to rewrite it in matrix form as

$$\min_{\mathbf{a},\mathbf{b}} \|[\mathbf{\Psi}_{P+1} \circ (\hat{\mathbf{h}} \otimes \mathbf{1}_{1\times(P+1)})]\mathbf{a} - \mathbf{\Psi}_{Q+1}\mathbf{b}\|^2, \tag{14}$$

where $\otimes$ represents the Kronecker product. Then, instead of trying to move from the frequency domain to the filter coefficient domain, in order to exploit the finite order of $\mathbf{b}$, we simply project out this term using the orthogonal projection matrix

$$\mathbf{P}^{\perp}_{\mathbf{\Psi}_{Q+1}} = \mathbf{I}_N - \mathbf{\Psi}_{Q+1}\mathbf{\Psi}^{\dagger}_{Q+1}, \tag{15}$$

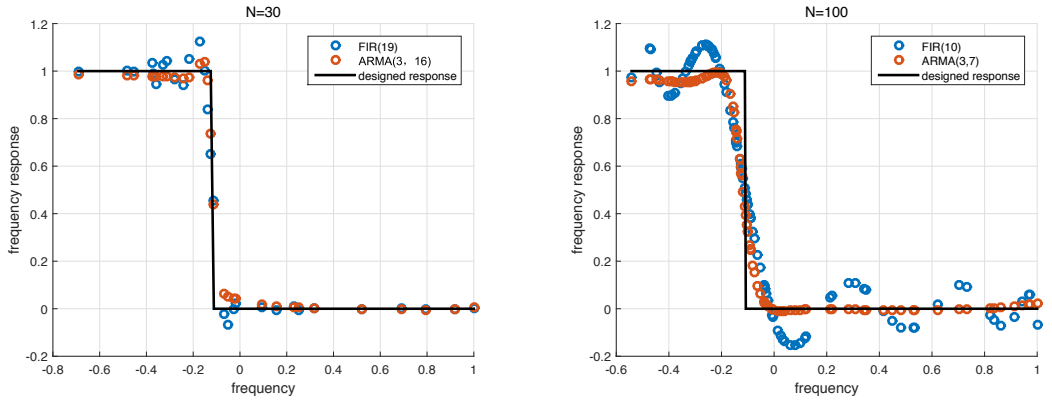
which is generally well-conditioned. Staying in the frequency domain not only preserves the optimality of the LS problem, but also simplifies the solution procedure. This allows us to transform (14) into the *equivalent* problem

$$\min_{\mathbf{a}} \|\mathbf{P}^{\perp}_{\mathbf{\Psi}_{Q+1}}[\mathbf{\Psi}_{P+1} \circ (\hat{\mathbf{h}} \otimes \mathbf{1}_{1\times(P+1)})]\mathbf{a}\|^2, \tag{16}$$

which can be used to solve for $\mathbf{a}$. The vector $\mathbf{b}$ can be estimated using (14) after plugging in the estimate for $\mathbf{a}$.

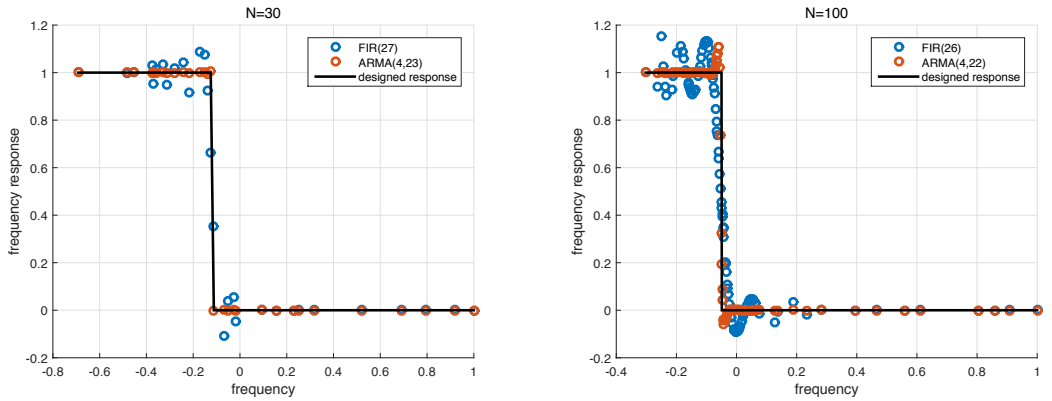
5 Numerical evaluation

The fundamental aim of our ARMA filter design is to create a good match to the desired response, but with less coefficients than an FIR filter with a similar match or to obtain a better match than an FIR filter with the same number of coefficients. For small graphs (N=20), the Vandermonde matrix is well-conditioned. Thence, an FIR filter with order $N-1$ can perfectly match any desired response depending on the inverse matrix of $\mathbf{\Psi}$. In this case, the FIR filter does better than the ARMA filter with a same filter order. The ARMA filter only works well when the order is higher than FIR filter, which brings more coefficients and computations. To test our ARMA filter design methods, we generate connected graphs by randomly placing 30 nodes and 100 nodes in a squared area, where two nodes are neighbors if and only if they are close enough to each other. We test both design methods. The first one is based on (13), whereas the second one is based on (16).



(a) 30 nodes (b) 100 nodes

Figure 1: Prony's method of ARMA filter with 30 nodes and 100 nodes graphs



(a) 30 nodes (b) 100 nodes

Figure 2: ARMA filter using projection matrix on graphs with 30 nodes and 100 nodes.

In Figure 1, we adopt the first method and show a comparison between the ARMA filter and FIR filter for the same number of coefficients. For a graph with 30 nodes, an FIR filter (28) is applied for the first step. We choose the order of FIR filter near the number of nodes which makes a good performance for the graph. The ARMA filter

design starts by fixing a specific order $P \ll N$ and then searching for the Q that leads to the best fit. Only the number of filter orders Q smaller than $K - P$ need to be investigated. The MSE between design response and ARMA filter response is used as evaluation index. Given an initial condition $P = 3$, the best performance comes from ARMA (3, 16). Comparing the performance of ARMA filter on graphs with different nodes, the result of large graph (100 nodes) depends on the selection of P and Q. With an initial condition $P = 3$, the ARMA filter shows obvious error even the result still better than the FIR filter with same order. As we mentioned before, the derivation of Prony's approach creates non-equivalent process for the filter coefficients. The LS solution for graph filter initially brings inevitable deviation to the course of solving the coefficients of ARMA filter.

In Figure 2, we focus on the second design method (i.e., (16)), which is appropriate for graphs of any size. Here we apply also our method to 30 nodes and 100 nodes graphs, which prevent us from using FIR filter orders of the same order of magnitude as the graph size. Same as the Prony's method, we first give the order $P \ll N$ and search for Q. Then we compare the resulting ARMA design with a FIR filter of order $P + Q = N$. Clearly, this ARMA filter design shows a better performance than the FIR filter design with the same number of coefficients.

The two methods only make sense here if Q is larger than P. For above two ARMA filters, all poles are in the unit circle and those filters are stable. However, the stability of ARMA filter depends on the initial P and the construction of graph. For same graph, when we change the order P, the best performance of ARMA filter may lose the stability. For different graphs, even we design with same order P, the stability is not a guarantee.

6 Conclusion

In this paper, we have considered the design of ARMA graph filters for a known graph. We have seen that Prony's method cannot be directly extended but needs some proper modifications in the graph context. We have seen that for large graphs and a relatively small number of coefficients, our ARMA filter design outperforms the FIR filter design for the same number of coefficients. We have not considered any stability issues in this paper related to the distributed implementation of our ARMA filter. This is left for future work.

References

[1] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 8398, 2013.

[2] A. Sandryhaila and J. M. Moura, "Discrete Signal Processing on Graphs," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1644-1656, 2013.

[3] Agaskar and Y. M. Lu, "A Spectral Graph Uncertainty Principle, " *IEEE Trans. Information Theory*, vol.59,no. 7,pp. 4338-4356, July 2013.

[4] D. K. Hammond, and P. Vandergheynst, "Wavelets on graphs via spectral graph theory", *Applied and Computational Harmonic Analysis* , 30 (2011) 129150.

[5] A. Sandryhaila, and J. M. F. Moura, "Classification via regularization on graphs", in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 495 - 498, Dec. 2013

[6] S. Chen, A. Sandryhaila, and J. M. F. Moura, "Signal recovery on graph: variation minimization ," *IEEE Transactions on Signal Processing*, vol. 63, no. 17, pp. 4609 - 4624, Jun. 2015.

[7] S. Chen, R. Varma, and A. Singh, "Signal recovery on graphs: Random versus experimentally designed sampling", in *IEEE, Sampling Theory and Applications (SampTA), 2015 International Conference*, pp. 337 - 341, May 2015

[8] A. Anis, A. Gadde, and A. Ortega, "Toward a sampling theory for signals on arbitrary graphs", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3864 - 3868, May 2014

[9] S. Chen, A. Sandryhaila, J. M. Moura, and J. Kovacevic, "Signal Denoising on Graphs via Graph Filtering", in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, GA, USA, December 2014, pp. 872-876.

[10] A. Sandryhaila, S. Kar, and J. M. F. Moura, "Finite-time distributed consensus through graph filters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 10801084.

[11] A Sandryhaila, JMF Moura, "Discrete signal processing on graphs: Graph filters", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6163 - 6166.

[12] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis", *IEEE Trans. Signal Process.*, vol. 62, no.12, pp. 30423054, 2014.

[13] X. Shi, H. Feng, M. Zhai, T. Yang, and B. Hu, "Infinite impulse response graph lters in wireless sensor networks", *Signal Processing Letters*, IEEE, 2015.

[14] A. Loukas, A. Simonetto, and G. Leus, "Distributed Autoregressive Moving Average Graph Filters", *IEEE Signal Processing Letters*, 2015, arXiv:1508.05808.

[15] E. Isufi, A. Loukas, A. Simonetto, and G. Leus, "Distributed Time-Varying Graph Filtering", arXiv preprint, arXiv:1602.04436v1, 2016.

[16] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, Georgia Institute of Technology, sec. 4.3, pp. 133-144, 1996.

Greedy Gossip Algorithm with Synchronous Communication for Wireless Sensor Networks

Jie Zhang, Richard C. Hendriks and Richard Heusdens
Signal and Information Processing Lab.,
Dept. of Microelectronics, Delft University of Technology,
2628 CD Delft, The Netherlands
{j.zhang-7, r.c.hendriks, r.heusdens}@tudelft.nl

Abstract

Randomized gossip (RG) based distributed averaging has been popular for wireless sensor networks (WSNs) in multiple areas. With RG, randomly two adjacent nodes are selected to communicate and exchange information iteratively until consensus is reached. One way to improve the convergence speed of RG is to use greedy gossip with eavesdropping (GGE). Instead of randomly selecting two nodes, GGE selects the two nodes based on the maximum difference between nodes in each iteration. To further increase the convergence speed in terms of transmissions, we present in this paper a synchronous version of the GGE algorithm, called greedy gossip with synchronous communication (GGwSC). The presented algorithm allows multiple node pairs to exchange their values synchronously. Because of the selection criterion of the maximum difference between the values at the nodes, there is at least one node pair with different information, such that the relative error must be reduced after each iteration. The convergence rate in terms of the number of transmissions is demonstrated to be improved compared to GGE. Experimental results validate that the proposed GGwSC is quite effective for the random geometric graph (RGG) as well as for several other special network topologies.

1 Introduction

Distributed signal processing in wireless sensor networks (WSNs) has many operational advantages. For instance, there is no need to have a fusion centre (or host) for facilitating computations, communication and time-synchronization. Positions of the network nodes are not necessarily known *a priori*, and the network topology might change as nodes join or disappear. For the design of fault-tolerant computation and information exchange algorithms over such WSNs, decentralized randomized gossip (RG) based averaging consensus is attractive, because it does not require any special routing, there is no bottleneck or single point of failure, and it is robust to unreliable and changing wireless network conditions. Moreover, the decentralized RG puts no constraints on the network topology and requires no information about the actual topology.

Since the original RG algorithm was proposed in [1], many derivatives were proposed to improve its convergence rate, and it has been employed into various applications (see e.g., [2] and references therein). Dimakies introduced a geographic gossip [3], which enables information exchange over multiple hops with the assumption that nodes have knowledge of their geographic locations, such that it is a good alternative for the grid network topology. In [4], a synchronous communication process was considered and improvements were made to the synchronous RG of [1] in a speech enhancement context. They allowed multiple node pairs to exchange their current values per iteration synchronously. Other improvements to increase the convergence speed are to use clique-based RG (CbRG) and cluster-based RG (see e.g., [5] and [6]), where cliques or clusters

are used to compress the original graph. Deniz *et al* presented a greedy gossip with eavesdropping (GGE) to accelerate the convergence [7]. Instead of randomly choosing two nodes, they chose the two nodes to communicate that have the maximum difference between values per iteration. Another more competitive broadcasting based algorithm was proposed in [8], although it cannot guarantee to reach the actual consensus surely.

To further increase the convergence speed in terms of transmissions, we present in this paper a synchronous version of the GGE algorithm, called greedy gossip with synchronous communication (GGwSC). Each time slot is divided into two time scales, one is the time used for node pairs selection, and the other is for the *gossip exchange* between every node pair. The simultaneous communicating node pairs are chosen recursively. Each time, one node selects the node from its neighbors that has the maximum difference. Then, the additional communicating node pairs are chosen recursively by excluding the node pairs that are already formed. Finally, the chosen node pairs communicate synchronously. Thus, unlike the synchronous gossip in [1] or [4], which performs updates completely at random, the GGwSC, like GGE, makes use of the greedy neighbor selection procedure. Whereas unlike GGE, we also permit multiple node pairs to communicate so as to accelerate the convergence rate. Experiments have demonstrated the effectiveness of the proposed method. The convergence rate in terms of the number of transmissions for random geographic graphs (RGGs) is accelerated compared to the GGE algorithm. Additionally, we also test the improvement on the convergence rate of the proposed method under different conditions in this paper, e.g., different initializations for the nodes and different network topologies.

2 Fundamentals of GGE

To guide the reader, we first give a brief overview of the GGE algorithm presented in [7]. We consider a network of N nodes and represent network connectivity as a graph, $G = (V, E)$, with vertices $V = \{1, 2, ..., N\}$ and edge set $E \subset V \times V$ such that $(i, j) \in E$ if and only if nodes i and j directly communicate. We assume that communication relationships are symmetric and that the graph is connected. Let $\mathcal{N}_i = \{j : (i, j) \in E\}$ denote the set of neighbors of node i (excluding i). Each node in the network has an initial value y_i, and the goal is to use only local information exchanges to arrive at a state where every node knows the average $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Each node is initialized with $x_i(0) = y_i$.

At the kth iteration of GGE [7], an activated node s_k is chosen uniformly at random. This can be accomplished using the asynchronous time model, where each node "ticks" according to a Poisson clock with rate 1. Then, s_k identifies a neighboring node t_k satisfying

$$t_k \in \arg\max_{t \in \mathcal{N}_j} \Big\{ \frac{1}{2}(x_{s_k}(k-1) - x_t(k-1))^2 \Big\}, \tag{1}$$

in other words, s_k identifies a neighbor that currently has the most different value from itself. This choice is possible because each node i maintains not only its own local variable, $x_i(k-1)$, but also a copy of the current values at its direct neighbors, $x_j(k-1)$, for $j \in \mathcal{N}_i$, because of eavesdropping with wireless communications. When s_k has multiple neighbors whose values are equally (and maximally) different from s_k's, it chooses one of these neighbors at random. Then the update is performed by enforcing the average $\frac{1}{2}(x_{s_k}(k-1) - x_{t_k}(k-1))$ to s_k and t_k, while all other nodes $i \notin \{s_k, t_k\}$ hold their values at $x_i(k) = x_i(k-1)$. Finally, the two nodes, s_k and t_k, broadcast these new values so that their neighbors have up-to-date information. If the values x_i on all sensors are stacked as a vector, i.e., $\boldsymbol{x}(k) = [x_1(k), x_2(k), ..., x_N(k)]^T$, we can formulate the above update as

$$\boldsymbol{x}(k) = \boldsymbol{U}_{GGE}(k)\boldsymbol{x}(k-1), \tag{2}$$

where $\boldsymbol{U}_{GGE}(k)$ is an $n \times n$ dimensional update matrix, which is dependent across time. For two communicating nodes x_{s_k} and x_{t_k} at iteration k, the update matrix is

$$\boldsymbol{U}_{GGE}(k) = \boldsymbol{I} - \frac{1}{2}(\boldsymbol{e}_{s_k} - \boldsymbol{e}_{t_k})(\boldsymbol{e}_{s_k} - \boldsymbol{e}_{t_k})^T, \tag{3}$$

where $\boldsymbol{e}_i = [0, ..., 1, 0, ..., 0]^T$ is an $N-$ dimensional vector with the ith entry equal to 1. Note that similar to the standard RG, the update matrix is doubly stochastic, which implies $\boldsymbol{U}_{GGE}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T\boldsymbol{U}_{GGE} = \mathbf{1}^T$ with $\mathbf{1}$ denoting a vector of all ones.

Given the initial vector of a network $\boldsymbol{x}(0) = [x_1(0), x_2(0), ..., x_N(0)]^T$, the theoretical consensus will be $\tilde{x}_{ave} = \mathbf{1}\tilde{\boldsymbol{x}}(\mathbf{0})/N$. To measure the convergence rate, we use the relative convergence error defined as

$$RE = \frac{\|\tilde{\boldsymbol{x}}(k) - \tilde{x}_{ave}\mathbf{1}\|}{\|\boldsymbol{x}(0) - \tilde{x}_{ave}\mathbf{1}\|}, \tag{4}$$

such that the iteration can be quitted when $RE \leq \varepsilon$ (or after a fixed amount of iterations).

3 GGwSC

In this section, we will present the proposed GGwSC algorithm based on GGE. As mentioned above, in GGE, a node selects a neighboring node whose state value is most different from its own value. This strategy can indeed accelerate the convergence at the cost of additional communication bandwidth compared to the original gossip algorithm [1], because it has to send (broadbast) the new values (eavesdrop) to all its neighbors. In spite of this, it still has a relatively slow convergence because only two nodes are allowed to exchange their state values at each iteration. In [4], a synchronous randomized gossip (SRG) was proposed for distributed delay and sum beamforming (DDSB) based speech enhancement in WSNs, where each node is permitted to communicate with one of its neighbors randomly at each iteration, such that the state values of multiple nodes are updated after each iteration. Given sufficient communication bandwidth, we combine the idea of GGE and SRG to further accelerate the convergence. Hence for the GGwSC, multiple node pairs can communicate at each iteration. These active node pairs are constrained to be disjoint, and the communicating node pairs are chosen according to **arg max** distance vectors.

This newly proposed GGWSC algorithm can generally be described as in **Algorithm 1**. For the practical realization, there are several points worthy to be noted:

- ◇ Given N (even) nodes, the desired case is that $N/2$ node pairs are chosen synchronously by the *SelectNodePair* function at each iteration. This would be most efficient. However, this will not always happen. For example, at kth iteration, when the node s_k is randomly activated, but all of its neighbors are selected already (i.e., $\mathcal{N}_{s_k} = \mathcal{O}$), s_k has a bye (i.e., $x_{s_k} = x_{s_{k-1}}$) and needs to wait for the next iteration $k + 1$.

- ◇ For the kth iteration, the update matrix $\boldsymbol{U}_{GGwSC}(k)$ is a manifold stochastic process approximately, that is, $\boldsymbol{U}_{GGwSC}(k) = \prod_{\{s_k,t_k\}\in V} \boldsymbol{U}_{GGE}^{s_k,t_k}(k)$.

- ◇ Note that for a communicating node pair, two transmissions are required during an iteration, e.g., s_k computes the average, such that s_k broadcasts it to its neighbors, and t_k also needs to broadcast the received average from s_k to its neighbors.

Algorithm 1: *GGwSC*

Input: $\boldsymbol{x}(0) = [x_1(0), x_2(0), ..., x_N(0)]^T$, $G = (V, E)$

while $RE > \varepsilon$ **do**
  function $SelectNodePair(G)$
    $s_k = N \times rand$;
    $t_k \in \arg\max_{t \in \mathcal{N}_{s_k}} \Big\{ \frac{1}{2}(x_{s_k}(k-1) - x_t(k-1))^2 \Big\}$;
    updating topology by excluding $\{s_k, t_k\}$ to $G' = (V', E')$;
    if $(V' \neq \mathcal{O})$ $SelectNodePair(G')$;
    else break;
  end function
  $\boldsymbol{U}_{GGE}^{s_k,t_k}(k) = \boldsymbol{I} - (e_{s_k} - e_{t_k})(e_{s_k} - e_{t_k})^T/2$;
  $\boldsymbol{U}_{GGwSC}(k) = \prod_{\{s_k,t_k\} \in V} \boldsymbol{U}_{GGE}^{s_k,t_k}(k)$;
  $\boldsymbol{x}(k) = \boldsymbol{U}_{GGwSC}(k)\boldsymbol{x}(k-1)$;
end
return $\boldsymbol{x}(k)$

3.1 Convergence Rate: GGwSC versus GGE

In the following, we investigate the convergence rate in terms of the underlying communication topology. The convergence rate for gossip algorithms [1] is typically defined in terms of the $\varepsilon-$avergaing time

$$T_{ave}(\varepsilon) = \sup_{\boldsymbol{x}(0) \neq \mathbf{0}} \inf \Big\{ k : \Pr\Big(\frac{\|\tilde{\boldsymbol{x}}(k) - \tilde{x}_{ave}\mathbf{1}\|}{\|\boldsymbol{x}(0) - \tilde{x}_{ave}\mathbf{1}\|} > \varepsilon \Big) \leq \varepsilon \Big\}. \tag{5}$$

The averaging time $T_{ave}(\varepsilon, Pr)$ is bounded by the second largest eigenvalue of the expected value of the update matrix $E[\boldsymbol{U}_{GGwSC}]$, that is [1]

$$\frac{0.5 \log \varepsilon^{-1}}{\log \lambda_2(E[\boldsymbol{U}_{GGwSC}])^{-1}} \leq T_{ave}(\varepsilon, Pr) \leq \frac{3 \log \varepsilon^{-1}}{\log \lambda_2(E[\boldsymbol{U}_{GGwSC}])^{-1}}. \tag{6}$$

Although this bound is suitable for the GGwSC as well, it is hard to relate it as a homogeneous Markov chain, and $T_{ave}(\varepsilon, Pr)$ is difficult to calculate as a function of $\lambda_2(E[\boldsymbol{U}_{GGwSC}])$, because $E[\boldsymbol{U}_{GGwSC}]$ depends on the network topology. Therefore, we use here an alternative bound to investigate the convergence rate, which is based on results from [7]. Given a graph $G = (V, E)$, we will have

$$E[\|\tilde{\boldsymbol{x}}(k) - \tilde{X}_{ave}\mathbf{1}\|^2] \leq A(G)^k \|\boldsymbol{x}(0) - \tilde{X}_{ave}\mathbf{1}\|^2, \tag{7}$$

where $A(G)$ is the graph-dependent constant defined as

$$A(G) = \max_{\boldsymbol{x} \neq \tilde{x}_{ave}\mathbf{1}} \frac{1}{N} \sum_{s=1}^{N} \Big(1 - \frac{\|\boldsymbol{g}_s(k)\|^2}{4\|\tilde{\boldsymbol{x}} - \tilde{x}_{ave}\mathbf{1}\|^2} \Big), \tag{8}$$

where $\boldsymbol{g}(k)$ is the subgradient function defined in [7]. Indeed, $A(G)$ is equivalent to $\lambda_2(E[\boldsymbol{U}_{GGWSC}])$ functionally. Obviously, the smaller of $A(G)$, the faster of the convergence rate. For the kth iteration of GGwSC, there is at least one node pair (s_k, t_k) communicating synchronously, such that $\boldsymbol{g}(k)$ has more than two elements unequal

to 0. Yet for the GGE algorithm, one node pair (s_k, t_k) is allowed to communicate per iteration, such that there are only two elements of the subgradient function unequal to 0. Therefore, we have the relationship between the subgradient functions, as $\|\boldsymbol{g}_{GGwSC}(k)\|^2 \geq \|\boldsymbol{g}_{GGE}(k)\|^2$, which leads to

$$A_{GGwSC}(G)^k \leq A_{GGE}(G)^k, \tag{9}$$

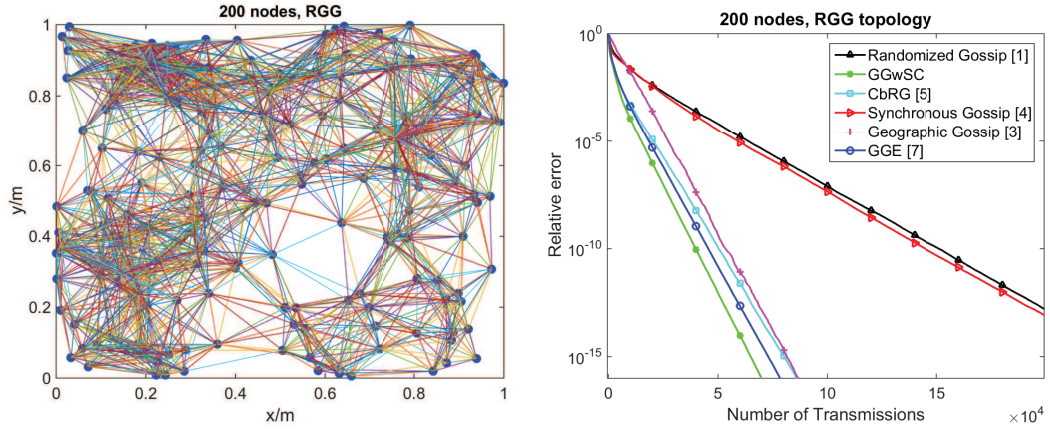
with equality if and only if only one node pair gossips per iteration. Consequently, we have demonstrated theoretically that GGwSC converges faster than GGE.

4 Performance Analysis

In this section, we present simulations to compare the GGwSC with several state-of-the-art methods, including Boyd's original RG [1], GGE [7], synchronous gossip [4], CbRG [5] and geographic gossip [3], by observing the convergence rate in terms of transmissions. We also investigate how this is effected by the network topology.

4.1 Random Geometric Graph (RGG)

Firstly, in order to observe the general performance of convergence, we place 200 nodes randomly in a (1×1) m enclosure. A Gaussian distribution $\mathcal{N}(0, 1)$, is used to initialize the values of $\boldsymbol{x}(0)$ on each sensor. The maximum number of transmissions is fixed to 20000, and the results are averaged over 100 realizations for the RGG. The transmission radius is set to be $\sqrt{\log N/N}$, which determines the RGG topology.



(a) RGG (b) Convergence

Figure 1: Convergence of relative error of the state-of-the-art methods for the RGG topology with 200 nodes.

Fig. 1(a) shows a typical RGG with 200 nodes, and Fig. 1(b) shows the corresponding convergence behaviours. We can see that our method achieves the fastest convergence rate, and randomized gossip and synchronous gossip are slowest.

4.2 Initialization

Secondly, we examine performance for four different initial conditions, $\boldsymbol{x}(0)$, which are consistent to those in [7], in order to explore the impact of the initial values on the

convergence behaviour. The first two of these cases are a Gaussian bumps field, and a linearly-varying field. For these two cases, the initial value $\boldsymbol{x}(0)$ is determined by sampling these fields at the locations of the nodes. The remaining two initializations consist of the "spike" signal, constructed by setting the value of one random node to 1 and all other node values to 0, and a random initialization where each value is i.i.d. drawn from a Gaussian distribution $\mathcal{N}(0,1)$ of zero mean and unit variance. The first three of these signals were also used to examine the performance of geographic gossip in [3].

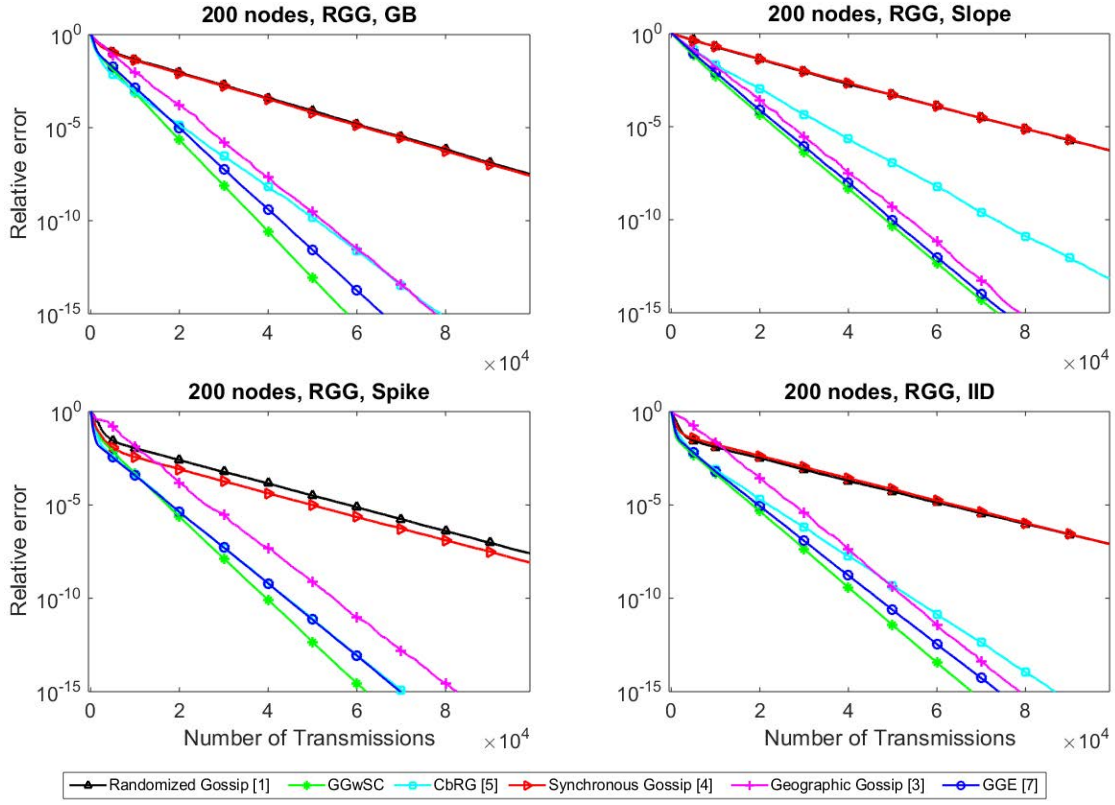


Figure 2: Comparison of the performance of the state-of-the-art methods with four different initializations of $\boldsymbol{x}(0)$.

Fig. 2 shows that GGwSC converges to the average at a faster rate asymptotically than the other state-of-the-art methods for all initial conditions. Out of these candidate initializations, the linearly-varying field is the worst case, because it improves the convergence rate least compared to GGE. This is not surprising since the convergence analysis in Section 3.1 suggests that constant differences between neighbors cause both GGwSC and GGE to provide minimal gain.

4.3 Special topologies

Finally, we investigate the influence of the network topologies on the convergence rates. We test three special kinds of topologies, including complete connected, grid, and a star topology. Note that for the grid network topology, the number of nodes must be a square. Some results are shown in Fig. 3 versus the number of transmissions. To this end, we can conclude:

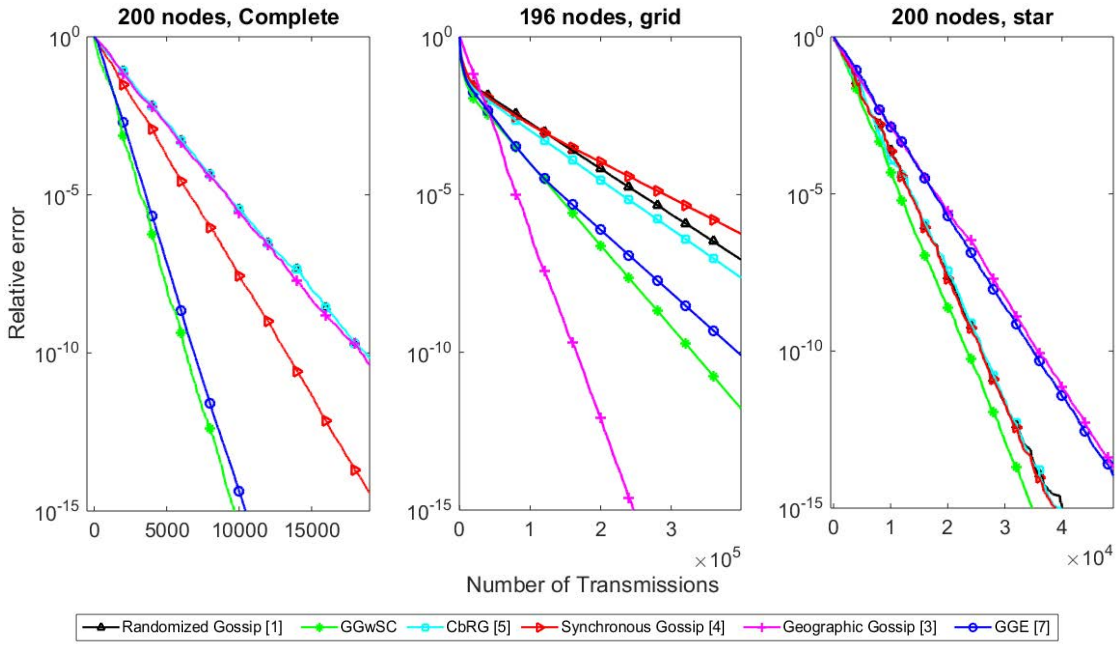


Figure 3: Comparisons of the performance of the state-of-the-art methods for three special network topologies (left: grid; middle: grid, where the number of nodes must be a square, e.g., 196; right: star).

- ◇ GGwSC is the most effective gossiping strategy, and it has the fastest convergence rate generally, except for the grid topology. For these grid-structured networks, geographic gossip has the best performance, because it is specified to thes kind of networks.
- ◇ Although both GGwSC and GGE perform gossiping according to the difference between neighboring nodes, through the synchronous communication strategy the former guarantees that at least one node pair has a value difference per iteration except in the case when the average is reached. That is why GGwSC is faster than GGE in terms of transmissions.

Accordingly, in general the proposed GGwSC algorithm obtains the fastest rate of convergence.

5 Conclusions

In this paper, we proposed a greedy gossip with synchronous communication (GGwSC) as an extension of the GGE algorithm [7] for averaging consensus. The convergence rate of GGwSC was analyzed theoretically as being faster than GGE. The experimental results demonstrated the effectiveness of the proposed method. Additionally, we also tested the performance on the convergence rate of our method under several conditions, e.g., different initializations for the nodes and different network topologies. In general, the proposed GGwSC algorithm obtained the fastest rate of convergence.

References

[1] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[2] A. G. Dimakis, S. Kar, J. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[3] A. G. Dimakis, A. D. Sarwate, and M. J. Wainwright, "Geographic gossip: efficient aggregation for sensor networks," in *ACM Int. Conf. Inform. Process. in Sensor Networks (IPSN)*, 2006, pp. 69–76.

[4] Y. Zeng and R. C. Hendriks, "Distributed delay and sum beamformer for speech enhancement via randomized gossip," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 260–273, 2014.

[5] Y. Zeng, R. C. Hendriks, and R. Heusdens, "Clique-based distributed beamforming for speech enhancement in wireless sensor networks," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, 2013, pp. 1–5.

[6] W. Li and H. Dai, "Cluster-based distributed consensus," *IEEE Transactions on Wire. Communicat.*, vol. 8, no. 1, pp. 28–31, 2009.

[7] D. Üstebay, B. N. Oreshkin, M. J. Coates, and M. G. Rabbat, "Greedy gossip with eavesdropping," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3765–3776, 2010.

[8] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione, "Broadcast gossip algorithms for consensus," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2748–2761, 2009.