Technical report for Collaborative Research Center SFB 876

Providing Information by Resource-Constrained Data Analysis

December 2017

Part of the work on this report has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis".

Speaker: Prof. Dr. Katharina Morik
Address: Technische Universität Dortmund
Fachbereich Informatik
Lehrstuhl für Künstliche Intelligenz, LS VIII
D-44221 Dortmund

Contents

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A1
Data Mining for Ubiquitous System Software

Katharina Morik Olaf Spinczyk

Machine Learning on Heterogeneous Hardware

Sebastian Buschjäger
Lehrstuhl für Künstliche Intelligenz, LS 8
Technische Universität Dortmund
sebastian.buschjaeger@tu-dortmund.de

With increasing volumes in data and more sophisticated machine learning algorithms, the demand for fast and energy efficient computation systems is also growing. To meet this demand, two approaches are possible: First, machine learning algorithms can be tailored specifically for the hardware at hand. Second, instead of changing the algorithm we can change the hardware to suit the machine learning algorithms better. This report briefly discusses my last years' work which focused largely on the first approach and quickly outlines some ideas for future research.

1 Introduction

To make machine learning universally applicable, we need to bring its algorithms to small and embedded devices including both - the training and the application of models. From a computer architectural point of view, we may optimize these two aspects separately. In model application we rapidly apply an already trained model for predictions and thus focus on the optimization of this inference. In model training however, we would like to train models on small devices directly, so that these devices dynamically adjust their prediction rules for new data.

2 Machine learning for embedded devices

For model application I looked at the filtering of sensor data by using Random Forests for data driven filtering rules. The Random Forest is among the most widely used machine learning algorithms. It consists of M decision trees, which are independently trained by using bootstrap aggregation [2]. In bootstrap aggregation, each tree is trained on a different sample of the data, so that samples are expected to overlap to some extent. This overlapping enables compelling statistical properties such as variance reduction and stability. For predictions, each tree is traversed until a leaf-node is found. Then, individual decisions are combined with a majority vote.
Once a machine learning expert found a specific Random Forest model for the data at hand which performs well, we would like to put this model into practice so that it continuously performs predictions.
Random Forests' model application is difficult from a computer architecture point of view, since the structure of decision trees breaks caching and pipelining mechanism of modern CPUs. As detailed in [3], we can implement decisions trees in four ways utilizing data cache, instruction cache and vectorization units differently. Depending on the specific structure of the tree, these implementations yield in different run-times. For example, there might be certain paths in the tree which will be used the majority of the time during inference and thus these paths should be kept inside the cache.
I formalized this by viewing each decision node as Bernoulli experiment, where the probability to follow the left child is p_{il} and the probability to follow the right child is $p_{ir} = 1 - p_{il}$ for node i. Note, that p_{il} is usually computed as a bi-product during training. This way, all decisions performed while traversing the tree can be modeled as a chain of Bernoulli experiments with probability p_{il} and p_{ir}. This statistical view of model application in turn allows to estimate the expected number of decisions each tree in the forest will need to perform during inference. Using this knowledge we then can compute the performance of different implementation schemes before implementing a tree based on its structure.

3 Machine learning on embedded devices

Going from model application to model training, I stayed in the realm of embedded systems and sensor data. Here we notice, that embedded devices are often tightly connected with other systems measuring a multitude of different sensor values. For humans, it can be difficult to make sense from this stream of measurements. To tackle this challenge, we may compute data summaries on-the-fly - that is, we extract small but highly informative samples from the original data stream while the data is generated. These summaries can then be reviewed by human experts and may offer new insights about the data. Data summaries should adhere theoretical guarantees to make sure that all the informative events are included in any case. The Sieve-Streaming algorithm [1] precisely offers such

guarantees by maximizing sub-modular functions on a data stream. In supervision of a bachelor thesis [7], we were able to reduce the computation cost of Sieve-Streaming by carefully deriving new bounds. These new bounds ultimately lead to significant speed-ups on embedded systems as shown in [4].

Based on these data summaries I was also able to derive a product of expert ensemble model for Gaussian processes in the context of regression. From a theoretical perspective, this method links an optimization-based view of Gaussian processes to the probabilistic view of product of expert models [5]. From a practical point of view, the method offers a way to compute small, local Gaussian Processes expert models, which can be executed on small devices. It scales well with large amounts of data and offers competitive performance. The corresponding paper is under review at the moment.

4 FPGAs in machine learning

FPGAs have a long history in machine learning, as well as in embedded systems. However, the combination of all three seems to be a recent idea. Deep Learning methods greatly benefit from the computation offered by FPGAs and are currently in focus of research [6].
In the last year, I supervised the Fachprojekt about Deep Learning on FPGAs and gained theoretical knowledge about Deep Learning methods. These theoretical insights ultimately lead to two talks on the SFB's summer school. Additionally, I currently supervise multiple bachelor thesis in the context of Deep Learning on smaller devices which aim to bring this theoretical knowledge closer to the application of Deep Learning on FPGAs and small devices.

5 Future research

In the future I would like to focus my work on ensemble methods such as Boosting or Bagging. Practice shows, that ensembles of models consistently outperform single models. With [3] I started to work on ensembles from a computer architectural point of view. In conjunction with the embedded systems group (LS 12) we currently extend this approach to include the memory layout of single trees with respect to caching.
From a theoretical point of view, ensemble methods already offer a solid theoretical groundwork, which makes the behavior of these methods transparent. I would like to use this transparency fine-tune these methods for embedded systems and FPGAs while still preserving their theoretical guarantees.

As a second research focus, I would like to apply the aforementioned data summarization techniques on different data. In [7] we were able to use Sieve-Streaming to find (artificial) anomalies in the FACT data from the C3 project. In practice, Sieve-Streaming is based on kernel similarities of the input data, which requires hand-tuning and a lot of domain specifics. With the gained knowledge in Deep Learning methods, I would like to train autoencoder networks on the FACT data. Autoencoder generate data embeddings, which can then be fed into classical kernel functions to measure the similarity between two data points. Additionally, with the runtime reduction of Sieve-Streaming it is possible to implement it on FPGAs bringing it closer to the measuring device.
Last, the expertise gained in the context of Deep Learning methods should be used to bring Deep Learning and FPGAs closer together. More specifically, I would like to bring Deep Learning inference to FPGAs similar to the framework developed in the B2 project for GPUs.

References

[1] Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Streaming submodular maximization: Massive data summarization on the fly. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 671–680. ACM, 2014.

[2] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[3] S. Buschjäger and K. Morik. Decision tree and random forest implementations for fast filtering of sensor data. *IEEE Transactions on Circuits and Systems I: Regular Papers*, PP(99):1–14, 2017.

[4] S. Buschjäger, K. Morik, and M. Schmidt. Summary extraction on data streams in embedded systems. In *ECML Conference Workshop IoT Large Scale Learning from Data Streams (to appear)*, 2017.

[5] Marc Peter Deisenroth and Jun Wei Ng. Distributed gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[6] Griffin Lacey, Graham W Taylor, and Shawki Areibi. Deep learning on fpgas: Past, present, and future. *arXiv preprint arXiv:1602.04283*, 2016.

[7] Maik Schmidt. Datenzusammenfassungen auf Datenströmen. Bachelor's thesis, TU Dortmund University, 7 2017.

LockDoc: Trace-Based Analysis of Locking in the Linux Kernel

Alexander Lochmann
Arbeitsgruppe Eingebettete Systemsoftware
Technische Universität Dortmund
alexander.lochmann@tu-dortmund.de

This report gives a brief overview of our research on lock analysis in the Linux kernel. It explains our approach on how we record and analyse execution traces. Furthermore, it describes how we check existing and documented locking rules, and how we derive new locking rules for undocumented data structures. Finally, we show a subset of our results.

1 Introduction

Every modern operating system has to deal with different kinds of synchronization. In particular, this often involves locking a critical section. Doing so, an operating systems must provide proper locking techniques as well as a locking rule documentation. Otherwise, a new kernel developer would not know which lock to acuire when accessing a certain data structure. As can be seen in the Linux kernel, the number of source code statements that access locks has increased during the last 6.5 years by 49% for spinlocks and by 78% for mutexes [5]. Eventhough this inscrease indicates a high relevance of correct locking, the related documentation is minimal. For example, the documentation of locking rules in the filesystem subsystem consists of a few comments scattered throughout the code using inconsistent wording, e.g., "holds" vs. "is held", or "inode lock held" vs. "i_lock". Some comments are provided down in central header files or important C files, e.g., include/linux/fs.h or fs/inode.c, but most of them are distributed across the remaining files. So it is up to a new kernel developer to dig himself through all the filesystem-related code to find the proper lock for accessing struct inode, for example. We have therefore developed a new trace-based approach for lock analysis to overcome

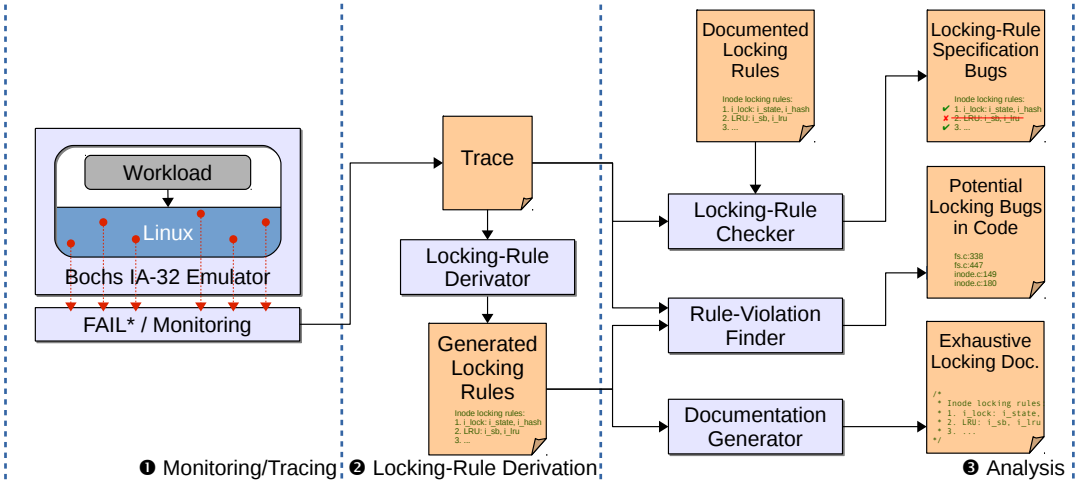


Figure 1: LockDoc overview: Based on a memory-access and lock acquisition trace from the Linux kernel, we infer the most probable locking rules for a specific set of data structures). Using this information, we look for locking-rule documentation bugs, locking bugs in the code, and generate exhaustive locking documentation. (taken from [5])

this sorry state. This work is in the context of lock analysis [1, 2]. Section 2 briefly describes this approach, while Section 3 shows a subset of our results.

2 LockDoc Approach

The *LockDoc* approach consists of three phases as shown in Figure 1. The first phase involves the recording of the traces, whereas, the actual analysis is performed in phase two and three.

Tracing To generate a trace, we run a certain workload inside a Linux virtual machine, and log heap allocations, operations that access the allocated memory as well as lock acquisitions and releases. Since there is no such thing as the one and only filesystem benchmark, we generate the traces using our own custom fs benchmark. It is composed of several micro benchmarks taken from the Linux Test Project [3], and some own tests like using pipes, dealing with symbolic links, and changing permissions. The guest system in the virtual machine logs every allocation, deallocation, acquisition and release via a virtual serial port. We use the Fail* fault-injection framework [6] in conjunction with the Bochs x86 emulator [4] to run the virtual machine and intercept the serial communication, and write it to a file. Using Fail*, we furthermore listen for memory accesses to heapallocated objects which have been logged as aforementioned.

Locking-Rule Derivation The purpose of the locking-rule derivator is to generate rules for every data-structure member which is observed during the benchmark run. Based on the asumption that locking in the Linux kernel is correct, the locking rule for a certain data-structure member must be in the execution trace. Hence, there must be a larger number of memory accesses that adhere to that rule. First, our heuristic extracts all possible locking rules from the dataset. Afterwards, it calculates the absolute support s_a, which corresponds to the number of observations, and the relative support s_r, which is the percentage of all observations, for each rule. To determine a winning rule, the relative support must be above the acceptance threshold t_{ac}. If, however, the derivator sees more than t_{nl} percentage of accessess without any lock held, it assumes that no lock is required at all.

Analysis The locking-rule checker uses the official and central documented locking rules and determines the relative and absolute support for each rule. It then identifies the amount of *correct* ($s_r = 1$), *ambivalent* ($0 < s_r < 1$), and *incorrect* ($s_r = 0$) locking rules. Based on the generated locking rules, the rule-violation finder determines potential locking bugs. In addition to the code location, it provides a list of locks held as well as the stacktrace. Finally, the documentation generator uses the generated locking rules to create a proper documentation for all data types and member which have been observed.

3 Results

We ran our experiments using a vanilla 4.10 Linux kernel which has been intstrumented to log allocations and deallocations of the following data types: backing_dev_info, block_device, cdev, inode, journalt_t, pipe_inode_info, super_block, and transaction_t. Moreover, we log calls to locking functions of different locks (*spinlock_t*, *rw_lock_t*, *semaphore*, *rw_semaphore*, *mutex* and *rcu*). Table 1 shows the results of the locking

Data Type	#R	#No	#Ob	✓(%)	~ (%)	✗
inode	14	3	11	27.27	36.36	36.36
transaction_t	42	12	30	70.00	20.00	10.00
journal_t	38	8	30	63.33	26.67	10.00

Table 1: Summary of validated locking rules: Each row shows how many locking rules are documented (#R), and how many of the corresponding members have not been observed (#No) and observed (#Ob). The last three columns denote the portion of correct ($s_r = 1$), ambivalent ($0 < s_r < 1$) and incorrect ($s_r = 0$) rules (cf. Section 2). (taken from [5])

rule checking. The Linux documentation contains 94 locking rules in total. We have seen 71 of them in our dataset (column "#Ob"). For struct transaction_t, for example, 70 % of the observed locking rules are *correct* ($s_r = 100$). Whereas, 10 % of them are *incorrect*. The reason for that is subject to future investigation.

References

[1] Peter T. Breuer and Simon Pickin. Checking for deadlock, double-free and other abuses in the Linux kernel source code. In Vassil N. Alexandrov, Geert Dick van Albada, Peter M. A. Sloot, and Jack Dongarra, editors, *International Conference on Computational Science (ICCS 2006)*, pages 765–772. Springer, May 2006.

[2] Peter T. Breuer and Marisol García Valls. Static deadlock detection in the Linux kernel. In Albert Llamosí and Alfred Strohmeier, editors, *Reliable Software Technologies – Ada-Europe 2004*, pages 52–64. Springer, 2004.

[3] Cyril Hrubis et al. Linux test project. `https://github.com/linux-test-project/ltp`. Accessed: 2018-01-24.

[4] Kevin P. Lawton. Bochs: A portable PC emulator for Unix/X. *Linux Journal*, 1996(29):7, September 1996.

[5] Alexander Lochmann, Horst Schirmeier, Hendrik Borghorst, and Olaf Spinczyk. LockDoc: Trace-Based Analysis of Locking in the Linux Kernel. In *2018 USENIX Annual Technical Conference (USENIX ATC 18)*, submitted 2018.

[6] Horst Schirmeier, Martin Hoffmann, Christian Dietrich, Michael Lenz, Daniel Lohmann, and Olaf Spinczyk. FAIL*: An open and versatile fault-injection framework for the assessment of software-implemented hardware fault tolerance. In *Proceedings of the 11th European Dependable Computing Conference (EDCC '15)*, pages 245–255, Piscataway, NJ, USA, September 2015. IEEE Press.

Probabilistic Models on the MSP430-FR5x

Nico Piatkowski
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
nico.piatkowski@tu-dortmund.de

Theoretical and empirical results show that our techniques can reduce the memory consumption, arithmetic requirements, and computational complexity of ordinary exponential family models. The conditional independence structure is kept intact, no new assumptions are introduced, and our proposed methods arise from the very definition of the exponential family. Reparametrization, integer models, and the stochastic quadrature are inherently connected via regularization, which allows us to transfer resource constraints to constraints or costs on the model's parameter space.

Here, we discuss our findings in the context of the Texas Instruments MSP430-FR5x microcontroller unit. The system contains a 16 bit CPU with reduced instruction set at 16 MHz clock rate. At runtime, it consumes 118 $^{\mu A}/_{MHz}$ (with 500 nA standby). It provides 256 kB of non-volatile ferroelectric random-access memory and 8 kB static random-access memory. We investigate if our proposed methods increase the size of models that can be learned on the MSP430-FR5x MCU.

Our sparse reparametrization is designed to allow more spatio-temporal models to fit into the main memory of a resource-constrained system. Regularized models enjoy a superior sparsity, compared to ordinary maximum likelihood estimates. If the regularization is too strong, the model quality begins to decrease. Our proposed reparametrization approach damps this effect by transferring information from non-zero parameters to parameters which are 0. Thus, parameters may be zero without sacrificing the model's quality. Exemplary results are summarized in Table 1. The table contains the average sizes in kilobyte of estimated parameter vectors, supposing a 32 bit floating-point representation, for models without regularization, with l_1-regularization, and with l_1-regularized reparametrization. Models on synthetic data use the inverse exponential decay and models on real-world

Table 1: Memory consumption (in kilobyte) of $\boldsymbol{\theta}$ (32 bit floating-point) without regularization, with l_1-regularization, and with l_1-regularized reparametrization (inverse exponential decay for synthetic data and rational decay for real-world data). Results shown are for $\lambda = 0.64$ and $T = 32$, averaged over all redundancy levels. Bold models would fit into the main memory of the MSP430-FR5x MCU, assuming 50 kB program code.

Data	d	**None**	l_1-**Reg.**	**Reparam.**
Chain	1066.68	4266.72	247.52	**202.08**
Star	1084.0	4336.0	**201.6**	**197.44**
Grid	1037.8	4151.2	277.92	**199.04**
Full	843.8	3375.2	257.92	**181.6**
Insight	2662025.0	10648100.0	5940.0	7130.0
Intel	62150.0	248600.0	**142.0**	430.0
Vavel	5610650.0	22442600.0	90196.0	30508.0

data the rational decay. Bold entries indicate that the parameter vector would fit into the main memory of our ultra-low-power architecture, where we assume that the inference algorithm, the optimization algorithm and additional variables consume ≈ 50 kB. First of all, we see that our regularization approach doubles the number of models which fit onto the MCU. Secondly, we know that the plain l_1-regularized models sacrifice the conditional independence structure and hence exhibit a substantially higher mean squared error than the reparametrized models. Thus, although two l_1-models would in principle fit into the systems memory, they are unusable in practice due to their low quality.

The second major obstacle when we try to learn undirected models on the MSP430-FR5x is the missing floating-point hardware. It is indeed possible to emulate 32 bit and 64 bit floating-point arithmetic in software, but this is accompanied by a very high computational overhead. Our integer undirected model solves this issue. We provide inference and optimization algorithms which work only on the integer domain—they do not require any floating-point arithmetic. Moreover, our theoretical derivations are valid for any subset $\{1, 2, \ldots, k-1\}$ of the non-negative integers, which implies that the results are also valid for small word-sizes. Experiments show that small values of k suffice to provide a practical model quality. Hence, we may choose the native 16 bit integers as underlying data type for our integer undirected model. Corresponding runtime results are summarized in Tables 2 and 3. The values in columns three and four represent the average runtime in milliseconds of one iteration of loopy belief propagation and one iteration of bit-length propagation executed on the MSP430-FR5x MCU. Table 2 contains LBP results for emulated 64 bit floating-point arithmetic, and Table 3 contains the corresponding

Table 2: Runtime in milliseconds of one iteration of message passing for LBP (64 bit floating-point), generation of one SQM sample (polynomial degree 2, 64 bit floating-point), and one iteration of BLprop (16 bit integer), on an MSP430-FR5x microcontroller unit.

Data	$\|E\|$	**LBP** (1 iter)	**BLprop** (1 iter)	**SQM** (1 sample)
Chain	15	4843.4	19.0	773.8
Star	15	4744.3	19.0	774.0
Grid	24	7713.9	29.5	1081.3
Full	120	40422.6	141.2	4704.2

Table 3: Runtime in milliseconds of one iteration of message passing for LBP (32 bit floating-point), generation of one SQM sample (polynomial degree 2, 32 bit floating-point), and one iteration of BLprop (16 bit integer), on an MSP430-FR5x microcontroller unit.

Data	$\|E\|$	**LBP** (1 iter)	**BLprop** (1 iter)	**SQM** (1 sample)
Chain	15	1156.2	19.0	350.3
Star	15	1140.4	19.0	393.1
Grid	24	1838.1	29.5	445.3
Full	120	9642.1	141.2	1549.7

results for emulated 32 bit arithmetic[1]. BLprop was executed with the MCU's native word-size of 16 bit. Both sets of results are unambiguous: compared to 64 bit emulation, our integer models are at least 250 times faster. In case of 32 bit emulation, the speedup is still 60. Having our results from [1] in mind, we see that the benefit of our integer models is higher the weaker the underlying computational architecture is. Moreover, our experiments show that good integer parameters can be found, even when the parameters of the data generating process are far from integrality. Thus, one may indeed expect to observe the same behavior in practice. The speedup goes hand in hand with a reduced energy consumption. When we assume that our microcontroller is battery powered, using integer models can hence significantly extend the uptime of our system.

Reparametrization and integer models may be combined, e.g., by choosing binary decay matrices. However, both rely on approximate message passing algorithms to perform

[1] 16 bit floating-point emulation is not supported by the MCU.

probabilistic inference. Such algorithms have several limitations like unknown convergence behavior on general cyclic graphs, and an unbounded approximation error. In certain settings, these properties may be undesirable and one has to resort to exact algorithms or approximation algorithms with error bounds. Those algorithms, like the junction tree algorithm or WISH, have an exponential complexity and are hence not well-suited for small systems. Our stochastic quadrature method combines polynomial approximation with Monte Carlo sampling for fast approximate inference. Moreover, the approximation error is bounded and depends on the norm of the parameter vector $\boldsymbol{\theta}$. Our results from [2] show, that small polynomial degrees and moderate sample numbers are often enough to achieve practical error rates. Exemplary results on the MSP430-FR5x can be found in the fifth column of Tables 2 and 3. We measured the average runtime to generate a single SQM sample for a polynomial degree of 2. The runtime is in between LBP and BLprop, which, in contrast to SQM, do not provide any guarantees on general loopy graphs. In practical applications, several samples have to be drawn. When we assume a sample size of 100, the accumulated runtime would be within $1/2$ and 8 minutes. This could indeed be prohibitive for some applications. Nevertheless, Theorem 7 from [2] allows us to find an upper bound on the error that depends on the number of samples. For any fixed model, we may thus adjust the sample size to find a reasonable tradeoff between resource consumption and quality. In contrast, the error of message passing on loopy graphs may oscillate, and a functional relation between the LBP error and the number of iterations is unknown. Moreover, SQM admits that several computational steps may be precomputed on a strong system. In the above example, $\boldsymbol{\theta}$ was known and we precomputed τ, the $\|\chi_\phi^i\|_1$ values, and the polynomial coefficients $\boldsymbol{\zeta}$. These values are valid for all parameters whose norm does not exceed $B = \|\boldsymbol{\theta}\|$. It is thus possible to reuse the same precomputed values during training, if the optimization procedure guarantees that the parameter norm does not exceed B, e.g., via regularization.

To sum up, our techniques reduced the memory consumption, accelerated inference and learning, and offer several ways to trade quality against resource consumption. Hence, a wider range of probabilistic models can be learned and applied on resource-constrained systems. This was the first investigation of exponential family models on ULP devices.

References

[1] Nico Piatkowski, Sangkyun Lee, and Katharina Morik. Integer undirected graphical models for resource-constrained systems. *Neurocomputing*, 173, Part 1:9–23, 2016.

[2] Nico Piatkowski and Katharina Morik. Stochastic discrete clenshaw-curtis quadrature. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 3000–3009. JMLR.org, 2016.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A2

Algorithmic aspect of learning methods in embedded systems

Christian Sohler Jens Teubner

Probabilistic Embeddings of the Fréchet Distance

Amer Krivošija
Lehrstuhl für Effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
amer.krivosija@tu-dortmund.de

The Fréchet distance is a popular distance measure for curves, but the computation complexity to determine the similarity between two given curves poses considerable computational challenges in practice. To address this problem we study distortion of the probabilistic embedding that results from projecting the curves to a randomly chosen line. We show that in the worst case and under reasonable assumptions, the discrete Fréchet between two polygonal curves in $\mathbb{R}^2$ or $\mathbb{R}^3$ of complexity n degrades by a factor linear in n with constant probability and show matching upper and lower bounds on the distortion (up to a constant factor).

Introduction

The Fréchet distance is a popular distance measure for curves which naturally lends itself to fundamental computational tasks, such as clustering, nearest-neighbor searching, and spherical range searching in the corresponding metric space. But the complexity of computation of the similarity between two given curves poses considerable computational challenges in practice, both for the discrete and the continuous Fréchet distance. To address this problem we study distortion of the probabilistic embedding that results from projecting the curves to a randomly chosen line.[1] Such an embedding could be used in combination with, e.g. locality-sensitive hashing [5], or (in case of the continuous distance) with the results on clustering in the one-dimensional space [4]. We show that in the worst case and under reasonable assumption that the curves are c-packed (for some

[1] Driemel, Krivošija, *Probabilistic Embeddings of the Fréchet Distance*, to appear, 2018

$c \geq 2$), the discrete Fréchet between two polygonal curves in $\mathbb{R}^2$ or $\mathbb{R}^3$ of complexity n degrades by a factor linear in n with constant probability. For the curve P we say that it is c-packed if for any point $p \in \mathbb{R}^d$ and any radius $r > 0$, the total length of the curve P inside the ball centered at p and with radius r is at most $c \cdot r$. We show upper and lower bounds on the distortion.

Given are input curves $P = \{p_1, p_2, \ldots, p_t\}$ and $Q = \{q_1, q_2, \ldots, q_k\}$ in $\mathbb{R}^d$. We consider the projections to a straight line chosen uniformly at random, and we are interested how does this projection affect the Fréchet distance $d_F(P, Q)$ of the curves P and Q (for both discrete and continuous Fréchet distance cases). Also we want to consider the same problem on the related similarity measure of dynamic time warping (DTW).

Let $\mathbf{u}$ be a unit vector that is chosen uniformly at random on the unit hypersphere and let L be the straight line through the origin that contains the vector $\mathbf{u}$. Let P' and Q' be the respective projections of P and Q to L. Given two curves P and Q in $\mathbb{R}^d$, the *traversal* T of P and Q is the sequence of pairs of indices (i, j) of vertices $(p_i, q_j) \in P \times Q$ s.t.

i) the traversal T starts with $(1, 1)$ and ends with (n, n), and
ii) the pair (i, j) of T can be followed only by one of $(i+1, j)$, $(i, j+1)$ or $(i+1, j+1)$.

We notice that every traversal is monotone. If $\mathcal{T}$ is the set of all traversals T of P and Q, then the discrete *Fréchet distance* between P and Q is defined as

$$d_F(P, Q) = \min_{T \in \mathcal{T}} \max_{(i,j) \in T} \|p_i - q_j\|. \tag{1}$$

Related similarity measure between two curves is dynamic time warping. It considers the *sum* of the used distances in the traversal (instead the maximum one). The continuous Fréchet distance observes the maximum over all possible monotone matchings of the points along P and Q (and not only the vertices of the curves).

Using dynamic programming the discrete Fréchet distance and the dynamic time warping distance of two polygonal curves of complexity n can be computed in $O(n^2)$ (Eiter, Mannila 1994). The state-of-the-art algorithm computes the discrete Fréchet distance in $O(n^2 \log\log n / \log n)$ (Agarwal *et al.* [1]).

Unless the Strong Exponential Time Hypothesis (SETH) fails, none of considered distance measures can be computed in strongly subquadratic time, shown by Bringmann [2] for Fréchet distance, and by Bringmann and Künnemann [3] for DTW.

Upper bound

Let p_i and q_j be two vertices of the curves P and Q. and let p_i' and q_j' be their projections to L. If the line segment $\overline{p_i q_j}$ is projected to the straight line L, supported by the unit vector chosen uniformly at random on the unit hypersphere in $\mathbb{R}^2$ or $\mathbb{R}^3$, the probability

that its length will be reduced by a factor greater than φ is at most φ. For the dimensions $d \geq 4$, it can be shown that in $\mathbb{R}^4$ there is an additional factor of $1+2/\pi$ to the probability ϕ. For dimensions $d \geq 5$ this factor grows in s.t. it cannot be bounded by a linear approximation.

We are interested to analyse the pairs of vertices of P and Q whose distance is larger or equal than some treshold, therefore we introduce the notion of the *guarding set*.

Definition 1. *For any two polygonal curves $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_n\}$ and a given parameter $\theta \geq 1$, a θ-guarding set B for P and Q is a subset of the set of pairs of vertex indices of P and Q s.t. a) for all $(i, j) \in B$, it holds that $\delta_{i,j} \geq d_F(P, Q)/\theta$, and b) for any traversal T of P and Q, it is $T \cap B \neq \emptyset$.*

A θ-guarding set B can be found by a simple algorithm that tests all possible traversals. Such set B can have a quadratic size in terms of the input curves. If we assume that the input curves are c-packed for some constant $c \geq 2$, this is no longer possible. Then our analysis carefully selects the entries of a guarding set B with a loss of a constant approximation factor. This implies the following lemma.

Lemma 2. *For the given c-packed curves $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_n\}$ from $\mathbb{R}^2$ or $\mathbb{R}^3$, let P' and Q' respectively be the projections to the one-dimensional space supported by the unit vector chosen uniformly at random on the unit hypersphere. For any $\gamma \in (0, 1)$ it holds that $Pr[d_F(P, Q)/d_F(P', Q') \in \mathcal{O}(n)] \geq 1 - \gamma$.*

Lower Bound

It may happen that for some two curves P and Q it holds that the ratio between Fréchet distance of the curves and the Fréchet distance of the respective projection curves P' and Q' is at least $\Omega(n)$, where n is the complexity of the curves. Let the curves P and Q be c-packed curves in a plane, for any constant $c \geq 2$. The following lemma holds for the discrete Fréchet distance (but it holds also for the continuous Fréchet distance and for the dynamic time warping distance).

Lemma 3. *For the given c-packed curves $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_n\}$, let P' and Q' respectively be the projections to the one-dimensional space supported by the unit vector chosen uniformly at random on the unit hypersphere. For any $\gamma \in (0, 1/\pi)$ it holds that $Pr[d_F(P, Q)/d_F(P', Q') \in \Omega(n)] \geq 1 - \gamma$.*

If the curves P and Q are not c-packed, for any constant $c \geq 2$, we may even reduce the ratio of the Fréchet distance between P and Q and their projection curves P' and Q' by a factor of at least $\Omega(n)$, and that can happen with probability 1. The same bound holds for the discrete Fréchet distance too.

Lemma 4. *There exist the curves $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_n\}$, such that if P' and Q' respectively are their projections to the one-dimensional space supported by the unit vector chosen uniformly at random on the unit hypersphere, then it holds that $d_F(P, Q) / d_F(P', Q') \in \Omega(n)$.*

Future work

We want to extend our results on the upper bound to the continuous Fréchet distance and the dynamic time warping distance. One idea would be to use simplifications of the input curves. A μ-simplification of a polygonal curve $P = \{p_1, \ldots, p_n\}$ is a curve whose set of vertices is a subset of the set of the vertices of P, and its edges have the length at least $\mu > 0$. Since the Fréchet distance is a metric, it holds that:

Lemma 5. *Given two c-packed curves P and Q and their μ-simplifications $\hat{P}$ and $\hat{Q}$ respectively, for $\mu > 0$. Then it holds that $d_F(P, Q) - 2\mu \leq d_F\left(\hat{P}, \hat{Q}\right) \leq d_F(P, Q) + 2\mu$.*

Unfortunately, projecting the curves P and Q to a straight line does not have to reduce the distance between the respective projections at all.

Another idea would be to use the idea of well-separated pair decomposition, in order to bound the sum of the distances of the matched points in the dynamic time warping distance. This approach has the similar problem as the simplification one.

References

[1] P. K. Agarwal, R. Ben Avraham, H. Kaplan, and M. Sharir. Computing the discrete Fréchet distance in subquadratic time. *SIAM J. Comput.*, 43(2):429–449, 2014.

[2] K. Bringmann. Why walking the dog takes time: Fréchet distance has no strongly subquadratic algorithms unless SETH fails. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 661–670, 2014.

[3] K. Bringmann and M. Künnemann. Quadratic conditional lower bounds for string problems and dynamic time warping. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015*, pages 79–97, 2015.

[4] A. Driemel, A. Krivošija, and C. Sohler. Clustering time series under the Fréchet distance. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 766–785, 2016.

[5] A. Driemel and F. Silvestri. Locally-sensitive hashing of curves. In *33st International Symposium on Computational Geometry, SoCG 2017*, pages 37:1–37:16, 2017.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A3

Methods for Efficient Resource Utilization in Machine Learning Algorithms

Peter Marwedel Jörg Rahnenführer

Stable Feature Selection for High-Dimensional Data

Andrea Bommert
Faculty of Statistics
TU Dortmund University
andrea.bommert@tu-dortmund.de

Finding a good predictive model for a high-dimensional data set can be challenging. For applications like genetic data it is not only important to find a model with high predictive accuracy, but it is also important that this model uses only few features and that the selection of these features is stable. This is because in bioinformatics, the models are not only used for prediction but they are also used for drawing biological conclusions which makes the interpretability and reliability of the model crucial. We suggest using three target criteria when fitting a predictive model to a high-dimensional data set: the classification accuracy, the stability of the feature selection, and the number of chosen features. To find out which measure is best for evaluating the stability, we first compare a variety of stability measures. We conclude that the Pearson correlation has the best theoretical and empirical properties. Also, we find that for the stability assessment behaviour it is most important if a measure contains a correction for chance or large numbers of chosen features. Then, we analyse Pareto fronts and conclude that it is possible to find models with a stable selection of few features without losing much predictive accuracy.

1 Stability Assessment

In some domains it is important, which features are chosen by a feature selection method, e.g. because these features will be subject to further research. Therefore, it is important that the results are reliable as such research is often very costly. Reliable results can

be obtained by a stable feature selection. This means that the sets of chosen features should be similar for similar data sets.

To quantify the stability of a feature selection, stability measures are used. We define one exemplary stability measure, others can be taken from [3]. We use the following notation: Assume there is a data set containing n observations of the p features $X_1, \ldots, X_p$. Resampling is used to split the data set into m subsets. The feature selection method is then applied to each of the m subsets. Let $V_i \subset \{X_1, \ldots, X_p\}$, $i = 1, \ldots, m$, denote the set of chosen features for the i-th subset of the data set and $|V_i|$ the cardinality of this set.

The Pearson correlation can be used as a stability measure. To do so, [4] define a vector $z_i \in \{0, 1\}^p$ for each set of selected features V_i to indicate which features are chosen. The j-th component of z_i is equal to 1 iff V_i contains X_j, i.e. $z_{ij} = \mathbb{I}_{V_i}(X_j)$, $j = 1, \ldots, p$. The resulting stability measure is

$$\mathrm{SC} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathrm{Cor}(z_i, z_j)$$

with $\mathrm{Cor}(z_i, z_j)$ denoting the Pearson correlation between z_i and z_j. The Pearson correlation measures the linear association between continuous variables. When applied to binary data like the vectors $z_1, \ldots, z_m$, the Pearson correlation is equivalent to the phi coefficient for the contingency table of each two of these vectors. Large values mean high stability, small values mean low stability, $-1 \leq \mathrm{SC} \leq 1$.

2 Comparison of Stability Measures

Figure 1 shows the empirical behaviour of twelve stability measures. We have applied many different feature selection methods. Some of them selected useful features for class prediction, others just chose features at random. We have performed resampling to generate similar datasets. For each feature selection, we have assessed the stability with all of the twelve stability measures and we have recorded the mean number of chosen features.

For all stability measures it is possible to take on small or large values if the mean sizes of the resulting models of the configurations are small. The measures SN, SO, SD, SZ, SJ, and SD-0 are not corrected for chance, i.e. they assign the higher stability values the more features are chosen. For the measures SD-1, SD-2, and SD-10 it is the other way round: They assign the lower stability values the larger the mean model sizes of the configurations are. The measures SL, SC and SS are corrected for chance, so the expected stability value is independent of the number of chosen features. However, for SL it is not possible to attain high values for a medium number of chosen features. For SS, even random

feature selections are not assigned the minimum value. SC shows the best empirical properties. For almost all number of chosen features, SC discriminates between stable and unstable feature selections. Additionally, [4] show that SC is a desirable stability measures from a theoretical point of view. We conclude that SC is a suitable measure for stability analysis.

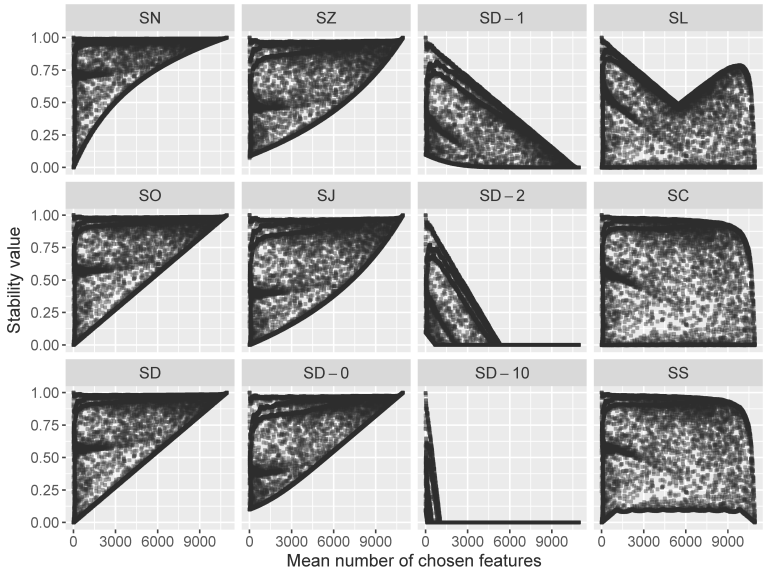


Figure 1: Scatter plot of the stability values and mean number of chosen features for twelve stability measures.

3 Finding Predictive, Sparse, and Stable Models for High-Dimensional Data

We have considered high-dimensional microarray and RNAseq datasets. We have combined filter methods and classifications methods in a way that the filter method is applied first and the classification method is learnt on the remaining features. So finding predictive, sparse, and stable models actually means tuning the hyper parameters of the augmented methods with respect to prediction accuracy, sparsity and stability of the feature selection. To do so, we have split the data into halves: training and testing data. On the training data, we have conducted a random search for desirable configurations. The testing data was used to determine the performance on new data of the configurations which were Pareto optimal on the training data.

Figure 2 shows the results for one of the datasets. It exemplarily shows that it is possible to choose configurations with a stable selection of few features without losing much predictive accuracy compared to model fitting only based on predictive performance.

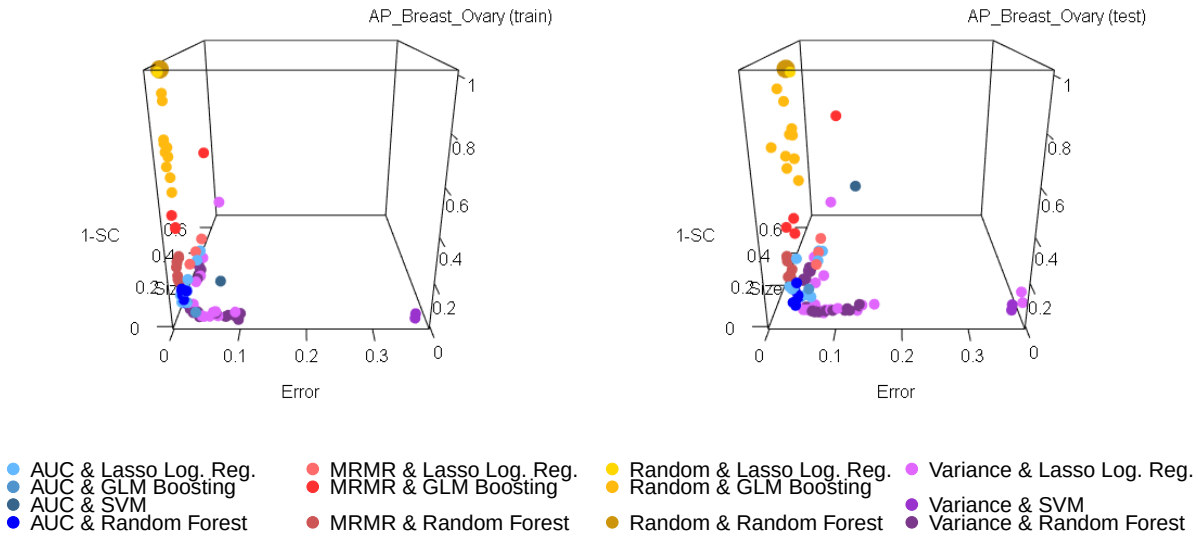


Figure 2: Pareto optimal configurations with respect to mean misclassification rate (error), mean model size (size) and stability measured by SC (1-SC for minimisation). The optimal point would be $(0, 0, 0)'$.

4 Future Work

In our future work, we will make the search for desirable configurations more efficient using MBO [1]. Also, we will integrate stability as a performance criterion in mlr [2], so that everyone can select configurations considering their stability and therefore obtain reliable results.

References

[1] Bernd Bischl, Jakob Bossek, Daniel Horn, and Michel Lang. *mlrMBO: Model-Based Optimization for mlr*, 2015. R package version 2.10.

[2] Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. mlr: Machine learning in R. *Journal of Machine Learning Research*, 17(170):1–5, 2016.

[3] Andrea Bommert, Jörg Rahnenführer, and Michel Lang. A multicriteria approach to find predictive and sparse models with stable feature selection for high-dimensional data. *Computational and Mathematical Methods in Medicine*, 2017:Article ID 7907163, 2017.

[4] Sarah Nogueira and Gavin Brown. Measuring the stability of feature selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 442–457. Springer, 2016.

RAMBO on Homogeneous Systems and Heterogeneous Embedded Systems

Helena Kotthaus
Computer Science 12
TU Dortmund University
helena.kotthaus@tu-dortmund.de

Model-based optimization (MBO) is a global optimization method for black-box functions that are expensive to evaluate. One of its uses is parameter tuning of learning algorithms. The goal is to find the configuration with the best performance within a limited time budget. Due to huge model spaces a large amount of resources is needed to evaluate the configurations. The goal is to execute MBO in a resource efficient way to enable the processing of larger problem sizes within a given time budget. We present resource-aware scheduling strategies that are included into our resource-aware model-based optimization framework, called RAMBO to efficiently map configurations to the underlying parallel architecture. We demonstrate the effectiveness of RAMBO, targeting homogeneous systems and the heterogeneous ARM big LITTLE architecture, commonly found in mobile phones.

1 Scheduling for MBO on Homogeneous Systems

MBO uses a regression model to approximate the objective function and iteratively proposes new interesting configurations for evaluation. Deviating from the original formulation, it is often indispensable to apply parallelization to speed up computation. This is usually achieved by evaluating as many configurations per iteration as there are workers (CPUs) available. However, if runtimes of the configurations are heterogeneous, resources might be wasted by idle workers. To accomplish efficient resource utilization a runtime estimation model is developed to guide the mapping of evaluations to available resources. In addition to the runtime estimates, the scheduling strategies use an execution priority reflecting the profit of an evaluation for finding the best configuration [3, 4].

	4 CPUs			16 CPUs		
Algorithm	0.5	0.1	0.01	0.5	0.1	0.01
asyn.eei	3.32 (2)	3.52 (1)	4.97 (2)	3.75 (3)	4.30 (3)	5.45 (3)
asyn.ei.bel	3.55 (3)	4.10 (3)	4.97 (2)	3.48 (2)	4.08 (2)	4.53 (2)
RAMBO	**3.17 (1)**	**3.85 (2)**	**4.57 (1)**	**3.13 (1)**	**3.93 (1)**	**4.47 (1)**
ei.bel	4.38 (4)	4.98 (4)	5.90 (5)	5.00 (5)	5.48 (6)	6.28 (6)
qLCB	4.52 (5)	5.03 (5)	5.63 (4)	4.72 (4)	5.17 (4)	6.10 (4)
rs	6.02 (6)	6.67 (6)	6.83 (7)	5.50 (7)	6.48 (7)	6.87 (7)
smac	6.22 (7)	6.70 (7)	6.82 (6)	5.32 (6)	5.47 (5)	6.17 (5)

Table 1: Ranking for accuracy levels 0.5, 0.1, 0.01 on 4 and 16 CPUs [4].

We extended the scheduling strategy proposed in [3] with an improved resource-aware scheduling algorithm [4]. This algorithm, which replaces the original simple first fit heuristic, is based on a knapsack solver to better handle heterogeneous runtimes. We compared our new approach to five established MBO parallelization strategies on a set of continuous functions with heterogeneous runtimes [4]. Compared to the considered approaches, our new RAMBO approach converges faster to the optima if the runtime estimates used as input for scheduling are reliable. Table 1 lists the aggregated ranks over all benchmark functions, grouped by MBO algorithm, accuracy level, and number of CPUs. We compare the approaches at three accuracy levels 0.5, 0.1 and 0.01 that represent the distance between the best found configuration at time t and a predefined target value. The target value is the best configuration found by any MBO method after the complete time budget. On average, RAMBO reaches the accuracy level first in 2 of 3 setups on 4 CPUs and is always fastest on 16 CPUs.

Figure 2 exemplary visualizes the mapping of the parallel configuration evaluations (jobs) for all MBO algorithms on 16 CPUs. The necessity of a resource estimation for jobs with heterogeneous runtimes becomes obvious, as qLCB and ei.bel can cause long idle times by queuing jobs together with large runtime differences. The knapsack based scheduling manages to clearly reduce this idle time. Overall, the resource utilization obtained by the scheduling in RAMBO leads to faster and better results, especially, when the number of available CPUs increases.

For future work we plan to develop a resource-aware scheduling strategy that combines the parallel asynchronous and synchronous MBO approaches to reduce the computational overhead of the asynchronous approach.

2 Scheduling for MBO on Heterogeneous Systems

We developed a resource-aware scheduling strategy for parallelizing MBO on heterogeneous architectures, like those commonly found in mobile devices. Such devices typically

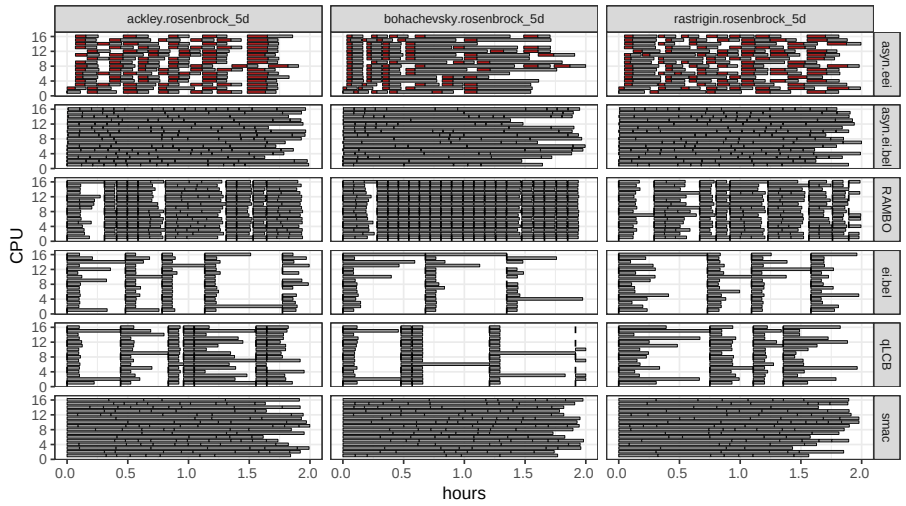


Figure 1: Scheduling of MBO approuches including `RAMBO`. Time on x-axis and mapping of jobs to CPUs on y-axis. Gray boxes represent jobs. Red boxes represent overhead for the asynchronous approaches. Gaps represent CPU idle time [4].

consist of different processors with different frequencies and memory sizes, and are characterized by tight resource and energy restrictions. The original parallel package [1] that is part of the R distribution, targets problems that can be decomposed into independent tasks that are then processed in parallel. It is also used for parallelizing MBO on homogeneous architectures. However, as the `parallel` package does not support heterogeneous architectures, it is ill-suited for the kinds of systems we are considering. To support heterogeneous systems within R, we improved the parallel package. Our changes are integrated into the devel-branch of the R distribution [2].

When MBO is conducted on heterogeneous systems the execution time of an evaluation of a configuration can vary heavily not only depending on the configuration but also on the underlying architecture. Key to our approach is a regression model that estimates the execution time of a configuration for each available processor type. In combination with a knapsack based scheduler allowing to allocate tasks to specific processors, we enable resource-aware scheduling of MBO to optimize its overall runtime. In cooperation with the B2 project, we demonstrated the effectiveness of our approach targeting the ARM big LITTLE architecture of the Odroid-XÚ3 platform, commonly found in mobile phones [5].

Figure 2 shows an exemplary result for one of the benchmark functions also used in [4]. `RAMBO` is able to outperform the default synchronous parallel MBO approach. It converges faster to the optimum.

For future work we plan to analyze the energy consumption of our resource-aware scheduling strategies. Furthermore we want to perform a comparison study with the asyn-

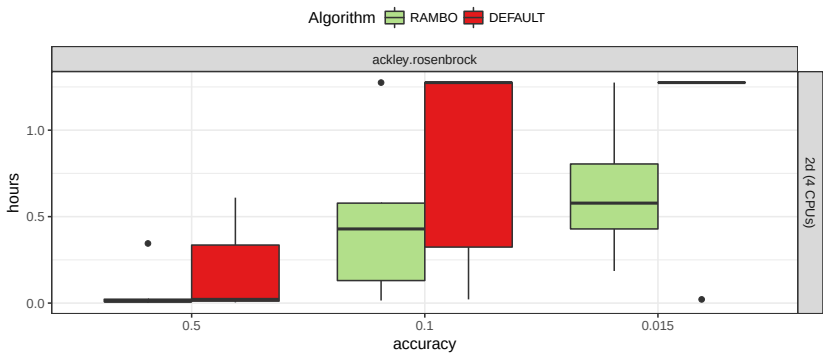


Figure 2: Accuracy level vs. execution time (lower is better)

chronous MBO approaches on the Odroid platform.

References

[1] Ripley B, Tierney L., Urbanek S., parallel: Support for Parallel Computation. R package first included in the R-core 2.14.0. 2015 URL http://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf

[2] Kotthaus H., Lang A., affinity.list: Support for Heterogenous Processors. Included in the Parallel R Package, R-devel branch, Sep. 18, 2017. URL https://developer.r-project.org/blosxom.cgi/R-devel/NEWS/2017/09/18n2017-09-18 Setting the CPU Affinity with mclapply: https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf

[3] Richter J., Kotthaus H., Bischl B., Marwedel P., Rahnenführer J., Lang, M., Faster Model-Based Optimization through Resource-Aware Scheduling Strategies. Proceedings of LION'10. LNCS vol. 10079, pp. 267-273, 2016

[4] Kotthaus H., Richter J., Lang A., Thomas J., Bischl B., Marwedel P., Rahnenführer J., Lang, M., RAMBO: Resource-Aware Model-Based Optimization with Scheduling for Heterogeneous Runtimes and a Comparison with Asynchronous Model-Based Optimization. Proceedings of LION'11. LNCS vol. 10556, pp. 180-195, 2017

[5] Kotthaus H., Lang A., Neugebauer O., Marwedel P., R goes Mobile: Efficient Scheduling for Parallel R Programs on Heterogeneous Embedded Systems. UseR! (to appear). Brussels, Begium, 2017

Energy measurement made simple on embedded systems

Olaf Neugebauer
Lehrstuhl für Technische Informatik und Eingebettete Systeme
Technische Universität Dortmund
olaf.neugebauer@tu-dortmund.de

Energy consumption is an important objective not only in the embedded domain. Even large server farms are affected by power budgets along with the thermal dissipation. Providing the necessary information in an easy way to the software developer is not always simple. Some techniques require deep understanding of electronics and special measurement equipment. In this report, we present easy solutions for software developers to determine the energy requirements of their application.

1 Energy measurement on Odroid-XU3

The Odroid-XU3 platform [1] uses the INA 231[8] to sense voltage, current and power. Separate sensors are connected to the A15, A7 CPUs as well as to the memory and GPU. The sensors are connected through a standardized I^2C interface with the MP-SoC. The sensors can be configured, for instances the averaging and conversion times. However, the original driver and Linux implementation shipped with the platform, does not support reconfiguration during runtime. Thus, we extended the driver provided by the manufacture to enable dynamic configuration of the sensors during runtime. This enables a rapid analysis with different sensor settings without time-consuming rebooting and reconfiguring of the target platform. To make the measured sensor values accessible, we developed two applications, the EnergyMeter and the Energy Relay Reader. In both applications, the systems measures the energy consumption in situ. The results are time coded values stored in a CSV file. This enables a straightforward graph generation e.g. with gnuplot like shown in Figures 1 and 2. Since the measurement applications use

standardized methods to read the sensor values, we think our applications can also be used for other sensors. Our driver extensions and measurement applications are available at [7].

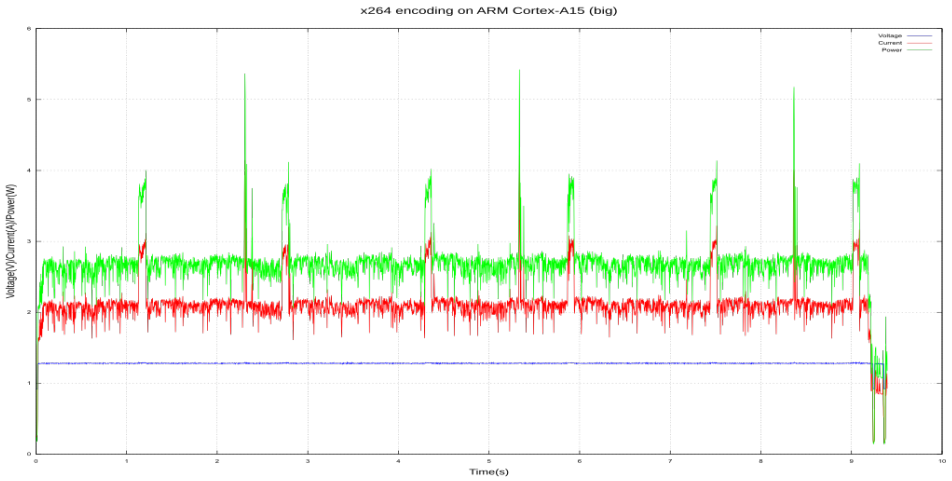


Figure 1: Voltage, current and power for x264 encoding on the A15 cores

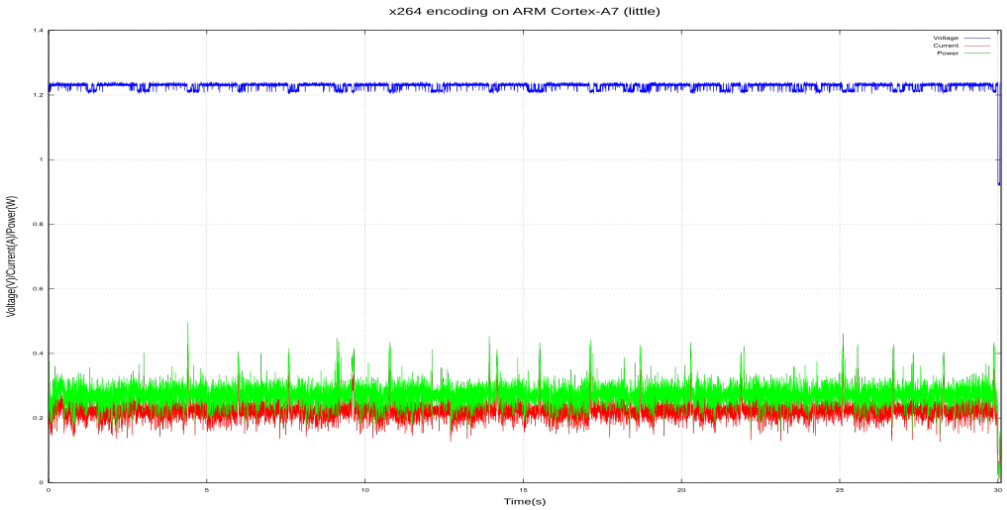


Figure 2: Voltage, current and power for x264 encoding on the A7 cores

2 EnergyMeter

The EnergyMeter implementation is derived from the original energy measurement application provided by the manufacturer of the Odroid platform. The original application required a graphical user interface and the measurement could not be controlled externally, e.g. to measure parts of an application. Our implementation removes these

limitation and we used it in [6, 4, 3]. The EnergyMeter uses polling to read out sensor values. Thus, it is able to sense dynamic changes in power requirements with a high resolution but produces therefore a high CPU workload. The sensing process can be controlled with external events. In this case, Linux named pipes are used to trigger the measurement. This feature provides the capability to measure parts of an application. We used this feature to exclude the initial setup phase of the virus detection algorithm[6] since it was the same over all optimization experiments and thus might hide the influence of certain optimizations to the energy consumption. Until the measurement application is terminated or an explicit measurement reset command is issued, it appends the energy values of multiple measurements.

3 Energy Relay Reader

A precise readout of the sensor data is not always required, thus, we developed the Energy Relay Reader which allows a relaxed and CPU friendly energy measurement. This application uses the relay feature of the Linux operating system. The basic idea is to create threads reading the energy sensor at specific points in time. These threads are then either scheduled to the worker or kernel thread queue. The queues differ in the priority they are processed by the operating system. Once the queues are processed, the results are added to the set of result files. This approach does not lead to a high CPU load since the CPU could idle between the measurements. A drawback of this approach is the possibility that due to queue processing, some readouts gets delayed drastically. It is also possible, that measurements are delayed until the termination of the experiments. This happens if measurement tasks remain, e.g. due to low priority, in the queues and get executed after the termination of the experiment. This behavior is shown if Figure 1 where the curves drop at the end and stay at idle. However, for long running applications or average measurements, these drawbacks are acceptable. We used the Energy Relay Reader in [5, 2].

References

[1] Hardkernel. *Odroid-XU3*. `http://www.hardkernel.com/main/products/prdt_info.php?g_code=G140448267127`. 2017.

[2] Helena Kotthaus et al. "R goes Mobile: Efficient Scheduling for Parallel R Programs on Heterogeneous Embedded Systems". In: *Abstract Booklet of the International R User Conference (UseR!) (to appear)*. July 2017.

[3] Peter Marwedel, Heiko Falk, and Olaf Neugebauer. "Memory-Aware Optimization of Embedded Software for Multiple Objectives". In: *Handbook of Hardware/Software Codesign*. Ed. by Soonhoi Ha and Jürgen Teich. Springer Netherlands, 2017, pp. 1–37. ISBN: 978-94-017-7358-4. DOI: 10.1007/978-94-017-7358-4_27-2. URL: https://doi.org/10.1007/978-94-017-7358-4_27-2.

[4] Olaf Neugebauer, Michael Engel, and Peter Marwedel. "A Parallelization Approach for Resource-Restricted Embedded Heterogeneous MPSoCs Inspired by OpenMP". In: *Journal of Systems and Software* 125 (2016), pp. 439–448. ISSN: 0164-1212. DOI: http://dx.doi.org/10.1016/j.jss.2016.08.069.

[5] Olaf Neugebauer, Peter Marwedel, and Michael Engel. "Quality Evaluation Strategies for Approximate Computing in Embedded Systems". In: *Technol. Innov. Smart Syst.* Ed. by Luis M. Camarinha-Matos, Mafalda Parreira-Rocha, and Javaneh Ramezani. Vol. 499. IFIP Advances in Information and Communication Technology. Springer International Publishing, 2017, p. 2017. ISBN: 978-3-319-56076-2. DOI: 10.1007/978-3-319-56077-9.

[6] Olaf Neugebauer et al. "Plasmon-based Virus Detection on Heterogeneous Embedded Systems". In: *Proceesings of Workshop on Software & Compilers for Embedded Systems, SCOPES*. 2015.

[7] SFB 876 - Software. *Resource optimizing real time analysis of artifactious image sequences for the detection of nano objects*. 2017. URL: http://sfb876.tu-dortmund.de/auto?self=Software.

[8] Texas Instruments Incorporated. *High- or Low-Side Measurement, Bidirectional CURRENT/POWER MONITOR with 1.8-V I^2C Interface*. SBOS644-FEBRUARY 2013. Texas Instruments Incorporated. Feb. 2013.

Parallel Model Based Optimization of Expensive Black-box Functions with Heterogeneous Runtimes

Jakob Richter
Faculty of Statistics
TU Dortmund University
richter@statistik.tu-dortmund.de

The optimization of parametrized algorithms with a long runtime is a challenging task. The performance of the algorithm is measured by an arbitrary criterion. The number of algorithm evaluations of various settings that can be calculated within an acceptable time bound is limited. Model-based optimization is a popular technique for expensive black-box optimization. It reduces the evaluations of the expensive objective function (e.g., the algorithm) by fitting a surrogate regression model on the set of already evaluated parameter configurations. Iteratively an infill criterion will be optimized on the cheaper surrogate leading to a proposed configuration that is than evaluated on the objective function. The outcome is added to the set of evaluated parameters, and the surrogate is updated. Often the evaluation of a single algorithm cannot make use of the complete parallel infrastructure available to further increase the evaluation speed. This motivates the evaluation of the algorithm with different parameter configurations in parallel which is the focus of our research. Therefore the optimization algorithm has to generate a batch of proposed configurations for each iteration. The task of the RAMBO-Framework is to handle the conflicts that arise in such setting: 1. The proposals should balance exploration and exploitation of the search space. 2. If different configurations of the algorithm lead to different runtimes idling should be avoided. 3. An overallocation of shared resources such as main memory through parallel execution should be avoided. Furthermore, the suitability of model-based optimization for a wide range of expensive black-box problems has to be further investigated. Especially the reliability on

complex search spaces that include categorical and hierarchical parameters. Applications reach from hyperparameter tuning for machine learning methods to an automatic pipeline optimization of embedded virus detection system.

1 Synchronous parallelization

Ordinary MBO is sequential by design. However, the need to evaluate more configurations within the same time and the growth of parallel computer infrastructure have driven the rapid development of extensions for parallel execution of multiple points. One possible extension is to derive multiple points $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_q^*$ from the surrogate in each iteration. Most techniques use adapted infill criteria, like the qLCB [6], the qEI [4]. Other heuristics like are universally applicable [4]): The proposed configuration $\boldsymbol{x}_1^*$ and its predicted outcome from the surrogate (*Kriging believer*) or a constant value (*constant liar*) are added as *fake values* to the set of evaluated configurations to derive the next proposal and so on. Usually, the number of proposed points N equals the number of available workers m.

However, these methods still do use parallel resources inefficiently if the runtime of the algorithm is heterogeneous. In those cases, the workers will idle after they completed their evaluation and wait until new proposals are generated. But within the synchronous framework, this only can happen after all evaluations are finished. To tackle this problem we use a further regression model to predict the runtime of each configuration based on the runtime of past configurations. This enables us to schedule the evaluations beforehand so that idling is reduced.

The generation of the points and the clustering is done as follows:

1. Generate $q = c \cdot m$ (here $c = 3$) proposals minimizing $\text{qLCB}(\boldsymbol{x}, \lambda_k) = \hat{\boldsymbol{y}}(\boldsymbol{x}) - \lambda_k \hat{s}(\boldsymbol{x})$ with $\lambda_k \sim \text{Exp}(0.5)$ for $k = 1, \ldots, q$. This results in a set of points that are likely close to the optimum if λ_k is small or that lead to an exploration of the search space if λ_k is big. Each point for its own is purposeful for the optimization. However, if e.g. two points are close to each other, the evaluation of the second point does not lead to significant information gain given that the first one is calculated. To avoid this, the points are reprioritized as follows:
2. Initialize an empty list P that stores the proposals in a new order according to their priority starting with the highest priority.
3. Use a hierarchical cluster method to generate a dendrogram that represents the clusters of those points.
4. Start the procedure with $k = 1$ cluster and continue with step 6.
5. Separate the set of points into k clusters. Ignore clusters that contain points in P.

6. Select the proposal $\boldsymbol{x}_k^*$ that was generated with the smallest value of λ_k.
7. Append $\boldsymbol{x}_k^*$ to P, set $k = k + 1$ and go to step 5.

Finally, we assign new priorities $\tilde{p}_j$ based on the order of P, i.e. the first job in P gets the highest priority q and the last job gets the lowest priority 1. This procedure avoids points being close to each other to be evaluated by giving a higher priority to points of separate clusters. To select the proposal with the smallest λ_k ensures that we get closer to the predicted optimum. Exploration is ensured by evaluating points of different clusters.

We use knapsack scheduling to select the best subset of jobs that maximizes the profit defined by the priorities under the restriction that no worker should take more time than $\hat{t}_1$, which is the predicted runtime of the job with the highest priority. Furthermore, each job should only be run once.

2 Asynchronous parallelization

In the asynchronous approach, each worker runs independently, and a new proposal is generated on the bases of all completed evaluations as soon as the evaluation is finished. Hence idling does not occur anymore. As suggested in [3] a proposal has to take into account the configurations ($\boldsymbol{x}_{busy}$) that are still being evaluated on other workers. Their outcome is unknown but can be estimated using the surrogate similar to the *surrogate believer* approach. The expensive to calculate Expected Expected improvement (EEI) [3] has been showed to perform slightly worse then the naive *surrogate believer* approach [7]. Especially for high degrees of parallelization, the cost of the EEI can affect the performance of the optimization, as the time that should be used to evaluate the target algorithm is used to generate proposals. Furthermore, the problem of parallel point proposals seems to be neglected in literature so far. When worker A finishes and uses its resources to generate a new proposal it can happen that worker B finishes during that process. Here there could be a decision for A to abort the proposal generation as the most recent result of worker B should be incorporated. This would lead to a dynamic switch to a partially synchronous parallelization on all workers that are available.

3 Further Collaborations

In collaboration with the B2 project and Pascal Libuschewski, we use heavily parallelized multi-criteria MBO [5] to optimize the pipeline of the PAMONO Sensor optical virus detection. It includes the choice of preprocessing parameters as well as hardware parameters to optimize for the best hardware architecture to solve this problem fast, energy efficient and with a virus high detection rate. The result of a multi-criteria optimization

is a Pareto front giving a set of points, of which each is optimal, meaning that further increasing one target criterion will decrease at least one other criterion. The user can then decide which trade-off he is willing to take.

I participate actively in the development of the R-Package **mlr** [1] and **mlrMBO** [2] amongst other contributions. The latter is the foundation of the RAMBO Framework.

References

[1] Bernd Bischl et al. "mlr: Machine Learning in R". In: *Journal of Machine Learning Research* 17.170 (2016), pp. 1–5. URL: http://jmlr.org/papers/v17/15-066.html (visited on 11/07/2016).

[2] Bernd Bischl et al. "mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions". In: (Mar. 9, 2017). arXiv: 1703.03373 [stat]. URL: http://arxiv.org/abs/1703.03373 (visited on 03/16/2017).

[3] David Ginsbourger, Janis Janusevskis, and Rodolphe Le Riche. "Dealing with Asynchronicity in Parallel Gaussian Process based Global Optimization". In: *4th International Conference of the ERCIM WG on computing & statistics (ERCIM'11)*. 2011, pp. 1–27.

[4] David Ginsbourger, Rodolphe Le Riche, and Laurent Carraro. "Kriging is Well-Suited to Parallelize Optimization". In: *Computational Intelligence in Expensive Optimization Problems*. Springer, 2010, pp. 131–162.

[5] Daniel Horn et al. "Model-based multi-objective optimization: taxonomy, multi-point proposal, toolbox and benchmark". In: *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2015, pp. 64–78. URL: http://link.springer.com/chapter/10.1007/978-3-319-15934-8_5 (visited on 11/07/2016).

[6] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. "Parallel Algorithm Configuration". In: *Learning and Intelligent Optimization*. Springer, 2012, pp. 55–70.

[7] Helena Kotthaus et al. "RAMBO: Resource-Aware Model-Based Optimization with Scheduling for Heterogeneous Runtimes and a Comparison with Asynchronous Model-Based Optimization". In: *Learning and Intelligent Optimization*. International Conference on Learning and Intelligent Optimization. Lecture Notes in Computer Science. Springer, Cham, June 19, 2017, pp. 180–195. ISBN: 978-3-319-69403-0 978-3-319-69404-7. DOI: 10.1007/978-3-319-69404-7_13. URL: https://link.springer.com/chapter/10.1007/978-3-319-69404-7_13 (visited on 11/13/2017).

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A4

Resource efficient and distributed platforms for integrative data analysis

Michael ten Hompel Olaf Spinczyk
Christian Wietfeld

Automated Measurement of Energy Consumption in Deeply Embedded Systems

Markus Buschhoff
Department of Computer Science 12
Technische Universität Dortmund
markus.buschhoff@tu-dortmund.de

Creating energy models from measurements for all components of a deeply embedded system is a challenging task. To construct valid models, it is necessary to drive every component of a system through all possible states while thoroughly observing the energy consumption on sophisticated and complex measurement devices. Additionally, the influence of hardware and application parameters, like the packet length when sending network packets, has to be analyzed. To apply the resulting model for online use, it is necessary to assign the measured values to system function calls, which imposes the requirement that system functions and their parameters structurally match the energy model. This requires an interface that is shaped to support energy model mapping. In the following, a concept is shown that supports the creation of energy-aware driver interfaces, exploits their structure to derive an automated measurement scheme, and maps the captured model to driver interface calls. An automatic analysis streamlines the results and calculates models for function parameters by utilizing regression algorithms. In subsequent iterations, the model can be refined until it matches a given accuracy requirement.

1 Introduction

To describe the energy consumption of peripheral hardware components like sensors or radio transceivers, priced and optionally timed automata (PTA) models are a common

solution in the literature. These models combine the simplicity of mapping functionalities of devices to automata states, which is a common practice in device documentations, and the integration of a non-functional cost model for each state and transition. Also, a functional device state can cover multiple non-functional states (i.e. if the energy consumption changes during a functional state). By using a timing model, these non-functional sub-states can also be covered. PTA based system drivers have shown highly accurate results with low overhead when used for energy accounting in low-power,deeply-embedded systems [1].

For the approach shown here, the following assumptions on peripheral devices are made:

1. A state transition can be triggered by a driver function.
2. A state transition can be triggered by a device-internal function and then is communicated to the CPU, e.g. by an interrupt request.
3. Other state changes do not occur. Internal, non-functional state changes are not considered for reasons of simplicity in this report.

This means that a transition can only occur on either a driver function call or before running an interrupt service routine. By that, the CPU is always aware of state changes within a device. By annotating the energy consumption for each state and transition within the driver code, the CPU is able to efficiently calculate and account the energy consumption for every device (including the CPU itself) during runtime [1].

2 System Setup

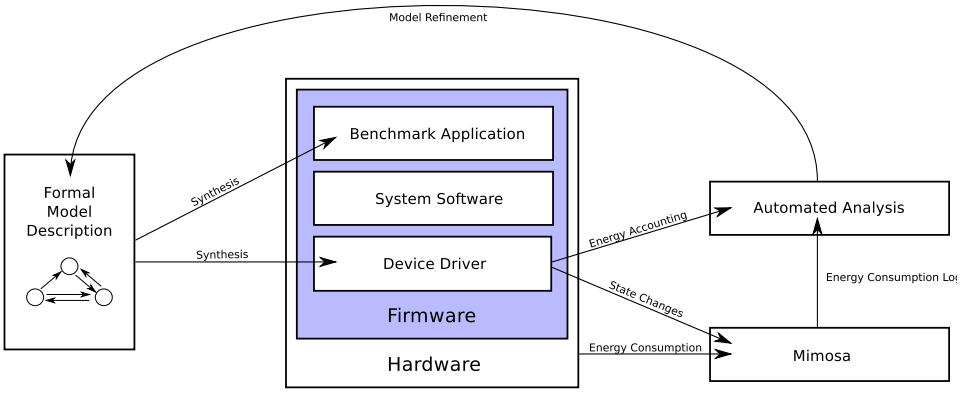


Figure 1: System overview

As an extension to this already existing method, a concept was evaluated that synthesizes system drivers for peripheral components based on an unannotated formal description of the functional automaton model of a device. This concept is depicted in Figure 1.

The automaton model contains all information necessary to create a device-driver software interface. In combination with a given automaton input sequence, a "benchmark" application that consequently iterates over all PTA states is generated. The benchmark code calls the synthesized system functions according to the automaton input. Additionally, the benchmark signals the driver calls to the system environment using an I/O pin of the CPU.

An external measurement device (the MIMOSA measurement equipment [2, 4] that was formerly constructed in the A4 project) is used to measure the energy consumption during consecutive benchmark runs. The collected results get analyzed by an automatic analysis toolchain which determines the average energy consumption during each state and transition. If the driver interface already contains energy annotations that were created during former iterations, the driver's energy accounting data is also gathered, so that the accounting accuracy can be compared to actual measurements.

3 Analysis

The analysis tool can deploy regression based methods to determine the dependency between function parameters and their respective energy consumption. To achieve this, the formal model description can be annotated with validity ranges for parameter values. The synthesis toolchain then creates code that not only iterates over all states and transitions, but also over all parameter ranges.

When a parameter is written into a register of a hardware device during a system function call, it might have a sustained effect on the device. For this reason it is necessary to identify the states and transitions that are influenced by a certain parameter. To achieve this, all parameters are considered global, and the deviation (noise) in energy consumption during runs with fixed parameters is compared to the deviation during measurement iterations with changing parameters.

Afterwards, for each parameter-affected part of the automaton model, a set of regression functions are calculated for the influencing parameters, among them linear, logarithmic, shifted logarithmic, hyperbolic exponential, square, root, and 0/1 bit-count regressions. From these functions, the one with the lowest residual sum of squares (RSS) is selected as best fit for further analysis steps.

Then, with all selected functions $F = (f_1, ..., f_m)$ and the corresponding parameters $p_1, ..., p_m$ influencing a given model property, a compound function $g(\vec{p})$ is created:

$$g(\vec{p}) = \sum_{F' \in \mathcal{P}(F)} \left(a_{F'} \cdot \prod_{f \in F'} f(\vec{p}) \right)$$

The coefficients $a_{F'}$ of the resulting function $g(\vec{p})$ are then optimized to minimize the *root mean square deviation* (RMSD) by regression. The quality of the result is determined by cross validation.

4 Results

The concept of automatic model measurement, generation and refinement was implemented and evaluated in [3].

It was shown that the calculated model functions for a diversity of different hardware components have an average error of below 1.5%. The largest error was seen in the model for a radio transceiver (nRF24), which has a maximum deviation of 6%.

The model refinement process for the analyzed components sometimes required minor manual interactions. But, in most cases the process was completely automatic.

The measurement and refinement of peripheral hardware took at least 20 minutes for simple hardware (LM75 temperature sensor), one hour for more complex hardware (nRF24 radio transceiver), and up to three hours for a complex transceiver module (TI CC1200 radio transceiver). The automated analysis of the gathered data alway took less than 2 minutes on on a Core i5-2320 CPU.

References

[1] Markus Buschhoff, Robert Falkenberg, and Olaf Spinczyk. Energy-aware device drivers for embedded operating systems. *SIGBED Rev.*, 2018. to appear.

[2] Markus Buschhoff, Christian Günter, and Olaf Spinczyk. MIMOSA, a highly sensitive and accurate power measurement technique for low-power systems. In *Real-World Wireless Sensor Networks*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013.

[3] Daniel Friesel. Automatisierte verfeinerung von energiemodellen für eingebettete systeme. Masterarbeit, Universität Dortmund, Germany, March 2017.

[4] Steve Kerrison, Markus Buschhoff, Jose Nunez-Yanez, and Kerstin Eder. Measuring energy. In Giorgos Fagas, Luca Gammaitoni, John P. Gallagher, and Douglas J. Paul, editors, *ICT - Energy Concepts for Energy Efficiency and Sustainability*, chapter 03. InTech, Rijeka, 2017.

Suitability of Bluetooth 5 for Realistic Internet of Things Activity Levels and Interference Scenarios

Stefan Böcker
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
stefan.boecker@tu-dortmund.de

Internet of Things applications have received a certain amount of interest, whereby especially many communication technologies are discussed in order to enable large-scale deployments. In this regard the Bluetooth Special Interest Group has proposed the next generation Bluetooth 5 specification in order to address the growing requirements for ease of use, cheap and energy efficient Internet of Things network solutions. Thus, Bluetooth 5 mainly increases the maximum communication range while adapting the corresponding data rates, which opens up new IoT applications for Smart Home or Building, Industry 4.0, as well as Smart Sity environments. This work aims to analyze the suitability of Bluetooth 5 for large-scale deployments realistic application activity levels and interference scenarios, based on an analytical models and supporting simulations.

1 Introduction

Internet of Things (IoT) applications have experienced significant attention and growth related to large scale deployments, establishing smart object and sensor networks, covering use cases of industrial, commercial, public as well as private domains. In order to address related functional and technical requirements many potential communication technologies are discussed to cover a best possible range of applications. Especially technologies tailored to specific IoT requirements, e.g. LTE evolutions NB-IoT or eMTC as well as extensive discussed Low Power Wide Area Networks (LPWAN), such as LoRa, needs to be considered.

The well-known and wide-spread Bluetooth Low Energy specification is still one of the market leaders to connect several sensor applications in the coverage area of Personal Area Networks (PAN). To close the gap to other network areas and to in order be ready for upcoming IoT markets, the Bluetooth Special Interest Group (SIG) provides the Bluetooth 5 specification (BT5), which improves data rates without increasing the energy consumption, broadcast messaging capacity, interoperability and long-range communication. Especially the substantially improved communication range enables IoT support and elevates Bluetooth 5 to a new level, covering new application areas beyond typical BLE PAN networks. In order to stress these new BT5 functionalities in large-scale deployments, this work presents a scalability analysis based on typical application activities and interference scenarios.

2 Approach and Implementation

As illustrated in Fig.1 the BT5 scalability analysis is performed by two independent modeling approaches. First an analytical model is developed to analyze the Packet Error Rate (PER) for n interfering BT5 devices, considering interference situations and thus usable spectrum for BT5 connectivity S_{BT}, as well as activity levels of individual piconets G. In order to verify the analytical model, this work implements a simulation model covering the BT5 Channel Selection Algorithm for each considered individual piconet and the above introduced activity levels G. Within a cross validation process reliability of both models is verified, whereby both models rely on the assumption that all packets of a specific communication interval are broken in case of an error within the corresponding interval. A detailed description of these models and insights are given in [1].

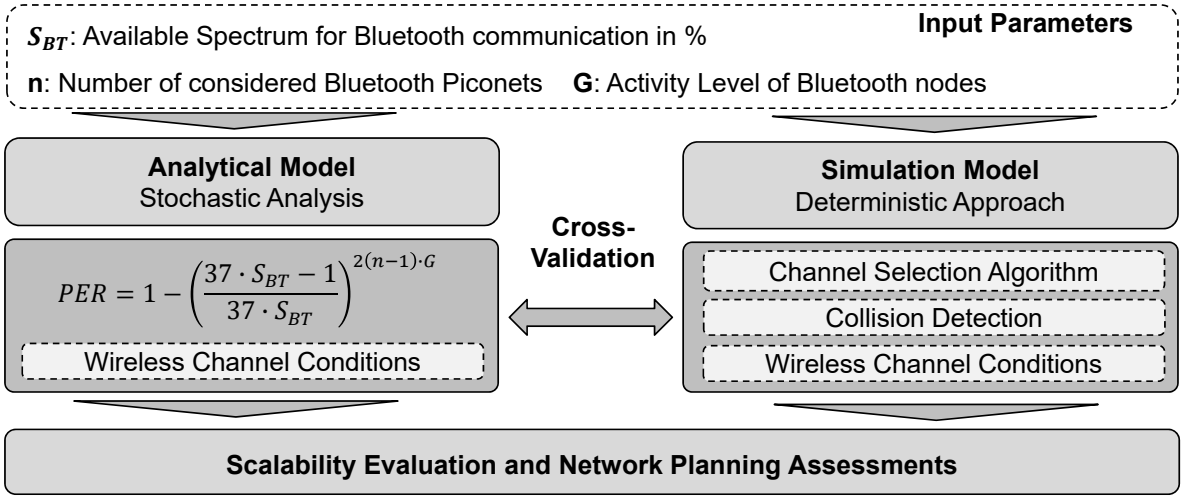


Figure 1: Modeling approaches for Bluetooth 5 scalability analysis

By adding empirical channel models to the approaches presented in [1], geographical scenario constellations of different environments can be considered as well. The consideration of receiver sensitivities for both, application as well as interfering activities, enables concrete network planning assessments.

3 Performance and Scalability Evaluation

Our performance evaluation is based on four activity levels that rely on realistic IoT use cases: Traffic congestion, Noise Monitoring, Tap Water Observation and Waste Management. Based on this, the activity levels vary from the highest communication effort of $G = 0.01$ for Traffic Congestion services (500 up 1000 monitoring events per day) down to $G = 0.00001$ for Waste Management services (1 or 2 events per day). The interference situation is covered by four different constellations of non-overlapping WiFi channels, that reduce the usable spectrum for Bluetooth communication S_{BT} from $S_{BT} = 100\%$ (no active WiFi channel detected) down to $S_{BT} = 100\%$ (3 competing WiFi channels detected).

Assuming a maximum allowed error rate of $PER = 1\%$, in order to guarantee a highly reliable IoT communication link, Fig.2 illustrates that the resulting PER is significantly varying related to the above introduced activity level as well as the interference situation within the ISM band. For the highest activity level of $G = 0.01$, the maximum PER limits the number of usable piconets to 20 ($S_{BT} = 100\%$) respectively 5 piconets ($S_{BT} = 25\%$). In case of the smallest activity level $G = 0.00001$, a significantly higher capacity limit is determined with a limit of little bit less than 19900 BT5 devices. A more detailed analysis is given in [1]

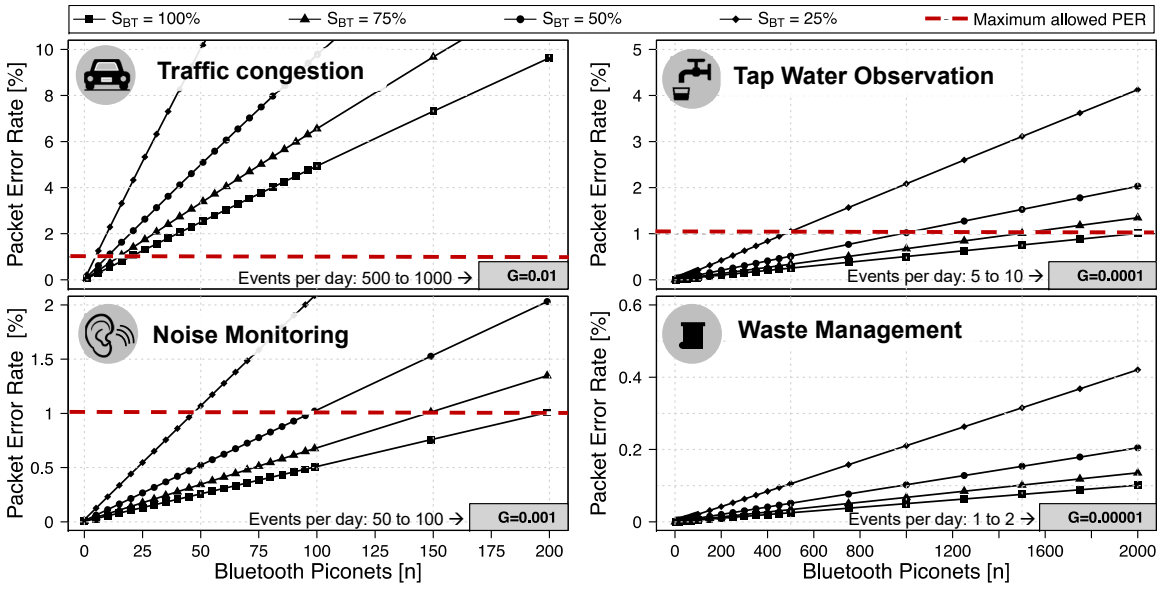


Figure 2: Packet Error Rate (PER) for relevant IoT activity scenarios [1]

4 Conclusion and Further Research

In this work a scalability analysis based on two modeling approaches is presented in order to evaluate the suitability of BT 5 for typical IoT activity levels and interference scenarios in the 2.4 GHz ISM band. Scalability results verify that BT5 is a feasible technology solution for implementation of IoT applications, whereby the activity level, expected coverage areas and the interference situation within the ISM band needs to be considered implicitly. Currently, real BT5 equipment is analyzed in lab and field trials (see Fig.3), in order to further validate and strengthen already achieved evaluation results. In

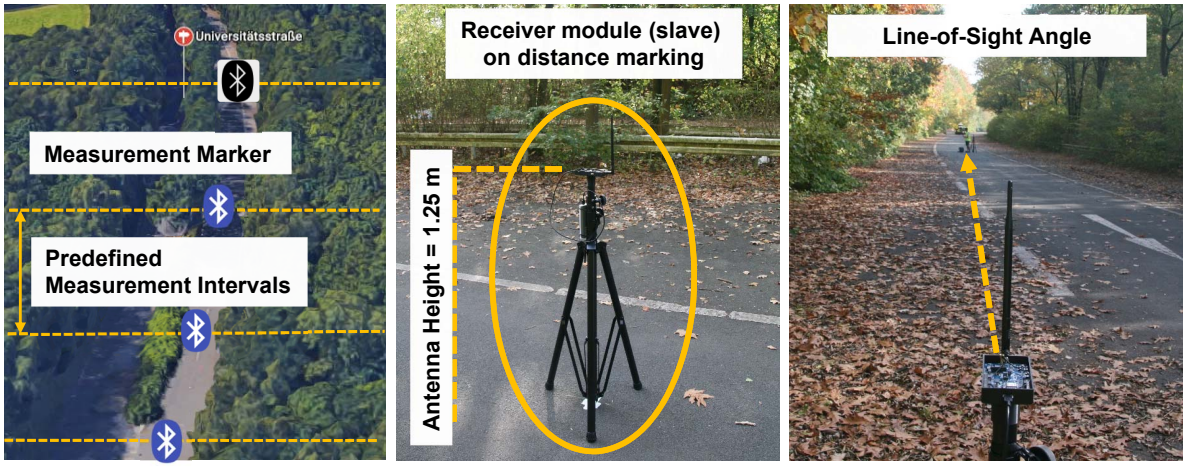


Figure 3: Packet Error Rate (PER) for relevant IoT activity scenarios [1]

future work, we aim at installing a BT5 Smart City Gateway, according to the approach in [2]. This allows to compare BT5 and LoRa technology performance in terms of IoT applicability and leads to a setup of a real large-scale deployment based on BT5 hardware.

References

[1] S. Böcker, C. Arendt, and C. Wietfeld. On the Suitability of Bluetooth 5 for the Internet of Things: Performance and Scalability Analysis. In *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC) - Workshop WS-07 on "The Internet of Things (IoT), the Road Ahead: Applications, Challenges, and Solutions"*, Oct 2017.

[2] P. Jörke, S. Böcker, F. Liedmann, and C. Wietfeld. Urban Channel Models for Smart City IoT-Networks Based on Empirical Measurements of LoRa-links at 433 and 868 MHz. In *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC) - Workshop WS-01 on "Communications for Networked Smart Cities (CorNer)*, Oct 2017.

Crystal Gazing: Passive Data-Rate Prediction in LTE User Equipment

Robert Falkenberg
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
robert.falkenberg@tu-dortmund.de

Choosing the most reliable network or transmitting data at the fastest data rate without actively probing the current links is a challenging task. Based on recent methods which give LTE User Equipment (UE) insight into distinct resource assignments of all other participants in a cell, we captured this data to train a neural network which is capable of reliably predict the expected data rate for a pending transmission over the observed network. With this technique we were able to predict the data rate of a UE with an estimation error below 1.5 Mbit/s in 93 % of cases in a public network.

1 Introduction

Since mobile networks, in particular Long Term Evolution (LTE), are popular candidates for connecting Cyber-physical Systems and or other Internet of Things (IoT) devices. While human network users are typically interested in best effort throughput and a high quality of experience, Machine-Type-Communication (MTC) has different demands to the network like energy-efficiency or latency. For example, MTC might be postponed or redirected to another network/technology under bad radio conditions or network congestion in order to reduce network stress. Another motivation, which arises from MTC under resource constraints, is the reduction of transmission-time to reside longer in power saving mode [3].

With this objective in mind, we focus on passive metrics an predictions, which enable anticipatory MTC in mobile networks. While radio networks typically provide passive indicators for an estimation of the radio signal strength or quality, in general no indicators for a network-load estimation are provided by today's technologies. Although, some use cases, e.g., continuous or frequent transmissions, might provide a direct data rate estimate from throughput measurements of ongoing or recent transmissions, sporadic

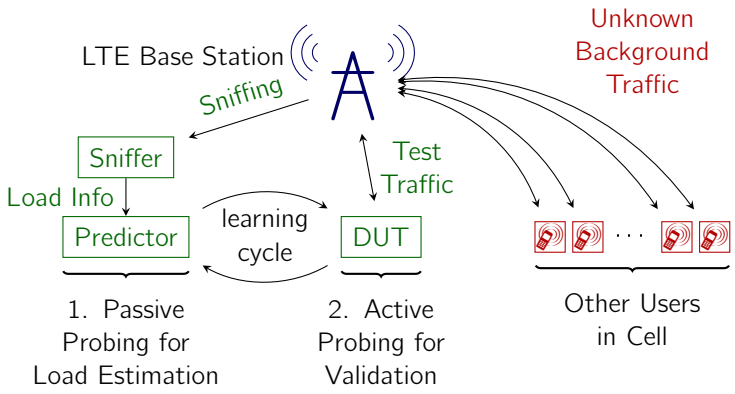


Figure 1: Overview of the scenario and measurement setup for LTE data rate predictions.

MTC cannot rely on this data. Since active probing is impractical for short transmissions, such devices have to take a different path.

Fortunately, radio resources of LTE networks are centrally organized by the base station (evolved NodeB (eNodeB)), which assigns Resource Blocks (RBs) to distinct UE in its coverage area. Although resource allocations are designed to be only decoded by the addressed UE, our Client-based Control Channel Analysis for Connectivity Estimation (C^{3}ACE) [2] proposed a method to make these allocations available on a passive sniffer in range of the eNodeB, hence giving the device insight in the radio-resource distribution of the entire cell and the number of currently active devices. This document describes the Enhanced Client-based Control Channel Analysis for Connectivity Estimation (E-C^{3}ACE) which including machine learning and additional features from the LTE sniffer [1].

2 Approach and Implementation

In order to extend our previous C^{3}ACE approach by additional data from LTE sniffing and utilize machine learning to bring them together to a data rate predictor, we setup the following data acquisition system, as shown in Fig. 1. The setup consists of an eNodeB, and several attached UE which stress the cell with various traffic patterns like full load or low rate background transmissions. In addition, we synchronized our passive sniffer to this cell, which continuously decodes the Physical Downlink Control Channel (PDCCH) and logs the sniffed resource allocations of all active cell-users into a logfile. At the same time, a Device Under Test (DUT) performs sporadic, best-effort transmissions, and measures the actually received data rate as a label for the machine learning process.

Fig. 2 gives a detailed insight into the data processing pipeline for the training process and the subsequent application phase. The first step brings together data from passive probing (sniffed data), which carries distinct resource allocations of other cell users $\vec{f}_s$, and data from active probing (by the DUT), which include channel quality indicators at the DUT and the actually achieved data rate r as the label. Furthermore, we calculated additional features (average and standard deviation) of the following values over a period of 1 s before the actual DUT transmission timestamp:

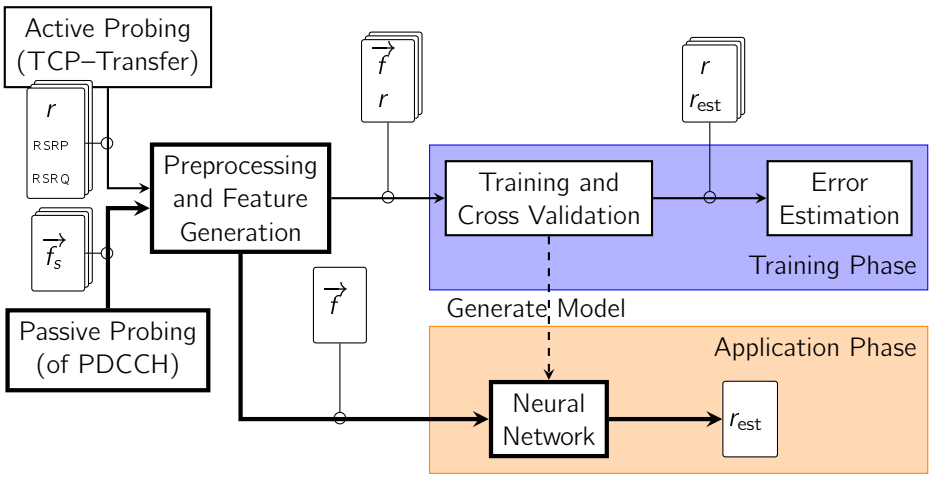


Figure 2: Processing pipeline for training of predictor model and application of the trained model for future predictions (bold components only).

Number of resource blocks: Reflects the current over-all utilization of the cell, whether the cell has spare ressources for additional traffic or already clips the resource demands of distinct participants to the available amount.

Number of active devices: Indicates the degree of competition in the cell. If many devices claim for radio resources, an additional device will receive a much smaller fraction of resources, than if only few devices have request for resources.

Transport Block Size (TBS), Modulation and Coding Scheme (MCS): Indicates the resource efficiency and user distribution in the coverage area. UEs in proximity to the eNodeB can transmit more payload in one resource block than cell-edge devices due to a strong radio signal. High average TBS and MCS indicate a concentration of active UE in the cell center, while low values indicate a distribution towards the cell edge.

Based on these generated and preprocessed features, we trained a neural network with the objective of minimizing the Root Mean Square Error (RMSE) between r and the estimate r_{est}. We achieved the lowest prediction error in cross validation with two-layer networks consisting of 10 neurons in the first layer and 5 neurons in the second layer.

3 Results

The results of the E-C^{3}ACE prediction accuracy on data from our laboratory setup are shown in Fig. 3 as green graphs. As a reference, we also included predictions from our previous C^{3}ACE approach (red), which rely only on the number of active devices in a cell and a reference value from an empty cell. The blue curves reflect the boundary of the C^{3}ACE approach by calculating the reference value retrospectively to minimize the RMSE over all captured samples.

The left figure shows the major prediction improvement of a machine learning model compared to the simple model. Although the true data rate (black) coarsely follows the $\frac{1}{n}$ relation from the simple model, it is much more precisely estimated by the neural network.

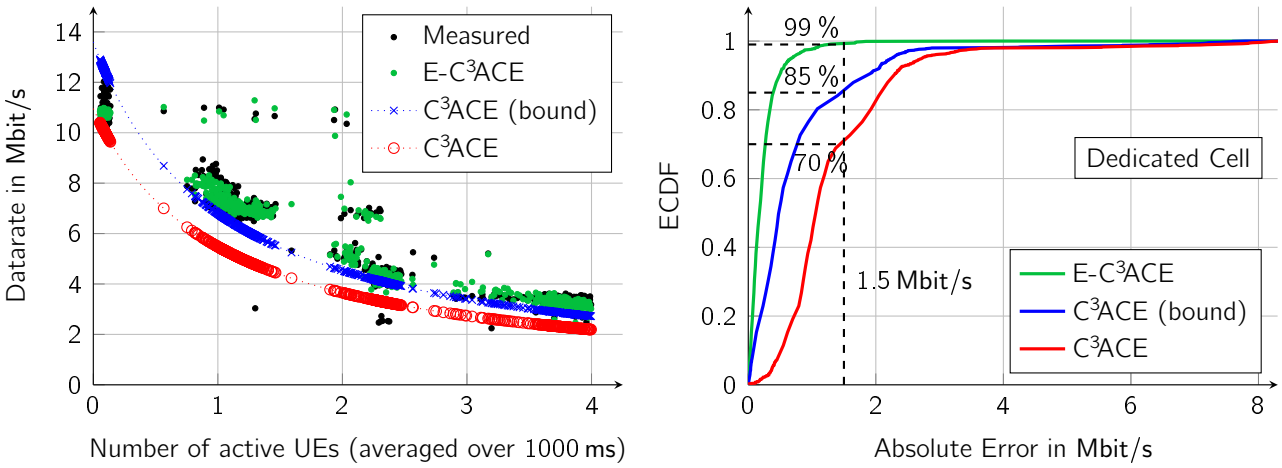


Figure 3: Date rate predictions in laboratory by C³ACE and E-C³ACE according to the number of active devices in the LTE cell. E-C³ACE (green) hits the real data rate (black) precisely and achieves in 99 % a prediction error below 1.5 Mbit/s.

Especially the scattering towards higher data rates is reflected by the trained model. This prediction improvement also manifests in a much smaller RMSE, as shown in the right figure as Empirical Cumulative Distribution Function (ECDF). The trained model clearly outperforms the simple approach and predicts in 99 % an error below 1.5 Mbit/s.

With this promising approach, we also performed field measurements in the public LTE network and still achieved in 93 % an error below 1.5 Mbit/s. Due to space restrictions, we kindly refer the interested reader to [1] for further details on this topic.

4 Conclusion and Further Research

Since we proved the functionality of the proposed data rate prediction method even in public LTE cells, we will further improve the prediction accuracy by application of recent methods for model based optimization (project B3) and deep learning (project A1) on larger datasets. In addition, we will integrate this approach in a live system to evaluate the energy- and resource-saving potential of this method to MTC for mobile vehicular applications in collaboration with project B4.

References

[1] Robert Falkenberg, Karsten Heimann, and Christian Wietfeld. Discover your competition in LTE: Client-based passive data rate prediction by machine learning. In *IEEE Globecom*, Singapore, dec 2017.

[2] Robert Falkenberg, Christoph Ide, and Christian Wietfeld. Client-based control channel analysis for connectivity estimation in LTE networks. In *IEEE Vehicular Technology Conference (VTC-Fall)*, Montréal, Canada, sep 2016. IEEE.

[3] Robert Falkenberg, Benjamin Sliwa, and Christian Wietfeld. Rushing full speed with LTE-Advanced is economical - a power consumption analysis. In *IEEE Vehicular Technology Conference (VTC-Spring)*, Sydney, Australia, jun 2017.

Utilizing Network Coding for Scalable Heterogeneous Link Aggregation

Karsten Heimann
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
karsten.heimann@tu-dortmund.de

Aggregation of multiple carriers and heterogeneous links in wireless networks as well as utilizing network coding in mesh networks are proven efficient and robust mechanisms to boost data rates. Hereby we present our new method *ScaIaNC* to aggregate heterogeneous links by incorporating network coding to adjust a balance of capacity, latency and data confidentiality. Through a parametrization, *ScaIaNC* can therefore be adapted to the user's needs. The performance of *ScaIaNC* is validated by an experimental setup showing the polarity of those different needs.

1 Introduction

Today, mobile devices often utilize multiple communication technologies for services such as cloud systems. Network coding is able to increase the packet throughput of those transfers by occupying the communication links more efficiently. Beside potentially harsh transmission channel conditions, mobile devices suffer from various resource constraints like battery lifetime or time critical applications. Furthermore, usage scenarios like emergency services ordinarily require secure and reliable access some backbone network.

Addressing those constraints, we developed the link aggregation protocol *ScaIaNC*, which loads heterogeneous links in a scalable fashion by means of network coding [1]. By doing so, an exploitation of each communication link capacity enhances the overall user experience in terms of transfer speed, real–time capability and security of confidential data. Therefore, multiple links are managed in parallel applying network coding on top to offer a scalability between maximum performance and maximum reliability.

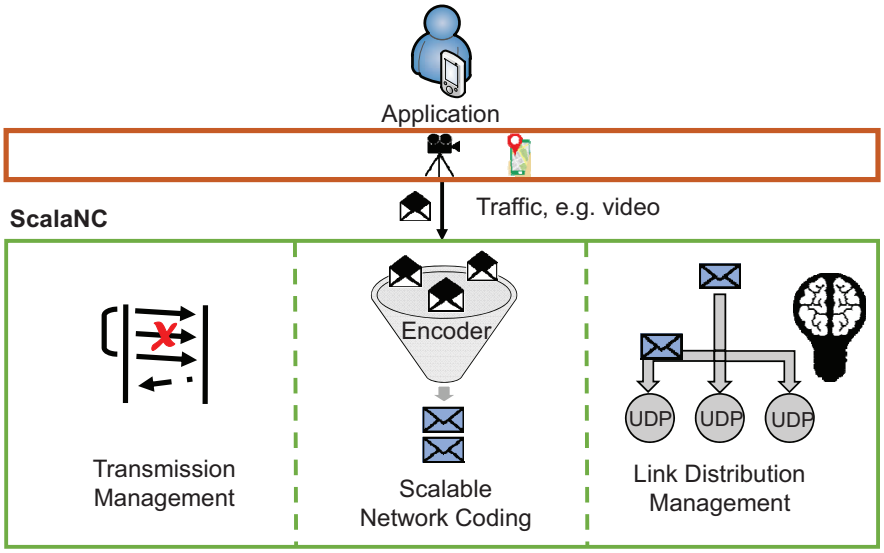


Figure 1: *ScalaNC* Overview

2 Scalable Link Aggregation enabled by Network Coding

The Scalable Link Aggregation enabled by Network Coding (*ScalaNC*) is located below the user application as depicted in Figure 1 and performs three different tasks:

First the user traffic (e.g. some file transfer or a video stream) is encoded by means of network coding. A predefined amount of packets, the generation, is network coded from those user data. To provide a more reliable transmission, some additional packets can be produced a priori to spend a forward error correction (FEC) functionality. However, further packets can also be derived posterior if packet errors necessitate retransmissions.

On the other hand our approach sets up UDP connections for each communication path and handles the transmission management. Although UDP is a connectionless and thereby unreliable protocol, this transmission management introduces some error handling and acknowledgement mechanism to cope with the demand for reliablity regarding file transfers for example.

The third part of *ScalaNC* is the link distribution management. We have designed a scalable method for splitting the produced packets dynamically onto the multiple connections. In a first step the user can choose between some predefined steps between an even distribution of the packets to all links and the highest achivable throughput. As an outlook this selection could be automated and dynamically fitted to the needs of the user in future. The even distribution comes with the highest security in terms of data confidentiality, because the data is spread uniformly to the different connections. Eavesdropping attacks on only a subset of links may only pick up some useless since encoded packets, so that the user data stay cryptically.

3 Experimental Setup and Performance Evaluation

Figure 2 shows our experimental setup consisting of the cloud system, which is connected via multiple interfaces and a network emulation to a mobile node. By means of the network emulation, capacity and packet error rate as path characteristics can be parameterized for each individual link. Those configurations are kept constant and the capacities of all links are known to both endpoints.

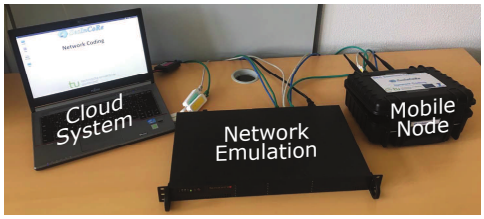


Figure 2: Our Experimental Setup for Multiple Communication Paths

To evaluate the performance of *ScalaNC*, we initially consider the parametrization of the FEC and the Link Distribution Management. Therefore, the experimental setup is configured for three heterogeneous communication paths with the characteristics as pointed in Table 1.

In Figure 3(a) goodput and latency are plotted against the FEC as the amount of additional redundancy. Although the goodput decreases with rising redundancy, the effect of redundant packets counteracting the PER becomes clear at a FEC of 100 %: Whereas naively a halving of the goodput to 39 Mbit/s is expected due to sending redundant information during the half of the time, a goodput of 44 Mbit/s was achieved. Inserting redundancy leads to a reduction of the latency as retransmissions can be obviated. In this setting a saturation is reached for a FEC amount of at least 50 %.

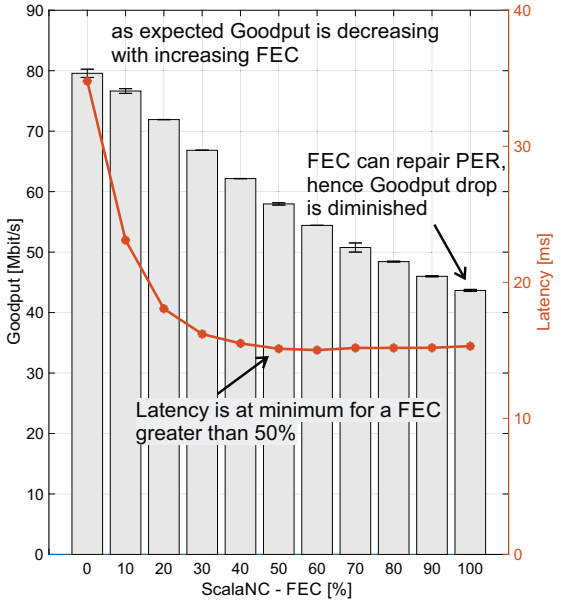
The link distribution management is analyzed in Figure 3(b) (without FEC). While a uniform load distribution limits the load of each path to the lowest available link capacity ($3 \cdot 5$ Mbit/s), it preserves the highest information distribution index (IDI). The IDI is derived from the established Theil–Index [2] to scale it to values from $[\frac{1}{N}, 1]$ with 1 indicating a evenly distribution and $\frac{1}{N}$ would rely on only one link:

$$\mathrm{IDI} = \exp(-\underbrace{\frac{1}{N}\sum_{i=1}^{N}\frac{x_i}{\mu}\ln\frac{x_i}{\mu}}_{\text{Theil–Index}}), \qquad \text{with } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

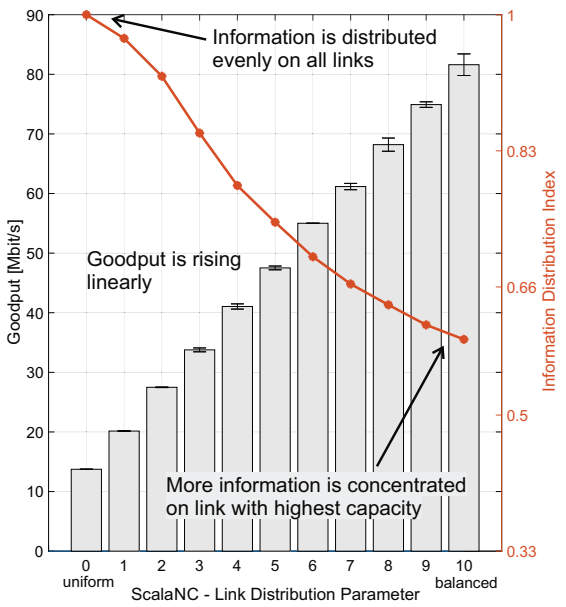
By easing the limitation of an evenly distribution, heterogeneous paths are aggregated more effective, which leads to a higher goodput to the downside of the IDI. If all links are fully utilized, the IDI amounts to 0.6.

Evaluation Setup	FEC			Link Distribution		
Capacity in Mbit/s	60	20	10	75	10	5
PER	15 %	10 %	5 %	5 %	5 %	5 %

Table 1: Parameter Sets for three Paths



(a) FEC vs. Goodput and latency



(b) Link Load vs. Goodput and IDI

Figure 3: Parameter Evaluation of FEC and Link Distribution

4 Conclusion and Further Research

Our network coding based, scalable link aggregation protocol *ScalaNC* promises robustness against packet errors as well as data confidentiality. It is especially suited to enhance communication performance at harsh channel conditions as in rural areas or highly mobile contexts.

Moreover, we plan to implement an automatically adaption to the channel conditions and also to consider different network coding variants to further reduce the latency.

References

[1] Daniel Behnke, Matthias Priebe, Sebastian Rohde, Karsten Heimann, and Christian Wietfeld. ScalaNC — scalable heterogeneous link aggregation enabled by network coding. In *13th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2017) - Fourth International Workshop on Emergency Networks for Public Protection and Disaster Relief (EN4PPDR'17)*, Oct 2017.

[2] H. Theil. The information approach to demand analysis. *Econometrica*, 33(1):67–87, 1965.

Empirical Measurements of LoRa-links at 433 and 868 MHz for Urban Channel Models in Smart City IoT Networks

Pascal Jörke
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
pascal.joerke@tu-dortmund.de

With the upcoming challenges of low power consumption, high communication ranges and large scalability in the Internet of Things (IoT), new technologies are evolved. These Low Power Wide Area Network (LPWAN) technologies will enable the digitalization of everyday's life. Though, city environments place great challenges on the LPWAN systems, caused by dense and large buildings leading to high path losses. Technologies like LoRa provide high communication ranges by using the 433 MHz ISM and 868 MHz SRD band for low attenuation and high receiver sensitivities to receive signals even with very low power. In our work we present urban channel models for Smart City IoT-Networks based on empirical measurements of LoRa-links at 433 MHz and 868 MHz.

1 Introduction

On the road to Smart Cities, existing communication technologies are faced with new challenges. Smart meters being placed in basements, smart sensors running on batteries and a huge amount of upcoming smart devices substantiate the need of high ranging and energy efficient communication technologies with great scalability. Therefore, new technologies have emerged, facing the rising Internet of Things (IoT) with billions of devices being developed digitally, enabling the establishment of suitable communication networks quickly, simply and affordable.

With LoRa an upcoming Low Power Wide Area Network (LPWAN) technology has been developed, promising long lifetime of battery powered end devices, low cost installation and high ranges with data rates between 0.3 kbit/s and 11 kbit/s. Though the

actual range depends on the data rate and frequency used as well as on the propagation conditions found on the installation site.

2 Approach and Implementation

For an efficient network planning of Smart City communication systems the knowledge of propagation characteristics is fundamental. Intelligent network cell placements will lead to reduced costs of a Smart City rollout, using a low number of LPWAN gateways. To evaluate the path loss characteristics of Smart City environments, an avaliability analysis of the LoRa technology is performed measuring the signal strength in the city of Dortmund, Germany, inhabited by 600.000 people on an area of 280 km^2, which leads to an average inhabitants density of 2143 people per qm^2. Therefore, the real world range analysis is based on a well urbanized area and thus a high attenuation is expected. Reflecting the measurement samples of this availability analysis, the suitability of established empirical path loss models such as Okumura Hata, ITU Advanced, Winner+ and 3GPP Spatial Channel Model for narrow band LoRa signals will be examined.

Most LoRa installations are using the 868 MHz band in Europe, which has significantly better propagation characteristics compared to the 2.4 GHz ISM band. Additionally, LoRa systems are also available for the 433 MHz ISM band, which further reduces path loss, but comes with a reduced maximal equivalent isotropically radiated power (EIRP) of 10 dBm in contrast to 14 dBm using 868 MHz. Our work examines both 433 MHz and 868 MHz systems, analyzing which system performs better. Therefore, LoRa gateways for both 433 MHz and 868 MHz are placed on the roof of a building in 30 m height on the campus of TU Dortmund University (Fig. 1).

Figure 1: Installation site of Smart City LoRa Gateway antennas in 30 m height.

3 Measurements and Results

The measurement consists of multiple runs using a car as a node moving through the network for the availability measurement. The LoRa nodes installed on the rooftop of the car receive pings which are periodic sent by the gateways and store the RSSI (Received Signal Strength Indicator), SNR (Signal to Noise Ratio) as well as the current GPS position of the car. These measurement samples then were used to examine the relation between distance and path loss of LoRa signals. Fig. 2 and 3 show the results of the

samples and the derived Dortmund path loss models as well as established path loss models.

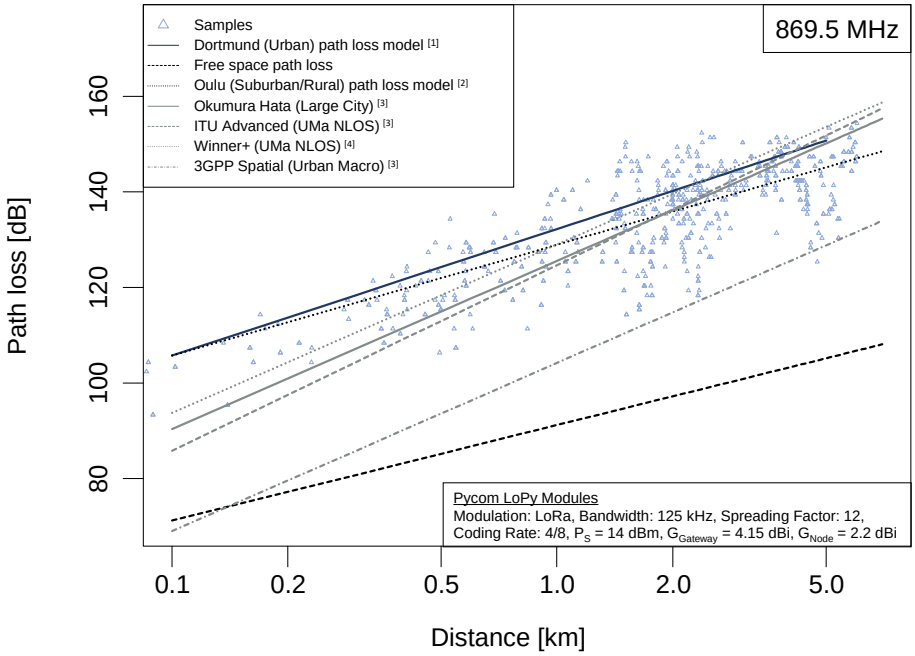


Figure 2: Comparison of path loss models with Dortmund (Urban) path loss model for 868 MHz Smart City Scenario in Dortmund, Germany [3] [1] [2] [4].

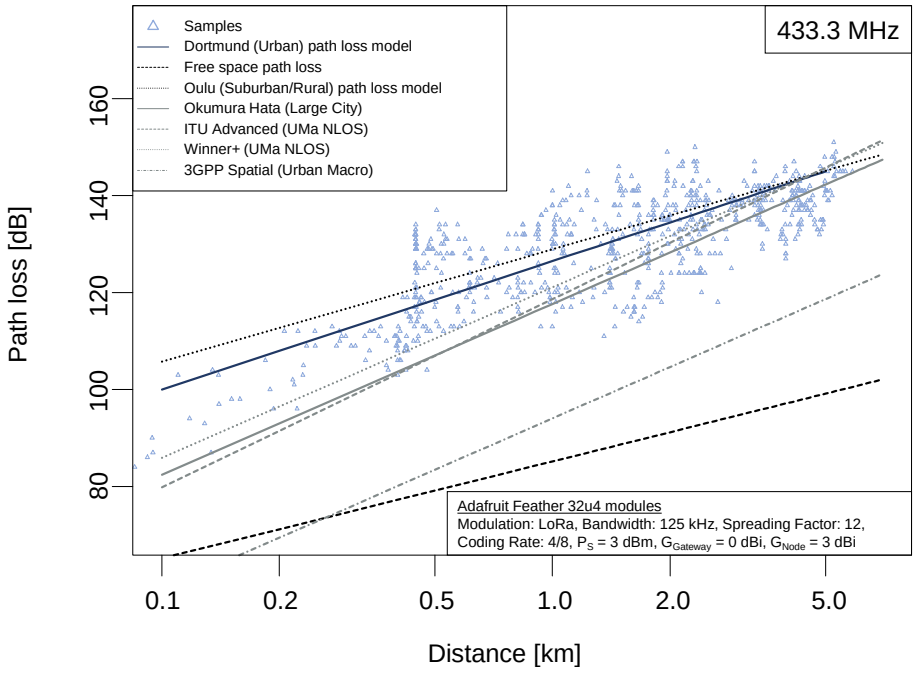


Figure 3: Comparison of path loss models with Dortmund (Urban) path loss model for 433 MHz Smart City Scenario in Dortmund, Germany [3] [1] [2] [4].

4 Conclusion

The comparison of the established path loss models and the measured path loss characteristics show that the considered path loss models are insufficiently predicting the expected path loss. Therefore new path loss models for 868 MHz and 433 MHz LoRa Smart City scenarios are proposed based on the extensive LoRa available measurements, given by these equations [3]:

$$PL_{LoRa868MHz} = 132.25 + 26.5 \cdot \log 10(d[km]) \quad (1)$$

$$PL_{LoRa433MHz} = 128.25 + 26.5 \cdot \log 10(d[km]) \quad (2)$$

5 Further Research

Our recent work was based on examining communication ranges of Low Power Wide Area Network technologies such as LoRa in Smart City environments. Further challenges of IoT scenarios include low power consumption for battery powered sensor applications like Smart Waste Management, where all garbage cans in the city inform the waste collection services, if they need to be emptied, enabling a better planning of services. Therefore, cellular LPWAN technologies like LTE Cat-M1 and LTE Cat-NB1 have been developed to enable battery powered devices running up to 10 years on a single battery by considering new power saving mechanisms like eDRX (extended Discontinuous Reception), which extends the time span between receptions windows, and PSM (Power Saving Mode), which enables the page monitoring between data transfers without losing the network registration. Our future research will focus on a detailed power consumption modeling of cellular LPWAN technologies such as LTE Cat-M1 and LTE Cat-NB1, determining how application, network and environmental parameters affect the power consumption of these sensors, predicting realistic battery lifetime for Smart City applications.

References

[1] M. Hata. Empirical formula for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology*, 29(3):317–325, Aug 1980.

[2] Andreas F Molisch. *Wireless communications*, volume 34. John Wiley & Sons, 2012.

[3] P. Jörke, S. Böcker, F. Liedmann, C. Wietfeld. Urban Channel Models for Smart City IoT-Networks Based on Empirical Measurements of LoRa-links at 433 and 868 MHz. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Oct 2017.

[4] J. Petajajarvi, K. Mikhaylov, A. Roivainen, T. Hanninen, and M. Pettissalo. On the coverage of LPWANs: range evaluation and channel attenuation model for LoRa technology. In *14th International Conference on ITS Telecommunications (ITST)*, pages 55–59, Dec 2015.

Data Collection for Modelling Power Chargers in Energy Harvesting

Mojtaba Masoudinejad
Lehrstuhl für Förder- und Lagerwesen
Technische Universität Dortmund
mojtaba.masoudinejad@tu-dortmund.de

An energy harvesting (EH) module is able to empower an embedded device and increase its life span. Such a system requires a charging module in addition to the harvester and energy storage. However, adding this structure increases the design complexity and requires detailed modelling. This paper provides a data collection and analysis process for state of the art charger devices. This collection enables data-based modelling of chargers.

1 Introduction

Non-stationary embedded devices will establish the future computing [4]. In spite of growing number of these devices [5], their power supply is still a challenging implementation aspect. Although increasing the storage size helps some applications, others such as an intelligent warehouse [1], cannot rely only on batteries. Logistics application [2] require device deployment in large scale, long life span with no maintenance costs.

Using EH expands devices' life cycle. However, it adds complexities to the hardware design. From one side dynamic voltage of the harvester has to be continuously matched to the battery's voltage. on the other hand, a control system has to keep both operational points near the optimum [3]. A solution including all these functionalities is mostly called a charger, which its overview is shown in Fig.1. Developing such charger with of-the-shelf components is possible. However, energy losses forces most implementations to use monolithic solutions. Currently, there are three chips from the BQ255XX series made by Texas Instruments (TI) and SPV1050 from ST available for this purpose. TI's BQ25505 and BQ25570 chips promise higher efficiency and have dominated the market.

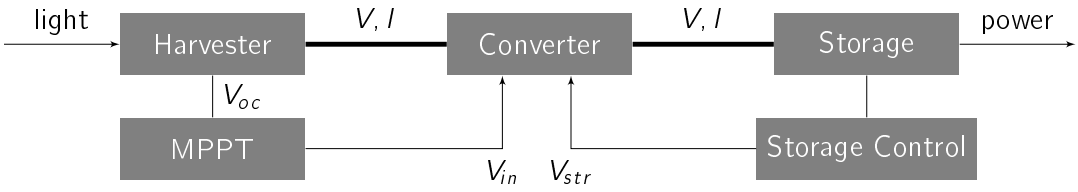


Figure 1: Complete power supply chain for a PV EH powered device

Any large-scale deployment necessitates a detailed power analysis of the whole system. Designer has to balance harvested energy with system's load using modelling and simulation. However, available knowledge from monolithic chargers is limited. Consequently, black-box modelling principle would be easier to avoid internal complexities. These techniques require only system's input/output at different operational conditions.

2 State of the Art

Reliable data collection requires understanding of possible operational conditions of the system. The overall operational state machine of these chips is presented in Fig. 2 [3] In a balanced system this device would be mostly working in the *NO* state. Therefore, data collection in this state has the highest priority.

From transitions in Fig.2 can be seen that the storage voltage (V_{str}) has to be larger than 1.8 V and smaller than the over-voltage limit which is 4.2 V for TI's evaluation boards. These boards are used to assure same design and setup as proposed by manufacturer.

3 Measurement

Two Source Measurement Units (SMU) from *Keysight* are connected to evaluation boards, responsible to emulate the harvester and storage. Measurement precision of both devices is 10 fA and 100 nV. One channel acting as a current source replicates a

CS:low storage voltage to run the MPPT
NO: MPPT and converter run normally
OV: feed to the storage is disabled

$C_1 : V_{str} < 1.8\,\mathrm{V}$

$C_2 : 1.8\,\mathrm{V} < V_{str} < V_{ov}$

$C_3 : V_{ov} < V_{str}$

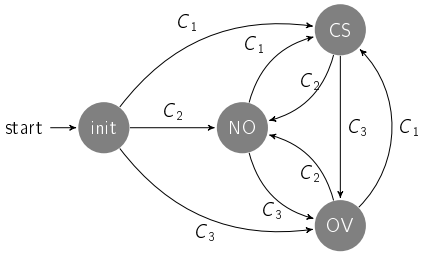


Figure 2: A simplified state machine representing charger's operational states

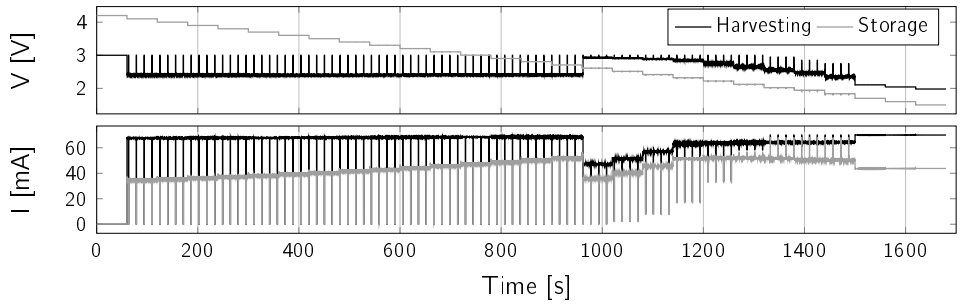


Figure 3: signals measured from TI BQ25505 at I_{in} =70 mA and V_{oc} =3.5 V

photovoltaic harvester with its open circuit voltage as voltage compliance. Simultaneously, another channel acts as a voltage source representing the battery. Voltage and current of both channels is measured concurrently.

For a proper modelling, two sets of independent data are required. One set will be used for developing the model, While the other set is used for its evaluation.

3.1 Modelling experiments

For collecting modelling data, input parameters (including I_{EH} and V_{oc}) are constant while V_{str} changes. V_{str} begins from 4.2 V and reduces to 1.5 V. An example of such data is presented in Fig. 3.

To cover all possible operational points, this process is automated. SMU is connected to a PC running a MATLAB program. This program communicates with the device using SCPI commands. It initializes the system, sets parameter values and acquires collected data from the device. Using this process, 540 datasets for each charger are collected.

3.2 Evaluation experiment

In addition to the cross validation during modelling, a secondary dataset is required here for validation. For this propose, a secondary experiment is proposed. In contrast to the modelling experiment with a static working condition during each time period, the evaluation experiment includes dynamics and is able to cover more operational points. During this experiment the I_{EH} changes over time as the variable parameter. This sinusoid signal starts at 50 mA, increases to 100 mA; then, reduces to zero and again to the initial value of 50 mA. This form uncovers possible hidden memory effects which causes different behaviours according to the signal derivation.

While input current changes during the experiment, V_{oc} and V_{str} are constant. An example of collected data for such experiment is shown in Fig. 4. The MPPT measurement periods

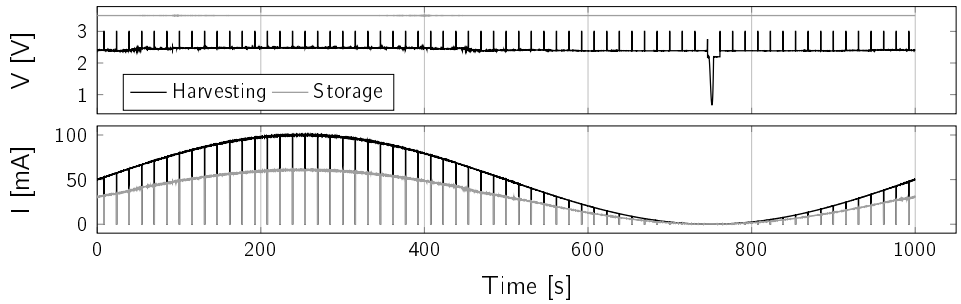


Figure 4: Example PSS signals measured from TI BQ25570 with a sinusoidal input current, 3 V open circuit voltage and a constant 3.5 V storage voltage

can be seen in these data as well. Using the automated collection process, 280 set of evaluation datasets are collected for each charger as well.

4 Sum up

This report addressed necessity of modelling energy harvesting based power supply. Two state-of-the-art monolithic power charger for this purpose are presented. Specification of two different data collection scenarios for modelling them is proposed.

References

[1] R. Falkenberg, M. Masoudinejad, M. Buschhoff, A. K. R. Venkatapathy, D. Friesel, M. ten Hompel, O. Spinczyk, and C. Wietfeld. PhyNetLab: An IoT-based warehouse testbed. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1051–1055, September 2017.

[2] Michele Magno, Fabien Ferrero, and Vedran Bilas. *Sensor Systems and Software: 7th International Conference, S-Cube 2016, Sophia Antipolis, Nice, France, December 1-2, 2016, Revised Selected Papers*. Springer, August 2017.

[3] Mojtaba Masoudinejad. Automated Data Collection for Modelling Texas Instruments Ultra Low-Power Chargers. Technical Report, TU Dortmund, Dortmund, 2017.

[4] Mojtaba Masoudinejad. A Power Model for DC-DC Boost Converters Operating in PFM Mode. Technical report, TU Dortmund University, 2017.

[5] J. Venkatesh, B. Aksanli, C. S. Chan, A. S. Akyürek, and T. S. Rosing. Scalable-Application Design for the IoT. *IEEE Software*, 34(1):62–70, January 2017.

Wireless Positioning based on IEEE 802.15.4a Ultra-Wideband for Logistic Applications

Janis Tiemann
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
janis.tiemann@tu-dortmund.de

Wireless positioning based on IEEE 802.15.4a gained attention for precise localization recently. However, for large scale logistic applications many aspects are not considered. This work aims to focus on optimizing the positioning accuracy, multi-user scalability and energy efficiency of such systems. Due to the simple nature of the most commonly used two-way ranging based systems, a significant amount of messages is required to obtain a position in a set of infrastructure anchors. This work proposes, implements and analyzes time-difference of arrival based positioning to improve the system capabilities to a level that supports autonomous robotic systems in logistic scenarios.

1 Introduction

Recent developments in ultra-wideband (UWB) wireless communication hardware enabled a wide range of research. The integration of precise time of arrival (TOA) estimation into commercially available low-cost receivers yields a variety of interesting opportunities for the research community. Our specific goal is to enable novel applications by the scalable and accurate integration of such systems into logistic scenarios. Previous work [1] enabled early integration with unmanned aerial vehicle (UAV) systems by the emulation of commonly used global navigation satellite system (GNSS) receivers. In other work [2] we could prove the applicability of UWB positioning in vehicular wireless power transfer scenarios. However, the aforementioned work uses two-way ranging based approaches

such as symmetric double-sided two-way ranging (SDS-TWR) to estimate the distance and hence the position of the mobile nodes to static anchor nodes. Since two-way ranging requires sequential message exchange of the participants, the scalability of such approaches is severely limited.

Therefore, our goal is to improve the multi-user scalability and overall system accuracy by using infrastructure based time-difference of arrival (TDOA) positioning. We evaluate the multi-user scalability and propose a method for high precision wireless clock synchronization, see [3]. In a second step we provided the core-components of our system as open-source to the research community, see [4]. In order to comparably analyze the systems capabilities, we participated in Track 4 of the annual EvAAL competition at the Seventh International Conference on Indoor Positioning and Indoor Navigation (IPIN2016) in Alcalá de Henares, Madrid, Spain. In the course of this work we prove the capabilities of our system design by putting it into challenging real-life applications.

2 Experimental Accuracy and Applicability Analysis

Our ultimate goal is enabling novel applications by the use of high precision wireless localization. One application that is close at hand is autonomous UAV indoor navigation for low-cost stocktaking and inventory keeping, see Fig. 1. In [5] we show that such a task can be achieved with our TDOA-based UWB localization system. We achieved a stable control loop using the UWB data as a accurate basis for the position, see Fig. 3. However, in order to enable productive usage of wireless indoor localization accuracy is not the only requirement.

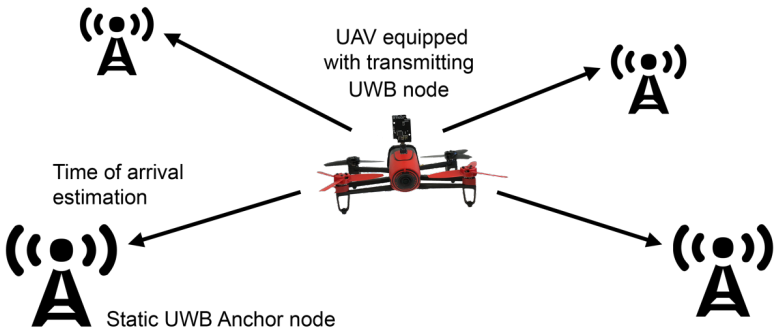
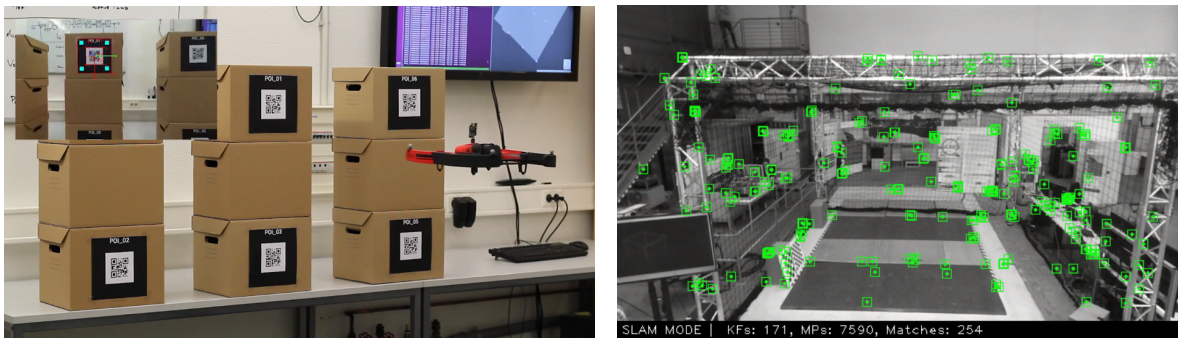


Figure 1: Illustration of the basic concept of TDOA-based UWB positioning on UAVs.

Another important aspect is the scalability in terms of multi-user access and the real-time requirements in terms of guaranteed update rate and propagation delay. The results showed that our UWB-based approach fulfills those requirements and we provide a proof

of concept video [6] illustrating an indoor UWB-based multi-UAV flight and stocktaking applications such as depicted in Fig. 2(a). The paper associated with those experimental results was recently awarded with the Best Paper Award at the eighth International Conference on Indoor Positioning and Indoor Navigation (IPIN) 2017 in Sapporo, Japan.



(a) Stocktaking application using QR-codes (b) SLAM-based flight outside UWB coverage

Figure 2: Stills of the proof of concept experiment videos [6].

To further extend the UWB-based UAV navigation capabilities, an extension based on monocular simultaneous localization and mapping (SLAM) is evaluated in the ongoing research. We are able to show that UAV navigation outside of the coverage of wireless localization systems is possible by augmenting the position information with additional data from the front-facing camera, see Fig. 2(b). The main challenge is to reference and scale the monocular SLAM data in a way that it can be used for navigation feedback.

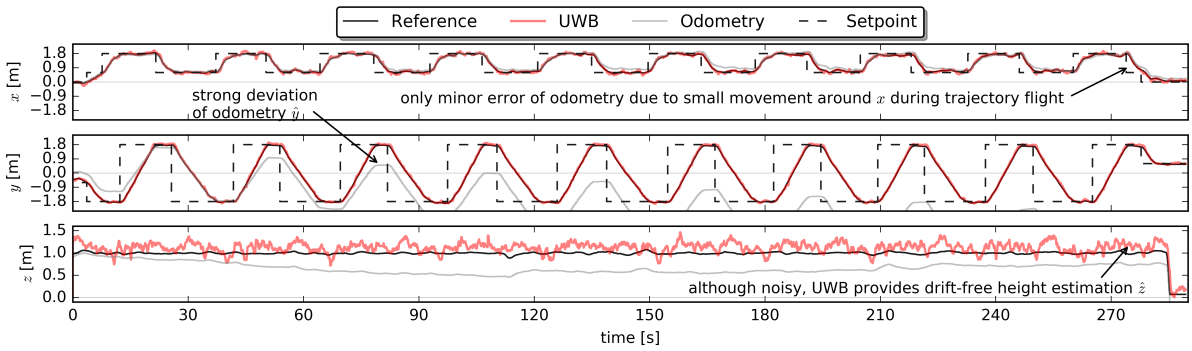


Figure 3: Time-series of long-term stability experiment for UWB-based UAV control.

Modern UWB transceivers yield advanced quality assessment parameters such as the raw channel response data that go way beyond the typically used received signal strength indication (RSSI). In an effort to improve the ranging accuracy, we analyzed this data and trained a neural network to estimate the distance error based on the channel response data, see [7]. Preliminary results show, that in known environments the channel response can yield valuable information that can improve the accuracy significantly.

3 Conclusion and Further Research

In this work, a scalable positioning method for IEEE 802.15.4a based UWB transceivers was proposed and evaluated. We could prove the advanced capabilities of the systems through practical applications using precise and scalable UWB-based UAV control. Future work will focus on improving the real-time capabilities through intelligent and energy-efficient multi-user scheduling to guarantee user-defined update rates.

References

[1] J. Tiemann, F. Schweikowski, and C. Wietfeld. Design of an UWB indoor-positioning system for UAV navigation in GNSS-denied environments. In *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, Oct 2015.

[2] J. Tiemann, J. Pillmann, S. Böcker, and C. Wietfeld. Ultra-wideband aided precision parking for wireless power transfer to electric vehicles in real life scenarios. In *IEEE Vehicular Technology Conference (VTC-Fall)*, Montréal, Canada, Sep 2016.

[3] J. Tiemann, F. Eckermann, S. Böcker, and C. Wietfeld. Multi-user interference and wireless clock syncronization in TDOA-based UWB localization. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Alcalá de Henares, Madrid, Spain, Oct 2016.

[4] J. Tiemann, F. Eckermann, S. Böcker, and C. Wietfeld. ATLAS - an open-source TDOA-based ultra-wideband localization system. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Alcalá de Henares, Madrid, Spain, Oct 2016.

[5] J. Tiemann and C. Wietfeld. Scalable and precise multi-UAV indoor navigation using TDOA-based UWB localization. In *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Sapporo, Japan, (Awarded Best Paper) Sep 2017.

[6] J. Tiemann. Scalable and precise multi-UAV indoor navigation using TDOA-based UWB localization, video: https://vimeo.com/205754093. Sep 2017.

[7] J. Tiemann, J. Pillmann, and C. Wietfeld. Ultra-wideband antenna-induced error prediction using deep learning on channel response data. In *IEEE Vehicular Technology Conference (VTC-Spring)*, Sydney, Australia, Jun 2017.

A Testbed of Heterogenous Systems for Industry 4.0

Aswin Karthik Ramachandran Venkatapathy
Lehrstuhl für Förder- und Lagerwesen
Technische Universität Dortmund
aswinkarthik.ramachandran@tu-dortmund.de

Industry 4.0 communication scenario for heterogenous systems and the networks are presented in this technical report. Due to an increase in heterogeneous systems, the networking between these systems have risen to be very important in a intra-logistics context. A research facility for intra-logistics research is presented where various systems with use cases are deployed in a research hall. This technical report describes the design and deployment strategies of research centre with the possibilities and the design insights for a futuristic Industry 4.0 material handling facility.

1 Introduction

With Industry 4.0, the technologies, services and methods used in industrial production and logistics are changing. Dynamic, real-time and self-organising value-added networks emerge, based on the availability of the relevant information in real-time through the networking of all parties involved in the entire value-added process and the interconnection of objects, humans and systems [2]. The current "technology push" in the design and introduction of autonomous CPS-based production systems, advancing digitisation, and automation lead to the development of new forms of services and work organisation [3]. This new forms of services with collaborative machines can be simulated to an extent but can only be understood deeply with a level of trust for adoption into industry when demonstrators are developed. The requirements for such a research centre is discussed with the scientific objectives followed by the description of all the systems that are deployed in the research centre with information about their flexibility and interoperability [6].

Figure 1: Structural insight into the research centre [6].

2 Scientific Objectives

Decentralisation of the heterogenous systems, methods for networking them and localising the deployed systems form the three basic pillars of an efficient hybrid work environment. With these three pillars a lot of other virtual elements can be included in the research focus such as creating a digital twin [1] [6]. This novel form of an interdisciplinary and cross-process research environment is intended to contribute, amongst others, to answer the following central scientific questions:

- Designing and organising advance human machine interfaces in a secure and purposeful way through the emergence of hybrid cooperation networks (HCN), in which humans and machines support and complement each other in a common work space [6].
- The innate abilities of employees (such as creativity, motor skills, experience, intuition) should be optimally combined with the abilities of technical (assistance-) systems (e.g. for data evaluation or information visualisation) [6].
- Technical systems should not only perceive but also analyse and evaluate their environment more intelligently, e.g. by means of big-data analysis or machine intelligence techniques [6].
- Emerging data volume should be transmitted securely, wirelessly and in a system guaranteed time with the increasing networking and decentralised control of entities in a HCN [6].

For performing research, to develop demonstrators and experimental setups the research centre is furnished with flexible reference and experiment systems and their interoperabil-

ity which is described in the following chapter. A 3D render of the research is shown in fig 1.

3 Heterogeneous Systems

There is optical reference system available in the hall which is also a real-time localisation system (RTLS). The RTLS provides millimetre precision of objects tracked in the hall. They communicate using TCP/IP using a streaming SDK provided by the manufacturer of the camera system (Vicon Cameras). There is a radio reference system, which comprises of 6 N210 Ettus research software defined radio (SDR). The 6 systems are networked within each other and also connected in the same network as the vicon camera API. This provides the possibility to connect these two systems. Another reference system is the laser project system which is mounted on the ceiling of the hall which can also plot points in the hall with millimetre position. The communication is established using an ARPNET protocol to directly render images from the laser sub system. There are 8 of these laser systems that can plot more than 60000 points per second. This provides a very good visual representation of the physical and virtual elements in the hall. The experiment systems are the robot and drone systems which can be programmed with communication interfaces with the reference system. The PhyNetLab is also integrate into this research hall providing multi-modal modelling of communication parameters [5] [4] [6]. Added to PhyNetLab, to increase the complexity of the testbed and also to create another kind of radio communication possibility, the floor is embedded with more than 400 CC1350 STK sensor. These devices can be freely programmed and the communication happens through a specially engineered cable through the shaft laid under the wooden floor. The devices can be flashed with the cable and also wired communication is possible. They are connected to a sync which is also connected to the same network. This enables the whole heterogenous system to communicate between each other [6].

4 Conclusion and Future Works

Industry 4.0 and IoT creates new forms of interaction of humans and machines. Based on CPS, a socio-technical work environment is created, in which humans and machines are in dialogue with each other and complete tasks together. The research centre will focus on creating decentralised logistics systems with emphasis on real-time localisation and navigation algorithms with heterogenous systems and network as a pivotal position for solving these problems [6].

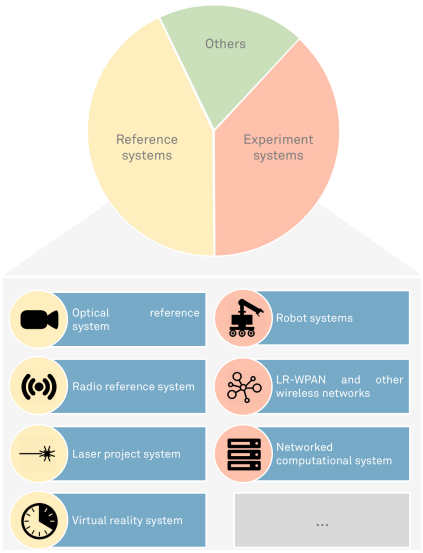


Figure 2: Inter-operable systems at the Research Centre [6].

References

[1] Adaption Intelligence of Factories in a Dynamic and Complex Environment.

[2] Peter Ittermann and Jonathan Niehaus. Industrie 4.0 und Wandel von Industriearbeit ueberblick ueber Forschungsstand und Trendbestimmungen. In Hartmut Hirsch-Kreinsen, Peter Ittermann, and Jonathan Niehaus, editors, *Digitalisierung industrieller Arbeit*, pages 32–53. Nomos, 2015. DOI: 10.5771/9783845263205-32.

[3] Nils Luft. *Aufgabenbasierte Flexibilitaetsbewertung von Produktionssystemen*. Praxiswissen Service, 2013.

[4] A. K Ramachandran Venkatapathy, A. Riesner, M. Roidl, J. Emmerich, and M. t Hompel. PhyNode: An intelligent, cyber-physical system with energy neutral operation for PhyNetLab. In *Smart SysTech 2015; European Conference on Smart Objects, Systems and Technologies*, pages 1–8, July 2015.

[5] A. K Ramachandran Venkatapathy, M. Roidl, A. Riesner, J. Emmerich, and M. ten Hompel. PhyNetLab: Architecture design of ultra-low power Wireless Sensor Network testbed. In *2015 IEEE 16th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–6, June 2015.

[6] A. K. R. Venkatapathy, H. Bayhan, F. Zeidler, and M. ten Hompel. Human machine synergies in intra-logistics: Creating a hybrid network for research and technologies. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1065–1068, Sept 2017.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A6

Resource-efficient Graph Mining

Kristian Kersting Petra Mutzel
Christian Sohler

Common Substructures between Molecules

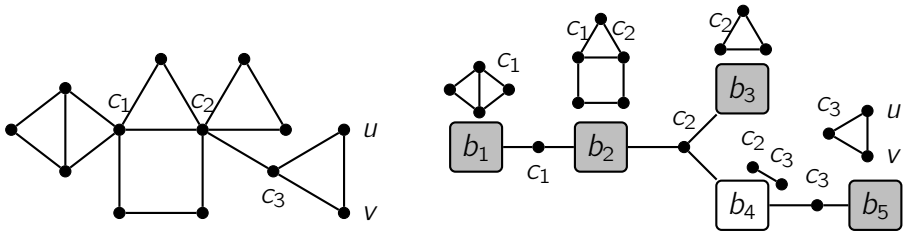
Andre Droschinsky
Chair of Algorithm Engineering (LS11)
TU Dortmund
andre.droschinsky@tu-dortmund.de

Finding the common structural features of two molecules is a fundamental task in cheminformatics. Most drugs are small molecules, which can naturally be interpreted as graphs. We can formalize this problem as maximum common subgraph problem. The vast majority of molecules yields outerplanar graphs. Unfortunately, computing a maximum common subgraph between outerplanar graphs is NP-hard.

Therefore, we consider a variation of the problem of high practical relevance, where the rings of molecules must not be broken, i.e., the block and bridge structure of the input graphs must be retained by the common subgraph. Given two outerplanar graphs subject to this constraint, our approach runs in time $\mathcal{O}(\Delta n^2)$ on n vertices with maximum degree Δ. Among others, this result, including an extensive experimental comparison, has been published at the *43rd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2017)* [3] and is presented in this report.

1 Maximum common subgraph

The maximum common subgraph problem arises in many application domains, where it is necessary to elucidate common structural features of objects represented as graphs. In cheminformatics this problem has been extensively studied [7, 8] and is often referred to as maximum or *largest common substructure problem*.



(a) Input graph (b) BC-tree of the input graph

Figure 1: A connected outerplanar graph (a) and its BC-tree (b). Block nodes are gray, while bridge nodes are white. The solid black nodes are cutvertices. The corresponding subgraphs of G are shown above the block and bridge nodes.

There are several variants of this problem. First, you may require the subgraphs to be connected. This reduces the number of possible solutions and therefore is generally faster in practice. Second, you may require that the subgraphs are induced or not. All four cases may be reduced to finding a maximum clique in a *product graph* [4, 7] and are NP-hard. Depending on your input, labels may be given, which must match in the common substructures of the input graphs. Alternatively, a weight defined between the edges and/or vertices of the two input graphs may be given. The objective depends on the specific application. It could be to maximize the number of vertices, the number of edges, the sum of both, or the weight, if a weight function is given.

If we restrict the graph classes, polynomial time algorithms are possible. The seminal work in this direction is attributed to J. Edmonds [5], who proposed a polynomial time algorithm for the maximum common subtree problem. Here, the given graphs and the desired common subgraph must be trees. Recently, we showed that this problem can be solved in time $\mathcal{O}(\Delta n^2)$ for (unrooted) trees on n vertices with maximum degree Δ [2]. Unfortunately, already for outerplanar graphs the problem is NP-hard. On the other hand, if both graphs are biconnected and outerplanar, we can compute a biconnected MCS in quadratic time, as we proved in [3]. A method to compute a common substructure between (not necessarily biconnected) outerplanar graphs was proposed in [8]. The constraint is called *block and bridge preserving* (BBP), which retains the maximal two connected components (blocks) and the bridges of the input graphs. The blocks and bridges are obtained from the BC-Tree of the input graphs. An example is shown in Figure 1. The authors provided a polynomial time algorithm for outerplanar input graphs. Their solution covers the connected, non induced case. They do support labels.

Our approach [3] features the induced variant with edge and vertex weights. The weights generalize labels. We proved a running time of $\mathcal{O}(\Delta n^2)$ for outerplanar graphs with n vertices and maximum degree Δ. We experimentally compared our algorithm to the method from [8] with the objective to maximize the number of vertices plus edges in the common subgraphs and required labels to match. The result showed that in more than 99% of the 29 000 randomly chosen molecule pairs from the NCI Open Database

Table 1: Upper half: Running times for our implementation (MCIS) and the implementation from [8] (MCES). Lower half: Relative differences in computation times.

Algorithm	Average time	Median time	95% less than	Maximum time
MCIS	1.97 ms	1.51 ms	5.28 ms	40.35 ms
MCES	207.08 ms	41.43 ms	871.48 ms	26 353.68 ms

Comparison	Average factor	Median factor	Minimum factor	Maximum factor
MCES / MCIS	83.8	25.6	1.8	28912.5

GI50 (http://cactus.nci.nih.gov) there was no difference between the induced and non induced variant. Further, our running time is much faster than their running time, see Table 1. Both programs have been written in C++ and were compiled and executed on the same machine (16 GB RAM, Intel i7 3770 CPU).

To be able to handle molecular graphs which are not outerplanar we decided to integrate the clique reduction using the algorithm for connected c-cliques from [1]. To maintain our BBP-property, we modified the reduction to report only cliques biconnected through c-edges. Further, we only use it for partial solutions if necessary. That is, if at least one of two blocks on which we compute a MCS is not outerplanar. In theory we cannot guarantee a polynomial time running time, since clique computation is NP-hard. In practice these non outerplanar blocks are often small and experiments showed that the support for those blocks increased the running time in the majority of the cases only marginally. For the few molecular graphs with high running times we added a command line parameter to use the best solution found so far within a user provided time limit.

In the field of chemistry, there are several examples of non isomorphic structures with similar chemical or physical properties. This feature is called bioisosterism, cf. [6]. During a master's thesis we integrated the detection of bioisosteres into our software. The results, both theoretical and practical, are not yet published.

2 Conclusion and outlook

We developed an algorithm to compute an induced BBP-MCS between two weighted outerplanar graphs [3]. Its practical running time is faster than a comparable (non induced) BBP-MCS algorithm [8] while delivering mostly identical results. Besides bioisosterism we currently investigate how to include disconnected parts into our MCS while maintaining polynomial running time. An example is shown in Figure 2.

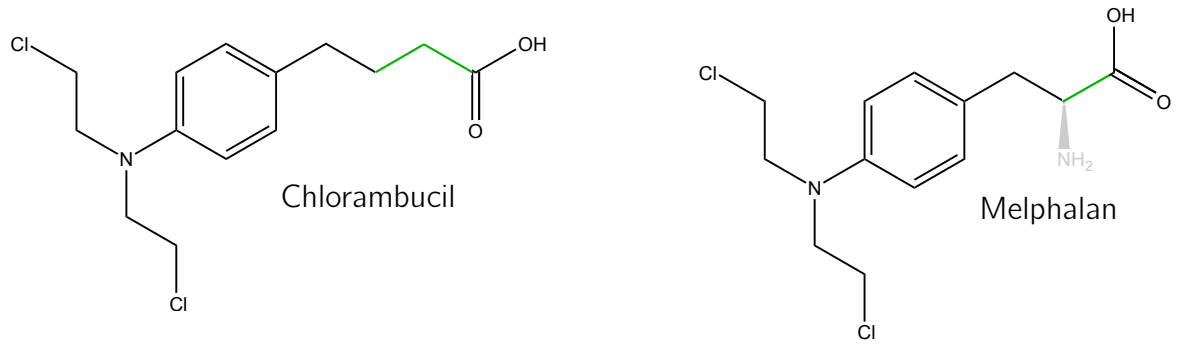


Figure 2: Two similar molecules: Chlorambucil and Melphalan. Computing a connected MCS will not add the isomorphic parts on the right side of the green edges to the solution.

References

[1] F. Cazals and C. Karande. An algorithm for reporting maximal c-cliques. *Theoretical Computer Science*, 349(3):484–490, December 2005.

[2] Andre Droschinsky, Nils Kriege, and Petra Mutzel. Faster Algorithms for the Maximum Common Subtree Isomorphism Problem. In *MFCS 2016*, volume 58 of *LIPIcs*, pages 34:1–34:14, 2016. `arXiv:1602.07210`.

[3] Andre Droschinsky, Nils Kriege, and Petra Mutzel. *Finding Largest Common Substructures of Molecules in Quadratic Time*, pages 309–321. Springer International Publishing, Cham, 2017.

[4] Ina Koch. Enumerating all connected maximal common subgraphs in two graphs. *Theoretical Computer Science*, 250(1–2):1–30, 2001.

[5] David W. Matula. Subtree isomorphism in $O(n^{5/2})$. In P. Hell B. Alspach and D.J. Miller, editors, *Algorithmic Aspects of Combinatorics*, volume 2 of *Annals of Discrete Mathematics*, pages 91–106. Elsevier, 1978.

[6] Nicholas A. Meanwell. Synopsis of some recent tactical application of bioisosteres in drug design. *Journal of Medicinal Chemistry*, 54(8):2529–2591, 2011. PMID: 21413808.

[7] John W. Raymond and Peter Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.*, 16(7):521–533, 2002.

[8] Leander Schietgat, Jan Ramon, and Maurice Bruynooghe. A polynomial-time maximum common subgraph algorithm for outerplanar graphs and its application to chemoinformatics. *Annals of Mathematics and Artificial Intelligence*, 69(4):343–376, 2013.

Higher-order Graph Classification

Christopher Morris
Chair of Algorithm Engineering
TU Dortmund University
christopher.morris@tu-dortmund.de

We report about our recent progress in developing graph kernels and graph feature maps that take global and local graph properties into account. In the first part of this report, we present the global-local Weisfeiler-Lehman graph kernel. In the second part, we outline some work in progress: neural graph feature maps that take global and local graph properties into account and are trained in an end-to-end fashion together with the classifier.

1 The Global-local Weisfeiler-Lehman Graph Kernel

Kernel methods are a broad class of algorithms for pattern analysis, which rely on a positive-semidefinite function, which measures the similarity between data objects. In several domains like chemo- and bioinformatics as well as social network and image analysis structured objects appear naturally. Graph kernels are a key concept for the application of kernel methods to structured data and various approaches have been developed in recent years, see [12, 9, 8] and references therein.

Most state-of-the-art graph kernels only take local graph properties into account, i.e., the kernel is computed with regard to properties of the neighborhood of vertices or other small substructures, see, e.g., [11, 10, 4]. In the past, few works considered graph kernels that use global graph properties. For example in [5] a kernel based on the Lovàsz number is proposed. Moreover, Kondor and Pan [6] proposed a graph kernel based on the graph Laplacian.

In [7], we introduced a kernel based on the k-dimensional Weisfeiler-Lehman kernel which is a well-known heuristic for the graph isomorphism problem. We considered the following

variant[1] of the k-dimensional Weisfeiler-Lehman algorithm (k-WL) for $k \geq 2$. Let G and H be graphs.[2] Instead of iteratively labeling vertices as the 1-dimensional Weisfeiler-Lehman algorithm, see, e.g. [1], the k-WL computes a labeling function defined on the set of subsets of cardinality k defined over the set of vertices $V(G) \cup V(H)$. In order to describe the algorithm, we define the neighborhood $N(t)$ of such a set $t = \{t_1, \ldots, t_k\}$:

$$N(t) = \{\{t_1, \ldots, t_{j-1}, r, t_{j+1}, \ldots, t_k\} \mid r \in V(G) \setminus \{t_j\}\}.$$

That is, the neighborhood $N(t)$ of t is obtained by replacing a vertex t_j from t by a vertex from $V(G) \setminus \{t_j\}$, see Figure 1a for a graphical illustration. In iteration 0, the algorithm labels each subset with its isomorphism type, i.e., two subsets s and t get the same label if the corresponding induced subgraphs are isomorphic. Now in iteration $i > 0$ we set

$$l^i(t) = \text{relabel}((l^{i-1}(t), \text{sort}(\{\!\{l^{i-1}(s) \mid s \in N(t)\}\!\}))),$$

where $\text{sort}(S)$ returns an (ascendantly) sorted tuple of the multiset S of labels and the bijection $\text{relabel}(p)$ maps the pair p to an unique label, which has not been used in previous iterations.

1.1 A Local Kernel Based on the k-dimensional Weisfeiler-Lehman Algorithm

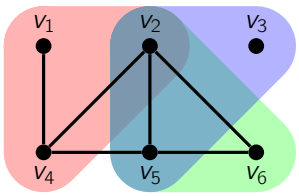
The k-WL is inherently *global*, i.e., it labels a set t by considering sets whose vertices are not connected to the vertices of t. Moreover, it does not take the sparsity of the underlying graph into account. In order to capture the local properties of a graph, we propose a *local* variant. The idea of our *local* k-WL (k-LWL) is the following: the algorithm again labels all subsets of cardinality k. But in order to extract local features we define the *local* neighborhood of a set $t = \{t_1, \ldots, t_k\} \subset V(G)$,

$$\begin{aligned} N^{\mathsf{L}}(t) = \{\{t_1, \ldots, t_{j-1}, r, t_{j+1}, \ldots, t_k\} \mid r \in V(G) \setminus \{t_j\}, \\ \text{and } \exists\, l \in [1\!:\!k]\colon (t_l, r) \in E(G)\}, \end{aligned}$$

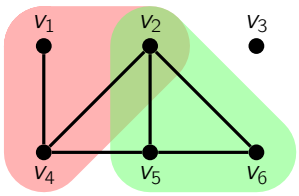
i.e., we consider a set s a local neighbor of t if s is in $N(t)$, and there is at least one edge between vertices of s and t, see Figure 1b for a graphical illustration. Now the local algorithm works the same way as the k-WL but in each iteration $i > 0$ considers the local neighbors of a set. After each iteration the algorithm computes a feature map where each component represents color counts of a certain color at the current iteration. The kernel is finally computed by taking the inner product in the corresponding feature space.

[1] In theoretical computer science it is usually defined on k-tuples. Due to scalability we considered k-sets instead.

[2] For clarity of presentation we omit the presence of node and edge labels. Observe that the algorithm can be extended to take node and edge labels into account.



(a) Illustration of the global neighborhood of the set $\{v_2, v_4, v_5\}$, for better clarity we only depict a subset.



(b) Illustration of the local neighborhood of the set $\{v_2, v_4, v_5\}$, for better clarity we only depict a subset.

Figure 1: Illustration of the local and global neighborhood of the set $\{v_2, v_4, v_5\}$

Since the algorithm operates on all subsets of cardinality k defined over the set of vertices of a graph the algorithm may not scale to large graph data bases. In order to overcome this limitation, we devised a stochastic version of the kernel with provable approximation guarantees using conditional Rademacher averages. On bounded-degree graphs, it can even be computed in constant time, i.e, the running time does not depend on the number of nodes or edges of the graph.

We supported our theoretical results with experiments on several graph classification benchmarks, showing that our kernels often outperform the state-of-the-art in terms of classification accuracies.

2 Neural Higher-order Graph Feature Maps

In recent years, graph feature maps based on ideas from deep learning have emerged. For example in [2] a neural version of the 1-dimensional Weisfeiler-Lehman algorithm has been proposed which uses a parameterized neighborhood aggregation function. These parameters are then trained in an end-to-end fashion together with parameters of the classifier, e.g., a neural net, to better adapt them to the given data distribution. Results, e.g, see [2] and [3], show that these methods improve over the classic kernel methods, e.g., [11] in terms of classification accuracy. Up to now these methods were only evaluated empirically. Hence, in order to understand the mechanisms of these algorithms better we want to conduct a theoretical analysis of the algorithm.

Moreover, we want to develop neural graph feature maps which take higher-order graph properties into account. For example we could develop neural variants of the algorithm sketched in the first section.

References

[1] L. Babai, P. Erdős, and S. M. Selkow. "Random Graph Isomorphism". In: *SIAM Journal on Computing* 9.3 (1980), pp. 628–635.

[2] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. "Convolutional Networks on Graphs for Learning Molecular Fingerprints". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2224–2232.

[3] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. "Neural Message Passing for Quantum Chemistry". In: 2017.

[4] S. Hido and H. Kashima. "A Linear-Time Graph Kernel". In: *9th IEEE International Conference on Data Mining*. 2009, pp. 179–188.

[5] F. D. Johansson, V. Jethava, D. P. Dubhashi, and C. Bhattacharyya. "Global graph kernels using geometric embeddings". In: *31st International Conference on Machine Learning*. 2014, pp. 694–702.

[6] R. Kondor and H. Pan. "The Multiscale Laplacian Graph Kernel". In: *Advances in Neural Information Processing Systems*. 2016, pp. 2982–2990.

[7] C. Morris, K. Kersting, and P. Mutzel. "Glocalized Weisfeiler-Lehman Graph Kernels: Global-Local Feature Maps of Graphs". In: *17th IEEE International Conference on Data Mining (ICDM)*. 2017.

[8] N. M. Morris C. Kriege. "Recent Advances in Kernel-Based Graph Classification". In: *European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases*. 2017.

[9] M. Neumann, R. Garnett, C. Bauckhage, and K. Kersting. "Propagation kernels: Efficient graph kernels from propagated information". In: *Machine Learning* 102.2 (2016), pp. 209–245.

[10] F. Orsini, P. Frasconi, and L. De Raedt. "Graph Invariant Kernels". In: *24th International Joint Conference on Artificial Intelligence*. 2015, pp. 3756–3762.

[11] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. "Weisfeiler-Lehman Graph Kernels". In: *Journal of Machine Learning Research* 12 (2011), pp. 2539–2561.

[12] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. "Graph Kernels". In: *Journal of Machine Learning Research* 11 (2010), pp. 1201–1242.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B1

Analysis of Spectrometry Data with Restricted Resources

Jörg Ingo Baumbach Jörg Rahnenführer

Confounder identification and correction in MCC-IMS

Salome Horsch
Statistical Methods in Genetics and Chemometrics
Technische Universität Dortmund
salome.horsch@tu-dortmund.de

Current research about diagnosing diseases through breath gas analysis via MCC-IMS often considers only the metabolites detected in breath. Clinical variables however are often not reported even though they might have crucial impact on the statistical classification. Furthermore, it can be observed that different MCC-IMS devices produce slightly different results. This could endanger the success of a classification model and especially the generalizability of models built on data from one particular device. In order to address the points of interest we conducted a small study with 49 healthy probands whose breath was measured on two devices. After the first measurement the subjects drank a glass of orange juice before the second measurement to simulate a diseased status for each subject. We show that the two devices achieve different results for some metabolites and how scaling of both devices separately makes up for the discrepancy.

In order to identify possible confounders in MCC-IMS analysis and their impact on classification results, a suitable dataset is needed, where possible confounders can be controlled (device) or at least observed (sex, smoking habits). In clinical studies, the device is usually not varied and each participant in a study is either a case or a control in a classification task. Therefore, we conducted a study, where we "simulate" a disease by changing the breath gas of a subject. This can easily be accomplished by letting the participants drink a glass of orange juice. That way each subject is first a control (before drinking the juice) and then a case (after drinking the juice), which allows direct comparison of the two samples. Each of the two measurements is processed by a different device (starting device is switched) such that the effect of the device can also be assessed.

Study design

The study was conducted over a period of four consecutive days in 2017 in the facilities of B&S Analytik GmbH in Dortmund. The 49 volunteering subjects were mostly Ph.D. students or staff from the TU Dortmund university and the labs where the measurements were taken. There were no criteria for exclusion except for pregnancy (due to lab regulations) and an allergy to citrus fruits (because a glass of orange juice had to be consumed).
The subjects answered a questionnaire about age and sex as well as their eating and drinking habits of the last 12 hours and their smoking habits. Breath gas of the subjects was measured on one MCC-IMS device first, then they drank a glass of orange juice before measuring the breath on the other device. Which device was used first was determined randomly with stratification for sex such that approximately half of both sexes started with each device. 26 of the 49 subjects were male, 23 female. The first measurement (without juice) was executed by device A for 13 males and 11 females, whereas 13 males and 12 females started with device B.
Each measurement consists of three sub-measurements. First, the breath sample was analyzed by the MCC-IMS device. Afterwards, a sample from the room air was taken and analyzed as well. During the last measured step a special gas was lead through the device in order to clean it from possibly remaining metabolites.
The MCC-IMS devices deliver a heat map image, where peak areas represent certain volatile organic compounds in the air. Detection of the peaks and merging them over different measurements to a well defined set of variables is achieved by two different approaches. The current gold standard is the semi manual examination of the heatmaps using the software VisualNow (VN). For comparison, we also use the results of an entirely automated peak detection (SGLTR/DBSCAN) that showed promising results in a following statistical classification task (see [1]).

Results

For broad application of the MCC-IMS technology it is vital that the measurements on different devices lead to comparable results. Figure 1 displays the first two principal components (PC) of the data, restricted to breath data and the automated peak detection (AP) results. The shapes of the symbols represent the device whereas the color marks, if the respective subject had consumed orange juice before the measurement. It can be seen, that the device contributes considerably to the variance in the data (second PC). The first PC however explains differences between the two groups (juice or no juice).
Closer investigation reveals two reasons for these strong differences. Some of the identified metabolites appear only in either one of the devices. This could go back to different

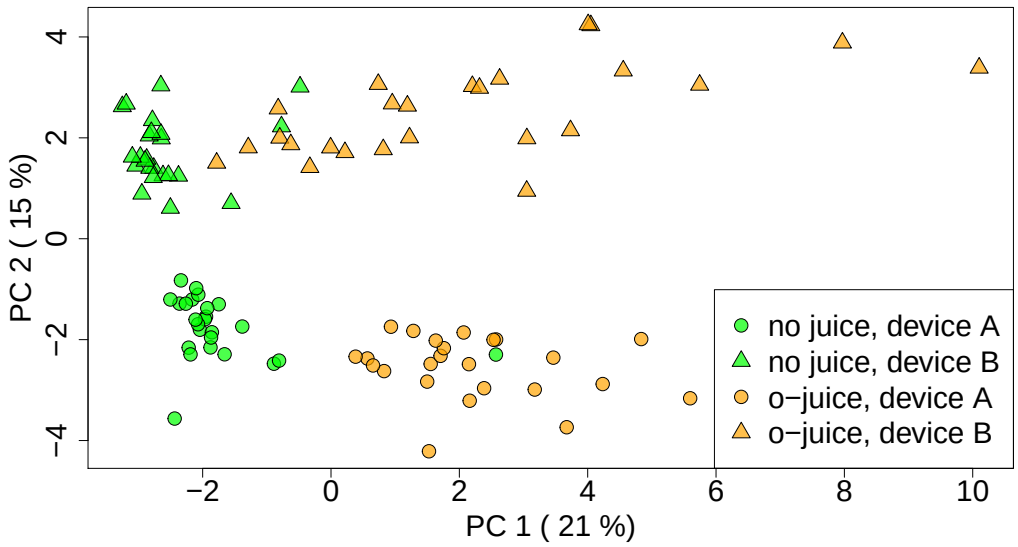


Figure 1: First and second principal components of the PCA for the breath sample data.

materials used in the device. Alternatively, it might be an artifact where the raw measurements of both devices are shifted such that during the peak clustering process the same metabolite is assigned to two clusters, one for each device. The other reason for the differences is a different scale of the measured intensities.
Figure 2 (top left) shows one example metabolite for both devices. Here, all measurements (breath gas, room air and flushing) are displayed in the order they were measured using the automated peak detection. It can clearly be seen that the intensities are remarkably lower for device B. Considering all peaks (45 for AP and 124 for VN) there is no obvious pattern for the relationship of the two devices, i.e. each of the devices has higher values than the other for some metabolites. Therefore, all variables are scaled separately in order to overcome systematic differences. A common approach is normalizing variables to mean zero and variance one. The result of normalizing both devices in this way is displayed in Figure 2 (top right). The results of the automated algorithm contain many zeros (meaning that a peak was not found or below a built-in threshold) that strongly affect mean and variance. As a result, the devices still show different intensity levels and the zeros on one device have no longer the same meaning as zeros on the other device. In order to avoid that, the scaling is improved by treating zeros differently and just normalizing the non-zero values. The zeros of both devices are shifted in the same way, in order to assure that they remain the smallest values. The results are shown in Figure 2 (bottom) for the normalization to mean zero and variance one (left) and a robust alternative normalizing to median zero and MAD (median absolute deviation from the median) one which allows for metabolites with outliers.

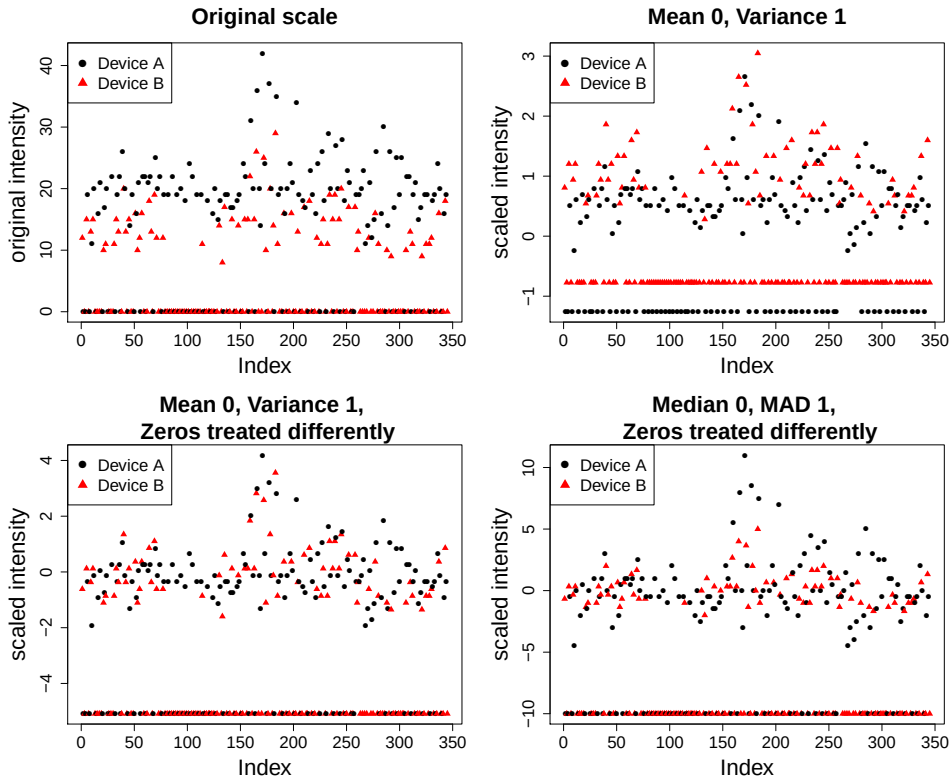


Figure 2: The original values of one metabolite (top left) and the results of three different scaling methods accounting for the two devices.

Outlook

Future research focuses on further understanding the differences between the devices and between the results of AP and VN. In comparison to AP, the VN dataset contains a lot more variables but fewer zeros since it doesn't use a threshold. Additionally, the importance of other variables like the sex of a subject will be examined. For assessing the success of potential corrections, the impact on the classification performance will be used. Since the classification of juice or no juice is very easy, the data will be altered in order to create a harder classification problem based on real data.

References

[1] S. Horsch, D. Kopczynski, E. Kuthe, J. I. Baumbach, S. Rahmann, and J. Rahnenführer. A detailed comparison of analysis processes for mcc-ims data in disease classification - automated methods can replace manual peak annotations. *PLoS One*, 12, 2017.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B2

Resource optimizing real time analysis of artifactious image sequences for the detection of nano objects

Jian-Jia Chen Roland Hergenröder
Heinrich Müller

Cache-Aware Investigation for Decision Tree Optimization

Kuan-Hsun Chen
Lehrstuhl für Technische Informatik und Eingebettete Systeme
Technische Universität Dortmund
kuan-hsun.chen@tu-dortmund.de

Due to "memory wall" phenomenon, hierarchical cache memories have been widely used in most modern architectures to hide the memory latencies. As a basic unit of random forests, Decision Tree seems clearly that abandons the rationale behind the cache memories. From the two typical implementations, the native tree and the if-else tree, we can see that the data excess pattern or the execution of the instructions totally behaves without locality. In this report, we present the experimental results, which roughly gets the margin how much we can improve from taking locality into consideration on a Raspberry Pi. This motivates us to figure out how to create the locality of the decision tree so that the advantage of cache memories can be taken.

As a promising and efficient method in machine learning domain, Random Forests have been applied to a variety of problems and are often referred to as one of the best black-box methods available offering high accuracy with only a few parameters to tune [2, 3]. However, we can observe that, the memory excess pattern even in its basic unit Decision Tree has no locality at all, which may experience extremely bad performance w.r.t. execution time on modern architectures. But why does the locality matter?

Due to "memory wall" phenomenon, hierarchical cache memories have been widely utilized in most modern architectures to hide the memory latencies. To take advantage of cache memories, application designers are expected to intensively consider the locality while designing their programs. There are two type of localities: the temporal locality and the spatial locality. The temporal locality means that the data/instructions which are just referred will likely be referred again soon. Likely a summation variable for recording the accumulating values will be referred in each iteration, by which this variable will be referred many time once it is loaded into the data-cache. The instructions representing

a cycle through the loop will be executed repeatedly, by which the instructions in the instruction-cache will be executed until the end of the loop.

The spacial locality means that items with nearby addresses tend to be referenced close together in time. A well-known example is that the array elements in succession will be likely referred. Therefore most programmers are educated to store and excess the memory sequentially, so that the advantage of cache memories can be fully exploited. Likely instructions are expected to be executed in sequence (i.e., in a basic block), so the prefetching which loads a few close instructions into the instruction-cache indeed increases the performance.

Let's investigate how the Decision Trees are implemented typically. There are two types of implementations: the native tree and the if-else tree [1]. The native tree uses a while/for loop to iterate over each individual node of a decision tree. In the state-of-the-art, those data of nodes will be stored into a continuous data structure, i.e., a *struct* array. The elements are usually indexed with a breadth-first traversal or a depth-first traversal, and the tree traverse goes through the data structure node by node. However, the locality issue here is that those loaded elements (nodes) with nearby addresses will not be utilized at all when the distance between each element in the traverse is greater than the maximum number of elements can be loaded into a cache line once.

The if-else tree is to use statically-generated if-else blocks to build up a Decision Tree, in which the data of threshold and the value are all hard-coded into the instructions with the constant value. Therefore, the latency incurred by indirect memory accesses can be avoided significantly. However, likely the hard-coded instructions are just loaded into cache once but only used once if the code size is too big, by which the advantage of the locality in the instruction cache is totally abandoned.

To have a sense that how bad the performance can be without locality, I have tried some dummy examples shown in Listing 1 on a Raspberry Pi Model B+. Function test1() is the reference with a good spatial locality. Function test2() shows that if the memory excess always evicts the loaded data in the same cache line on purpose. It should results in the same situation that if the spatial locality does not exist. Please note, that those numbers in the experiments are designed carefully depending upon the properties of the data-cache of the targeted platform. In ARM reference manual pg. B6-6, the specs of L1-cache are detailed as follows:

- 16KB L1 data cache / 16KB L1 instruction cache
- It is 4-Way Set associative.
- The number of sets is 128.
- The length of a cache line is 32 bytes.

Therefore we know that the cache block in one way is 4096 bytes = 32 bytes $\times$128. To simplify the system, we disable L2 cache and only consider L1 cache. We use Raspbian with Linux kernel 4.9. The cache replacement policy we use the default policy, which is random. We use Linux original performance meter named perf to evaluate our results.

```
int test1(){
        int array[1048577];
        int c=0;
        for(int i=0; i<1000000; i++)
                for(int j=0; j<16; j+=1)
                        c = array[j];
        return 0;
}
int test2(){
        int array[1048577];
        int c=0;
        for(int i=0; i<1000000; i++)
                for(int j=0; j<1048577; j+=1024)
                        c = array[j];
        return 0;
}
```

Listing 1: Example for dummy testing the locality impact

```
Performance counter stats for 'test1.out 1048576' (5 runs):

267,956          L1-dcache-stores
34,128           L1-dcache-load-misses
28,809,444    cycles

0.047688283 seconds time elapsed
Performance counter stats for 'test2.out 1048576' (5 runs):

744,188          L1-dcache-stores
38,538           L1-dcache-load-misses
162,531,937  cycles

0.238092822 seconds time elapsed
```

Listing 2: Example for dummy testing the locality impact

Consequently, as shown in Listing 2, we can see that Test2() is roughly five times slower than Test1(). This really motivates us to think about how to optimize the structure of the Decision Tree, especially for the native tree and the if-else tree. Up to here, we can

conclude that the utilization of cache does really matter. Clearly we know Decision Tree is efficient and widely used in machine learning domain. If this problem can be solved properly, this kind of essential improvement can bring us great benefit in the future.

References

[1] N. Asadi, J. Lin, and A. P. de Vries. Runtime optimizations for tree-based machine learning models. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2281–2292, Sept 2014.

[2] R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 96–103. ACM, 2008.

[3] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res*, 15(1):3133–3181, 2014.

Morse-Smale Complexes For Ridge Detection of 3D Real Functions

Thomas Kehrt
Informatik 7, Graphische Systeme
Technische Universität Dortmund
kehrt@ls7.cs.tu-dortmund.de

An algorithm is outlined which approximates the Morse-Smale cell complex of a real function D with a 3D-domain as the cell complex induced by integral curves of the gradient field of D of equal behavior. In the underlying application, D results from a point set of surface samples by low-pass filtering.

1 Introduction

Even though the approach presented in the following works in general it will be explained with respect to a specific application. The application is the reconstruction of a surface in $\mathbb{R}^3$ from a set P of noisy point samples p_i with limited density. Convolution of the samples with a low-pass filter with kernel K results in a density function $D(x) = \sum_i K(x - p_i)$ on $\mathbb{R}^3$. The domains of the ridges of D are considered as the reconstruction of the surface from the point set.

If K is chosen as radial Gaussian with aperture σ, the ridges are rather "hilly" and contain maxima and saddle points. This gives rise to decomposing the domain of D by means of a Morse-Smale cell complex [5,6]. A mesh approximating the domain of a ridge, and thus representing the reconstructed surface, is induced by the given point samples as nodes, and edges which connect a pair of nodes if the Morse-Smale cells containing them are neighboring, cf. the report of last year for an illustration of the 2D-case.

The Morse-Smale cell complex (MSC) of a real function consists of cells of equal behavior of the integral curves of its gradient vector field ∇D. An integral curve, or stream line,

of a vector field is a curve which follows the vector field, i.e. whose tangents are co-linear to the field vectors at the curve points. The cells of the complex result by aggregating all points of the domain whose traversing integral curves share a certain property. For ridge construction only those cells are relevant whose integral curves terminate in a common maximum or in a common saddle point of D, respectively. Cells of the first type are full-dimensional, i.e. of dimension 3. Cells of the second type are of lower dimension, i.e. 2, 1, or 0. They separate the interior of two full-dimensional cells and together form the so-called separatrix.

In the following an algorithm for approximate construction of those cells will be presented. A particular challenge is numerical instability of the separatrix which causes difficulties with respect to reliable calculation. Even a slight offset from the separatrix implies an integral curve terminating in a maximum rather than in a saddle point.

2 Cell Construction

In the first step, the full-dimensional cells are calculated from D. For this purpose, a cuboid containing the point cloud is partitioned by a regular hexahedral mesh, and D is discretely represented by its values at the mesh vertices. The cells are determined by a labeling algorithm. The labeling algorithm, which will be outlined in chapter 3, assigns labels to the mesh vertices so that all vertices with the same label together define a cell.

The spacing of the mesh vertices is set to $0.9\sqrt{3}\sigma$ in all dimensions. This is derived from the observation that the sum of two Gaussians induce a single maximum if their distance is smaller than 2σ. Such maxima can be captured if the volume is sampled at least twice in all directions, in accordance with the sampling theorem of Nyquist and Shannon. This implies that the diagonals of the cubic mesh cells have to be shorter than σ, and hence their edges to be shorter than $\sqrt{3}\sigma$. The factor 0.9 is chosen arbitrarily.

The second step determines the separatrix cells. Given the labeling of the first step, a surface separating the mesh vertices labeled differently is calculated by a marching cubes algorithm with non-binary labels, like e.g. the multi-material marching cubes by Wu and Sullivan [8]. The resulting surface is represented by a triangular mesh which generally is not manifold, i.e. it may have edges with more than two incident triangles. The cells of dimension 2 are obtained by evaluating D at the mesh nodes and applying the labeling algorithm to the mesh, with exception of the nodes on non-manifold edges.

The non-manifold edges lead to the cells of dimension 1 and 0. The 0D-cells are defined by the branching vertices of the graph induced by the non-manifold edges, and the 1D-cells are obtained by applying the labeling algorithm to the rest graph without the branching vertices.

3 Labeling

The input of the labeling algorithm is a manifold mesh with real function values at its nodes. Relevant mesh types in the context here are hexahedral meshes for 3-manifolds, triangular meshes for 2-manifolds, or polygonal chains for 1-manifolds. The labeling algorithm calculates a forest, i.e. a sub-graph of the edge graph of the given mesh which consists of a union of trees. In each step the algorithm takes a node of the mesh not yet processed, and looks for an adjacent node with maximum D-value. If the value is higher than the one of the nodes itself, it adds the edge between the two nodes to the result. If no such node exists, the node is a local maximum and defines the root of a tree. After processing all vertices, a unique label is assigned to every tree, and this labeling is reported as result.

The idea of the algorithm is that the selected edge is a discrete approximation of the gradient ∇D at the node under consideration, and the path from a node to the root of its tree approximates an integral curve. This heuristic approach avoids the conventional approximation of integral curves by e.g. a Runge-Kutta scheme which is costly due to the evaluation of ∇D for each iteration.

4 Discretization Artifacts, Their Suppression and 2-Manifold Enforcement

The 2D-cells approximate the real separatrix only up to a precision of $\sigma/2$. The evaluation of the density function is effected negatively by the distance to the real separatrix. The labeling algorithms sporadically fails when generating the integral curve trees such that too many cells are produced. In those cases we introduce an artificial 1D-cell using the marching triangles algorithm to separate them.

1D-cells are even more prone to noise. The reason is that the behavior of the density function is more complex at the 1D-cells. While the saddles on the 2D-cells can be approximated by a quadric, i.e. their behavior can be sufficiently described by the Hessian matrix, this is not the case for the saddles of the 1D-cells. 1D-cells result from the interaction of more than two 2D-cells. In terms of Morse-Smale-systems their upper link has more than two maxima (see [2, 3]). This produces more than one maximum per non-manifold polygonal line which compete in connecting 2D-cells.

Another source of artifacts is the vertex spacing of the 3D-mesh. In fact, the theoretical considerations only hold near the real surface. Due to low-pass filtering, the sought surface contracts towards its concave side thus deforming 3D-cells inwards.

One possible solution to cope with the above mentioned artifacts could be a finer mesh resolution, but this would increase computational complexity considerably. Due to the

chosen mesh spacing, the number of nodes of the hexahedral grid is just proportional to the number of points in the point cloud. However, evaluating D has time complexity $\mathcal{O}(|P| \log |P|)$, P the set of sample points, using an octree [1] and the fact that the kernel function evaluates numerically to zero at a distance less than 10σ when using single precision arithmetic.

A second problem is that if P represents the surface of a physical object the surface to be reconstructed must be 2-manifold.

Both problems can possibly be solved by analyzing the Reeb-graph of the cell complex. Current efforts concern the development of a set of rules to reduce the full Reeb-graph to a subset having the desired features. In the process a subset of the graph nodes shall be eliminated which will not be considered in the result. However, this is work in progress and no results are yet available to prove the concept.

References

[1] Josh Barnes and Piet Hut. A hierarchical o (n log n) force-calculation algorithm. *nature*, 324(6096):446–449, 1986.

[2] Herbert Edelsbrunner, John Harer, Vijay Natarajan, and Valerio Pascucci. Morse-smale complexes for piecewise linear 3-manifolds. In *Proceedings of the Nineteenth Annual Symposium on Computational Geometry*, SCG '03. ACM, 2003.

[3] Herbert Edelsbrunner, John Harer, and Afra Zomorodian. Hierarchical morse complexes for piecewise linear 2-manifolds. In *Proceedings of the Seventeenth Annual Symposium on Computational Geometry*, SCG '01, pages 70–79, New York, NY, USA, 2001. ACM.

[4] R Forman. A user's guide to discrete morse theory. In *Séminaire Lotharingien de Combinatoire*, 48, 2002.

[5] A. Gyulassy, P. T. Bremer, B. Hamann, and V. Pascucci. A practical approach to morse-smale complex computation: Scalability and generality. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1619–1626, Nov 2008.

[6] A. Gyulassy, V. Natarajan, V. Pascucci, and B. Hamann. Efficient computation of morse-smale complexes for three-dimensional scalar functions. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1440–1447, Nov 2007.

[7] J Milnor. *Morse Theory*. Princeton University Press, 1963.

[8] Ziji Wu and John M Sullivan. Multiple material marching cubes algorithm. *International Journal for Numerical Methods in Engineering*, 58(2):189–207, 2003.

Real-Time Low-SNR Signal Detection and Processing of Irregular Structured Data with CNNs

Jan Eric Lenssen
Department of Computer Graphics
TU Dortmund University
janeric.lenssen@tu-dortmund.de

We report advances in two different subjects. First, a real-time capable data processing pipeline combining different Convolutional Neural Networks with traditional methods for PAMONO image processing has been developed and deployed on mobile and desktop hardware. The methods improve the state of the art in low signal-to-noise ratio (SNR) signal detection and allow the automatic detection and classification of signals with an SNR below 1.

Second, advances in the field of geometric deep learning for irregular structured input data have been made. We introduced a novel B-spline based convolution operator to apply CNNs on data like meshes, (embedded) graphs or point clouds, achieving state-of-the-art results in benchmark tasks of image graph classification, graph node classification and shape correspondence.

1 Real-time ultra low SNR signal detection

The PAMONO-sensor is a biosensor that allows the detection of nano-particles (e.g. viruses, virus-like particles or fine dust) utilizing the effect of Surface Plasmon Resonance (SPR) [4, 10]. We developed a real-time capable signal processing pipeline that is able to detect signals in PAMONO images with an SNR below 1.0 [6]. The pipeline is shown in Figure 1. In addition to traditional image processing methods, the pipeline contains four different CNNs performing signal detection, classification and nano-particle size estimation [5]. In order to achieve the real-time property, different concepts

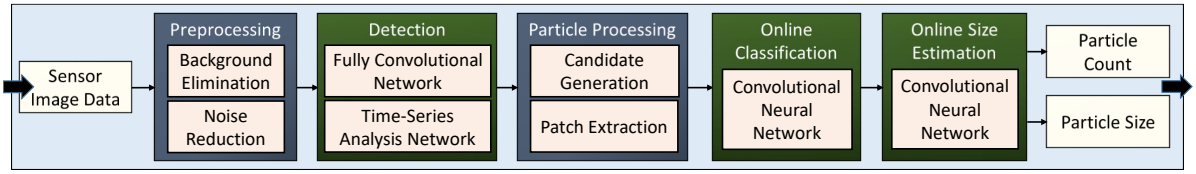


Figure 1: Real-time GPU image processing pipeline containing different DNNs.

Table 1: Detection results evaluated for the previous method M^{Baseline} and two different methods $M2^{\text{FCN-Det}}_{\text{CNN-Cl}}$ and $M4^{\text{TS-Det}}_{\text{CNN-Cl}}$, which are enhanced by deep neural networks.

Data set / method		M^{Baseline}	$M2^{\text{FCN-Det}}_{\text{CNN-Cl}}$	$M4^{\text{TS-Det}}_{\text{CNN-Cl}}$
80 nm Dataset 1	Precision	0.330	**0.840**	0.829
	Recall	0.549	**0.707**	0.428
	F_1-score	0.412	**0.768**	0.564
80 nm Dataset 2	Precision	0.053	**0.801**	0.729
	Recall	**0.561**	0.553	0.355
	F_1-score	0.097	**0.655**	0.478

for resource-efficient CNNs have been used, namely depthwise separable convolutions, 1×1-convolutions and feature map reduction layers. Together with *deepRacin*, our own OpenCL-based CNN inference library, we are able to process PAMONO image data in real-time, detecting, classifying and estimating the size of nanoparticles with size down to 80 nanometers. Selected results for signal detection in comparison to previous methods are shown in Table 1.

2 SplineCNN: Spline-based convolution for CNNs

In the context of our collaboration and future participation in project A6, we developed a variant of CNNs for irregular structured data using a novel B-spline based convolution operator [2]. The resulting SplineCNNs are able to achieve state-of-the-art results in tasks on graph and embedded graph data.

Preliminaries. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{U})$ be a directed graph, with $\mathcal{V} = \{1, \dots, N\}$ being the set of nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the set of edges, and $\mathbf{U} \in [0,1]^{N \times N \times d}$ containing d-dimensional pseudo-coordinates $\mathbf{u}(i,j) \in [0,1]^d$ for each directed edge $(i,j) \in \mathcal{E}$. $\mathbf{U}$ can be interpreted as an adjacency matrix with d-dimensional, normalized entries $\mathbf{u}(i,j) \in [0,1]^d$ if $(i,j) \in \mathcal{E}$ and $\mathbf{0}$ otherwise. For a node $i \in \mathcal{V}$ its *neighborhood* set is denoted by $\mathcal{N}(i)$. Let $\mathbf{f} : \mathcal{V} \to \mathbb{R}^M$, with $\mathbf{f}(i) \in \mathbb{R}^M$, denote a vector of M input features for each node $i \in \mathcal{V}$. For each $1 \leq l \leq M$ we reference the set $\{f_l(i) \mid i \in \mathcal{V}\}$ as *input feature map*.

Method	Accuracy [%]
MoNet [9]	91.11
SplineCNN	**95.22%**

(a) MNIST superpixel dataset

Method	Accuracy [%]
ChebNet [1]	87.12
GCN [3]	87.17
CaleyNet [7]	87.90
SplineCNN	**89.48%**

(b) Cora citation graph.

Method	Accuracy [%]
MoNet [9]	88.48
FMNet [8]	97.88
SplineCNN	**99.20%**

(c) FAUST dataset

Table 2: Classification accuracy of SplineCNNs and previous best performing methods for three different tasks: (a) image classification on 75 MNIST superpixels, (b) graph node classification on the Cora citation graph and (c) exact shape correspondence on meshes of the FAUST dataset.

Convolution Operator. We define an operator that convolves the irregular sampled graph signal $\mathbf{f}$ with a continuous kernel function $\mathbf{g}$ as

$$(\mathbf{f} \star \mathbf{g})(i) = \frac{1}{|\mathcal{N}(i)|} \sum_{l=1}^{M} \sum_{j \in \mathcal{N}(i)} f_l(j) \cdot g_l(\mathbf{u}(i,j)), \tag{1}$$

where $\mathbf{g}$ is a trainable B-spline surface with local supporting basis functions. The convolution aggregates weighted node features in each local neighborhood while the weights are sampled from the trainable, continuous kernel $\mathbf{g}$ based on local relations $\mathbf{u}$. For further details and information about the construction and training of the kernel function we refer to the published work [2].

Using this operator, we construct a convolutional layer that can be used in deep neural network architectures to train models on (embedded) graph data. Selected results for three different tasks (image graph classification, graph node classification and shape correspondence) are shown in Table 2. We also implemented the presented SplineCNNs on the GPU to provide fast training and inference algorithms. Therefore, inference on graph and mesh data can be performed under stronger resource constraints compared to related methods.

3 Future Work

Both presented fields will be further explored. For processing of PAMONO image data, the training based on generated training data is interesting for practical application of the method. Also, transfer learning for slightly variated data will be explored. The SplineCNNs hold potential to be applied to a large number of different datasets and deep learning problems. Also, additional methods from traditional CNNs can be transferred to this novel deep network.

References

[1] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pages 3837–3845, 2016.

[2] Matthias Fey, Jan Eric Lenssen, Frank Weichert, and Heinrich Müller. SplineCNN: Fast geometric deep learning with continuous b-spline kernels. *CoRR*, abs/1711.08920, 2017.

[3] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*, 2017.

[4] Jan Eric Lenssen, Victoria Shpacovitch, Dominic Siedhoff, Pascal Libuschewski, Roland Hergenröder, and Frank Weichert. A review of nano-particle analysis with the pamono-sensor. *Biosensors: Advances and Reviews, IFSA Publishing*, pages 81–100, 2017.

[5] Jan Eric Lenssen, Victoria Shpacovitch, and Frank Weichert. Real-time virus size classification using surface plasmon resonance and convolutional neural networks. *Bildverarbeitung für die Medizin 2017*, pages 98–103, 2017.

[6] Jan Eric Lenssen, Anas Toma, Albert Seebold, Victoria Shpacovitch, Pascal Libuschewski, Frank Weichert, Jian-Jia Chen, and Roland Hergenröder. Real-time low snr signal processing for nanoparticle analysis with deep neural networks. *BIOSIGNALS 2018 (accepted for publication)*, 2018.

[7] Ron Levie, Federico Monti, Xavier Bresson, and Michael M. Bronstein. CayleyNets: Graph convolutional neural networks with complex rational spectral filters. *CoRR*, abs/1705.07664, 2017.

[8] Or Litany, Tal Remez, Emanuele Rodolà, Alexander M. Bronstein, and Michael M. Bronstein. Deep functional maps: Structured prediction for dense shape correspondence. *CoRR*, abs/1704.08686, 2017.

[9] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Jan Svoboda, and Michael M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model CNNs. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[10] Victoria Shpacovitch, Vladimir Temchura, Mikhail Matrosovich, et al. Application of surface plasmon resonance imaging technique for the detection of single spherical biological submicrometer particles. *Analytical Biochemistry*, 486:62 – 69, 2015.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B3

Data Mining on Sensor Data of Automated Processes

Jochen Deuse Katharina Morik

Active Learning For Reducing Process Simulation Cost

Amal SAADALLAH
Lehrstuhl für Künstliche Intelligenz, LS 8
Technische Universität Dortmund
amal.saadallah@tu-dortmund.de

This report sums up our recent research to use Active Learning for reducing industrial processes simulations running costs. Due to their computational time, it is most often difficult to collect large amounts of data for learning tasks from these simulations. In our work, we propose an active learning approach that iteratively selects the most informative simulation scenarios for a given learning model. We evaluate our framework using mechanized tunnelling simulation for surface settlement calculation data. Experimental results are detailed in this report, followed up by our future work within the same topic.

1 Introduction

Process simulations are increasingly required to analyse processes regarding efficiency, stability and resulting outputs quality. Due to their computational time, these simulations are costly to run and also inappropriate for real-time applications. New trend in applied machine learning aim to replace simulations with surrogate models that are more appropriate for real-time application. However, building these models requires large amounts of simulation data. To accomplish this step with minimal costs, we propose an active learning approach that investigates both simulation input parameters space and the performance of the surrogate model with simulation data that have been collected so far. In other words, given an initial small number of simulation scenarios, our framework is able to iteratively select the next combinations of input parameters to run next that guarantee a continuous improvement in the performance of the surrogate model. We

evaluated our framework using data collected from a 3D numerical simulation for mechanised tunnelling. In the following, we give a brief overview of active learning. Then, we formally present our methodology. The obtained results are detailed and briefly discussed. Finally, future work is presented.

2 Active Learning

The main motivation behind active learning is that a learning model trained on a small training set is able to perform efficiently as a model trained on a larger number of examples randomly chosen, while being computationally smaller [1, 2]. Following this idea, active learning exploits the user–machine interaction to increase the model accuracy using an optimized training set. Starting from a small and non-optimal training set, the active learning procedure aims to select iteratively unlabelled data points whose inclusion in the training set improves the performance of the learning model. Formally, starting from an initial small labelled set L, the unlabelled samples set U are evaluated and sorted according to a selection criterion f_{AL}. From the sorted samples U_s, the first N_s samples are selected, labelled by an oracle and finally added to the training set L. The entire process is iterated until a stopping criterion is met (e.g., the total number of samples to add to the training set is reached, or the accuracy improvement on an independent calibration/validation set over the last iterations becomes insignificant).

Traditional active learning approaches are generally applied to learning problems with large amounts of unlabelled data where the cost of labelling is high [2]. However, it can also be viewed as a tool for informed sampling from a given distribution of data points.

3 Method: The Active Learning (Hybrid-DAL) Framework

The training set in our case is decomposed from different simulation scenarios (i.e. different combinations of input parameters). Let L be the initial training set, composed of n labelled scenarios (i.e. Simulations that we have already run from random combinations of input parameters) and U a learning set, composed of $m(m >> n)$ unlabelled scenarios. θ denotes the vector of simulation input parameters. N_s is the number of scenarios to select from the n labelled scenarios and N_u is the number of scenarios to add at each iteration of the active learning procedure.

In a first step we train the surrogate model with the labelled training set L. As our method investigates both labelled and unlabelled instances spaces unlike previous works that have been focusing only on the unlabelled instances space [3], in a second step we perform the first selection that will be performed over the labelled scenarios. This selection will define the subset of labelled scenarios to consider in the next iteration of the active learning procedure. First we compute the prediction error for each labelled scenarios S_i in L, $e_i(\theta_i)$. Second, we choose the N_s scenario with the highest error values. In a third step, we select the next N_u simulation scenarios to run. First, we compute the Euclidean distances between the N_s selected training scenarios and each scenario S_u $(u = n+1, n+2, ..., n+m)$ from the learning set U.

$$Ed_u \in \mathbf{R}^n = [Ed_{u,1}; Ed_{u,2}; ...; Ed_{u,N_s}] \tag{1}$$

Second, we identify the training scenario $tr_{MIN,u}$ closest to each scenario S_u. Third, we consider the distance value $Ed_{MIN,u}$ associated with the training scenario $tr_{MIN,u}$.

In our setting, we suppose for convention that the criterion f_{AL} has to be minimized.

$$f_{AL}(u) = -Ed_{MIN,u} \tag{2}$$

Finally, Select and label the N_u most distant samples. The whole procedure is iterated until the predefined stopping condition is satisfied.

The surrogate model that we have trained is a L1-regularized multivariate vector autoregressive model *(VAR)* which is is one of the most commonly applied tools to handle dependencies among multiple random processes for forecasting purposes [4]. In a *VAR* model, each variable has an equation explaining its evolution based on its own lagged values, the lagged values of the other model variables, and an error term. L1-regularization was applied to the *VAR* loss function during the training.

4 Experiments

Our experiments are conducted on data collected from a 3D numerical simulation based on ekate model developed specifically for process-oriented computational simulations of shield tunnelling processes at the Institute for Structural Mechanics of Ruhr University Bochum [5]. The simulation produces time series of surface settlement for 154 monitoring soil points. We consider 20 initial labelled scenarios and the remaining 130 scenarios as unlabelled. We add 5 scenarios at each iteration. 10 scenarios are kept for the test set. The performance of our surrogate model on the test set is evaluated at a given time step using the same error measure used in [5].

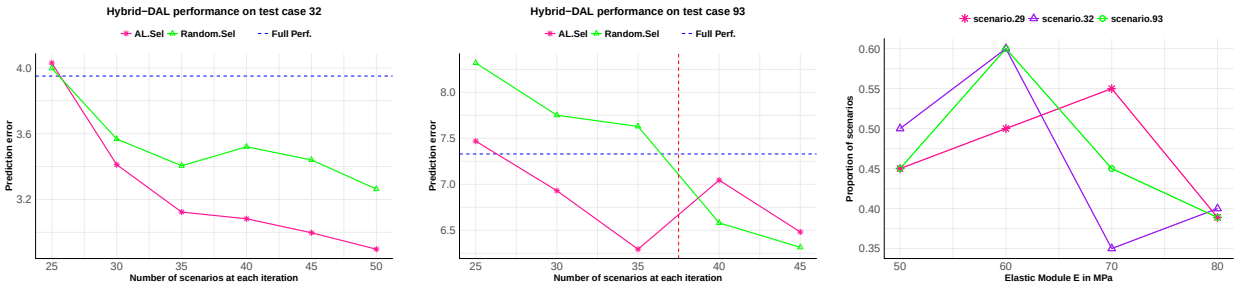


Figure 1: Illustration of the performance of Hybrid-DAL framework

As it is shown in Figure1, Our framework is able to reduce the number of required simulation scenarios up to 66% , improve the prediction accuracy up to 25% and identify sensible bounds for some input parameters, such as the Elastic Module $E(50MPa - 60MPa)$.

5 Future work

Within the same topic, we plan the use Generative Adverserial Networks (GANs) for realistic generation of industrial time series processes using initial samples from the original simulations. We will also combine active learning with (GANs) to enhance the quality of generated samples. In addition, we will investigate further active learning for feature selection.

References

[1] Cohn, David A., Zoubin Ghahramani, and Michael I. Jordan. "Active learning with statistical models." Journal of artificial intelligence research (1996).

[2] Settles, Burr. "Active learning literature survey." University of Wisconsin, Madison 52.55-66 (2010): 11.

[3] Douak, Fouzi, Farid Melgani, and Nabil Benoudjit. "Kernel ridge regression with active learning for wind speed prediction." Applied energy 103 (2013): 328-340.

[4] Tsay, Ruey S. Multivariate Time Series Analysis: with R and financial applications. John Wiley & Sons (2013).

[5] Cao, Ba-Trung, Steffen Freitag, and Günther Meschke. "A hybrid RNN-GPOD surrogate model for real-time settlement predictions in mechanised tunnelling." Advanced Modeling and Simulation in Engineering Sciences 3.1 (2016): 5.

Conceptual framework for control strategies in interlinked manufacturing processes based on quality predictions

Jacqueline Schmitt
Institute of Production Systems
TU Dortmund University
jacqueline.schmitt@ips.tu-dortmund.de

This report presents a conceptual framework on how to develop control strategies for interlinked manufacturing processes based on quality prediction. Various methods like finite element simulation, causal networks, and mathematical optimization have to be deployed in order to create a framework which allows to derive optimal control decisions based on quality predictions.

1 Introduction

Within the Collaborative Research Center 876 the subproject B3 focuses on the analysis of sensor data using data mining methods under real-time constraints. Learning methods have been developed for realtime quality predictions which now have to be integrated into advanced manufacturing process control. The developed methods have been applied to a hot rolling process in the steel industry. Based on various process parameters, e.g. rolling force, the anticipated final quality of the steel bars is predicted between intermediate production stages [1]. The label represents a quality measure, e.g. tensile strength. The prediction enables an early alignment of the predicted quality and the stipulated quality requirements. In case of unsatisfied quality requirements, control decisions on discharging the product or adapting impending process parameters can be made at an early stage of the process chain to prevent additional added value to defective products. In addition, those timely control decisions allow for a reduction of waste of energy and other resources [3].

2 Conceptual framework for control strategies

The underlying hypotheses implies that the anticipated final quality of a steel bar can be improved through adaptions of process parameters of impending production steps. Therefore, feasible parameter adjustments and appropriate combinations must be known in order to derive the right control decisions. The anticipated conceptual framework mainly consists of the following three steps:

1. First, a data set of process parameter value combinations and resulting quality has to be established. Due to the deficient availability of real process data in the considered hot rolling process, material forming simulation is used to generate data.
2. Secondly, the causalities of the system are represented by a causal network. Based on the simulation data causal influences of process parameters on the resulting product quality and inbetween different process parameters are identified.
3. Lastly, a mathematical optimization model is developed in order to derive cost-optimal control decisions on whether to remove defective products from the process chain or adapt impending process parameters to meet final quality requirements.

Material forming FEM simulation
As the real process is quite costly, the experiments to generate a respective data set of parameter value combinations have to be performed using material forming simulations. These simulations are based on the finite element method (FEM) and allow to simulate even marginal parameter values and combinations. FEM divides the solution space into smaller subspaces so that highly complex continuous problems can be reduced to simpler ones by discretization [4]. However, the remaining mathematical complexity is still high so that the FEM simulation requires high computational time to find approximate solutions numerically. Therefore, the simulation cannot be integrated into an online process control system directly and has to be conducted offline. Instead, causal influences of process parameters on product quality parameters will be depicted in a causal network.

Concept of causal networks
A causal network is a directed acyclic graph, consisting of nodes $\{X_1, ..., X_n\}$ that are connected by arcs. The acyclicity implies, that there are no feed-back loops in the system which means in mathematical terms that there is no direct path $X_i \rightarrow ... \rightarrow X_j$ such that $X_i = X_j$. The nodes represent random variables. In the underlying example of process control in the hot rolling process, the nodes $X_i, i = 1, ..., k$ represent process variables and the nodes $X_i, i = k + 1, ..., n$ represent the product quality variables. X_i is called a parent of X_j if there is a direct path $X_i \rightarrow X_j$. The arc can be interpreted as that X_i has a direct causal influence on X_j. For each node X_i with parents $\{Pa_1(X_i), ..., PA_{m_i}(X_i)\} \subset \{X_1, ..., X_n\}$, the effects that the parents have on the node can be quantified by the conditional probability distribution $P(X_i \mid Pa_1(X_i), ..., Pa_{m_i}(X_i))$. [2] Figure 1 shows an example of a causal network with process parameters $X_i, i = 1, ..., 4$ and product quality

measure X_5. The network implies that for example X_1 and X_2 causally influence X_3 which in turn can influence the product quality X_5 whereas X_4 has no causal influence on the product quality.

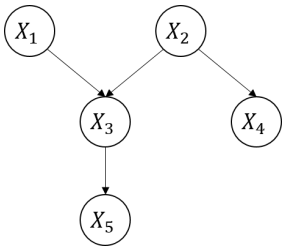


Figure 1: Example of a causal network

To derive the causal network based on the simulation data of the rolling process the following procedure must be performed [2]:

1. **Variable selection:** In the complex hot rolling processes multiple process parameters are recorded and assumed to influence the product quality. Since only a specific quality problem is studied, a subset of parameters is selected based on domain knowledge.

2. **Continuous variable discretization:** For most algorithms all variables must be discrete so that continuously recorded sensor data have to be discretized using specific discretization methods.

3. **Dimension reduction of the state space:** In order to increase computational efficiency domain knowledge is used to reduce the dimension of the state space.

4. **Temporal order of variables:** Since the rolling process consists of multiple production stages, a temporal order of variables is induced by their affiliation to a certain process stage. This information can be used to further constrain the model search and reduce computational complexity.

5. **Engineering-specified adjacencies:** In the hot rolling process there are variables that depend on each other e.g. the rolling force and rotation speed always have natural boundaries for a certain material or temperature of the steel bar.

6. **Identification of structural zeros:** When analyzing parameter combinations using a contingency table cells with zero counts can imply "sampling zeros" due to an insufficient sample size or "structural zeros" due to combinations of variable states that are impossible, e.g. a certain type of material can only be processed at a certain temperature.

Derivation of optimal control decisions by mathematical optimization

After causal factors of the rolling process chain have been identified and mapped in a causal network, optimal control decisions can be derived from a cost-based mathematical optimization model which ensures compliance of final quality requirements at lowest possible costs.

The objective function of the basic model aims to minimize occuring costs due to control decisions made upon the quality prediction:

$$\min x \cdot \sum_{j>k}^{n} \left(y(X_j) \cdot C^{Adapt}(X_j) \right) + (1 - x) \cdot C^{Scrap} \tag{1}$$

where $C^{Adapt}(X_j)$ denotes the costs for adapting the process parameter X_j and C^{Scrap} denotes the costs for removing a steel bar from the process chain. The decision variable $x \in \{0, 1\}$ denotes whether to remove ($x = 0$) a steel bar from the process chain or to continue processing ($x = 1$). The decision variable $y(X_j) \in \{0, 1\}$ denotes if the value of parameter X_j should be adapted ($y(X_j) = 1$) or not ($y(X_j) = 0$).
The constraints of the model ensure that if the processing is continued, final quality requirements are met and process parameter values are within their natural boundaries.

3 Conclusion

An effective process control of complex manufacturing processes, e.g. hot rolling of steel bars, requires the deployment of multiple sometimes complex methods like FEM simulation, causal networks and mathematical optimization. The implied complexity can be managed and reduced by applying structural concepts and splitting the complex overall problem into smaller, less complex partial problems and solve them successively.

References

[1] Benedikt Konrad, Daniel Lieber, and Jochen Deuse. Striving for zero defect production: Intelligent manufacturing control through data mining in continous rolling mill processes. In Katja Windt, editor, *Robust Manufacturing Control*, volume 1 of *Lecture Notes in Production Engineering*. Springer, 2012.

[2] Jing Li and Jianjun Shi. Knowledge discovery from observational data for process control using causal bayesian networks. *IIE Transactions*, 39(6):681–690, 2007.

[3] Daniel Lieber, Benedikt Konrad, Jochen Deuse, Marco Stolpe, and Katharina Morik. Sustainable interlinked manufacturing processes through real-time quality prediction. In *Leveraging Technology for a Sustainable World*, Proceedings of the 19th CIRP Conference on Life Cycle Engineering, pages 393–398. Springer, 2012.

[4] Robert L. Taylor, Jianzhoug Zhu, and Olgierd C. Zienkiewicz. The finite element method: Its basis and fundamentals. 2005.

Collaborative Research Center SFB 876 - Relieving bottlenecks in production processes by means of machine learning

Mario Wiegand
Institut für Produktionssysteme
Technische Universität Dortmund
mario.wiegand@ips.tu-dortmund.de

Final quality testing at the end of the production line is commonly used in manufacturing companies to guarantee high quality products. Due to long test times End-of-Line-testing (EoL-testing) often becomes the bottleneck of the process chain. In this report we present a strategy to abbreviate the time consuming process of EoL-testing by machine learning. In order to train well performing prediction models, we have to cope with highly imbalanced data.

1 Introduction

Within the Collaborative Research Center 876 project B3 deals with the time-constrained analysis of sensor data using machine learning techniques. So far the development and application of learning algorithms as well as process control strategies focused on a hot rolling mill in steel industry [2–5]. In high-volume production of injectors EoL-testing is a very time consuming process because every injector is tested under various conditions in a variety of testing points. Thus, EoL-testing is limiting production performance, i.e. it is reducing the output of the process chain. In production management a process limiting the capacity of the process chain is called bottleneck.

Traditional methods of production planning and scheduling are not able to further reduce test time. Therefore, new strategies are required to increase the process' ouput and improve productivity.

2 Strategies for test time reduction and first results

The manufacturing process of injectors consists of various steps including laser drilling of the nozzle as well as assembly of injector body, actuators, electronic control unit and injector nozzle. After manufacturing the injectors' characteristics and functionality are checked in EoL-testing. Depending on the test program, the process time of EoL-testing is 5 to 6 times the cycle time of the production line. That is why, more than one test station is integrated into the production line to ensure the required output. For every injector a type-specific test program is conducted consisting of different features that are measured under specific conditions in several testing points. [1] Figure 1 shows an overview of the production line.

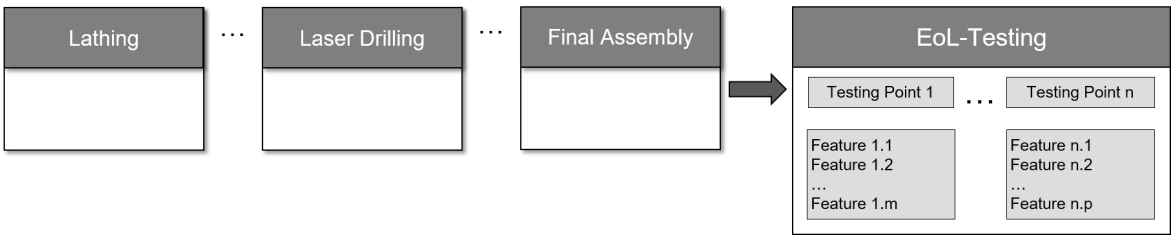


Figure 1: Manufacturing process of injectors

Within the highly automated manufacturing process as well as the EoL-test station a huge amount of data is recorded by various sensors and stored in a manufacturing execution system (MES). The analysis of this data by means of machine learning offers new opportunities for test time reduction, in order to relieve the process' bottleneck and increase output and productivity. Different strategies for reducing test time by applying prediction models can be derived [1]:

- One strategy is to predict the features of testing points based on features of testing points recorded upstream in the EoL-testing. If a high overall accuracy is achieved, subsequent testing points can be substituted by predicted values. It is also possible to decide on substituting subsequent testing points only for some of the products depending on the model's confidence. Whenever a previously defined threshold of the confidence is exceeded, the physical measurement of the predicted features can be skipped.
- Another strategy is to train a prediction model with no or low false negative rate. As a result, only products that are classified NOK have to be physically tested at the considered test points. If a product is classified OK, no feature measurements are necessary.
- The aforementioned strategies refer to the prediction of testing features based on testing features measured at previous testing points in the EoL-testing. In addition

to the information from quality testing, process data from manufacturing steps can be used as input for training to improve prediction accuracy. Furthermore, as a consequence process control decisions can be made earlier during manufacturing, so that defective products could be ejected early from process chain.

The success of each strategy depends on the available data and the models' predictive abilities. In a first step we train prediction models only based on data from EoL-testing and try to minimize the false negative rate. The data set consists of 217,632 injectors described by 120 features [1]. Due to a huge amount of missing values we deleted 39 sparse features from the data set. Only a small proportion of NOK products is included in the data leading to a highly imbalanced class distribution. To adequately cope with the class imbalance we apply different strategies like undersampling of the majority class and MetaCost-learning. MetaCost-learning allows to define a problem specific cost function that we used for penalizing false negatives overproportionally. In addition to this, the false negative rate respectively the class recall serves as primary performance criterion because we are interested in detecting all defect products, in order to only deliver high quality products to the customer.

For model training several supervised learning algorithms like Naive Bayes, Decision Trees, Random Forest and Gradient Boosted Trees are applied. Some of the algorithms are combined with MetaCost-learning. Besides, we apply the different algorithms with and without undersampling. Table 1 shows an extract of the results [1].

Table 1: Performance of different learning algorithms

Learning Method	Accuracy	Recall	Precision
Naive Bayes	96.15	33.83	61.65
Decision Tree (with MetaCost)	97.00	44.20	78.38
Random Forest	98.03	61.93	90.46
Gradient Boosted Tree	98.20	67.38	90.77
Decision Tree with undersampling	97.30	80.65	53.75
Gradient Boosted Tree with undersampling	97.74	81.20	72.13

Referring to the overall accuracy all learning algorithms provide similar performance. The high accuracy mainly originates from the heavily imbalanced data set. Considering recall as primary performance criterion the ensemble learning strategies Random Forest and Gradient Boosted Trees perform much better than Naive Bayes and Decision Trees. However, the biggest performance boost can be achieved by balancing the data using undersampling.

3 Future Work

The results achieved in the first experiments are not sufficient for substituting parts of EoL-testing with predicted values in the real process. Nevertheless, the results could serve as a basis for further research. Therefore, in future work we want to improve the results by expanding data preprocessing by the extraction and selection of relevant features. Additionally, we will test further methods to cope with the high class imbalance and also focus on the other aforementioned control strategies for relieving the bottleneck.

References

[1] Jochen Deuse, Jacqueline Schmitt, Marco Stolpe, Mario Wiegand, and Katharina Morik. Qualitätsprognosen zur engpassentlastung in der injektorfertigung unter einsatz von data mining. In *Schriftenreihe der Wissenschaftlichen Gesellschaft für Arbeits- und Betriebsorganisation (WGAB) e.V.*, 2017.

[2] Michel Eickelmann, Mario Wiegand, Benedikt Konrad, and Jochen Deuse. Die bedeutung von data mining im kontext von industrie 4.0. *Zeitschrift für wirtschaftlichen Fabrikbetrieb (ZWF)*, 110(11):738–743, 2015.

[3] Daniel Lieber, Marco Stolpe, Benedikt Konrad, Jochen Deuse, and Katharina Morik. Quality prediction in interlinked manufacturing processes based on supervised and unsupervised machine learning. In *Procedia CIRP - 46th CIRP Conf. on Manufacturing Systems*, volume 7, pages 193–198. Elsevier, 2013.

[4] Marco Stolpe. *Distributed Analysis of Vertically Partitioned Sensor Measurements under Communication Constraints*. PhD thesis, TU Dortmund University, Dortmund, 2017.

[5] Mario Wiegand, Marco Stolpe, Jochen Deuse, and Katharina Morik. Prädiktive prozessüberwachung auf basis verteilt erfasster sensordaten. *at-Automatisierungstechnik*, 64(7):521–533, 2016.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B4

Analysis and Communication for Dynamic Traffic Prognosis

Kristian Kersting Michael Schreckenberg
Christian Wietfeld

Analytical Calculation of Capacity-related Quantities of the Railway Operations Research

Merlin Becker
Physics of Transport and Traffic
Universität Duisburg-Essen
merlin.becker@uni-due.de

In urban public transport networks, the calculation of transportation times is a common question. In this report, a solution for the calculation of railway running times is presented. Due to its analytical nature, the solution is fast and precise.

1 Introduction

In the field of railway operations research there is ongoing research on the efficient calculation of running time related quantities. As the most approaches utilize rudimentary approximate calculation methods to determine the results, they are neither fast nor precise [1].

Our presented method aims at this weakpoint. As a new contribution, we solved the integral of the acceleration process analytically. It follows that the whole calculation process can be expressed in an analytical form. It is presented in the following. The different modes of motion are graphically shown in fig. 2.

2 Analytical Formulae

The acceleration of a train can be described by a second order polynomia. In fig. 1 the typical behavoir of the formula is shown.

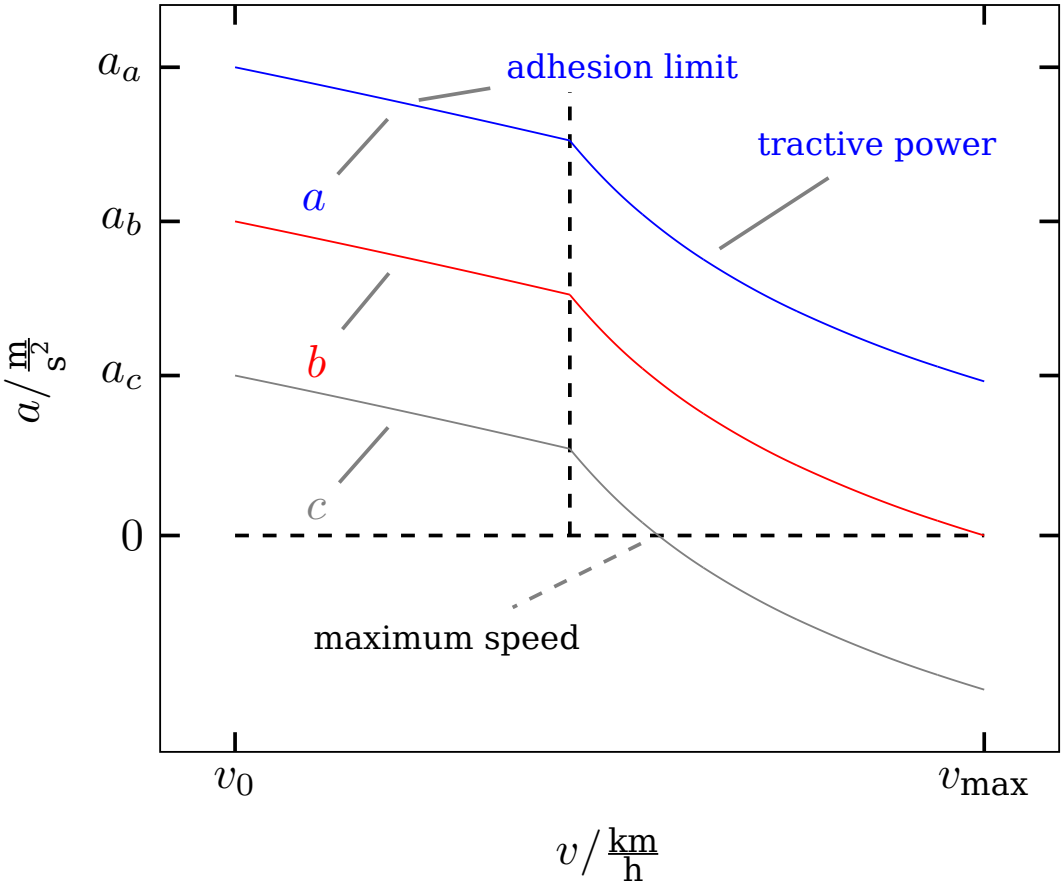
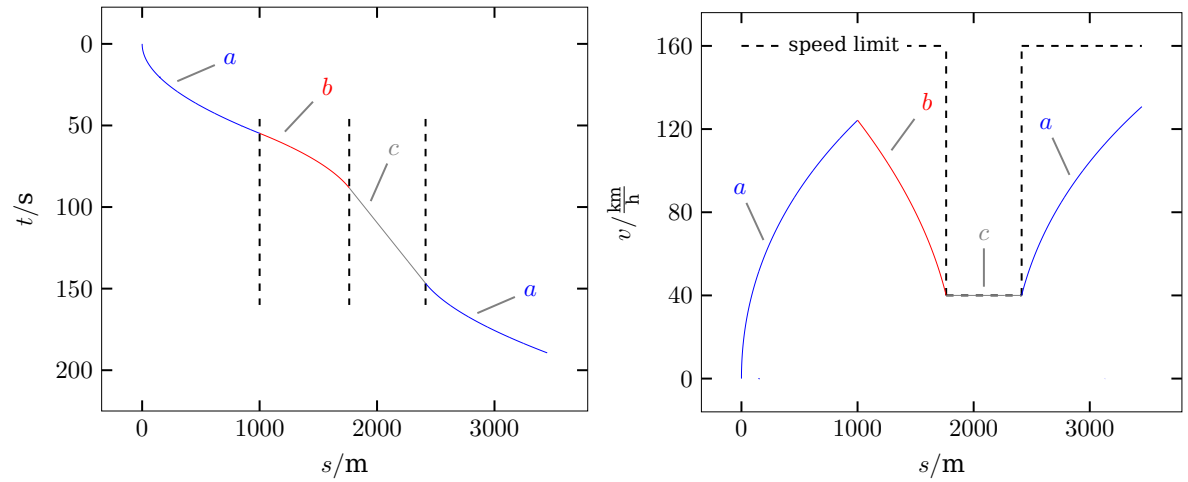


Figure 1: Acceleration-velocity-diagram. The acceleration-velocity-diagram contains information on the acceleration process. As the effective acceleration is limited by different power limits, it decreases with a higher speed driven. The different curves represent the effective acceleration under the consideration of different gradients. Gradients can limit the maximum speed.

$$a(v) = c_0 + c_1 v + c_2 v^2. \tag{1}$$

c_0, c_1 and c_2 are train-specific constants which depend on the chosen train. Under the consideration that these parameters are real numbers, the integrals for the time and the distance in dependance on the start and end speed driven, can be calculated. Note, that the constants are not necessarily constant for the whole acceleration process.

$$t = \frac{2}{\sqrt{4c_0c_2 - c_1^2}} \arctan\left(\frac{c_1 + 2c_2 v}{\sqrt{4c_0c_2 - c_1^2}}\right)\Bigg|_{v(t_0)}^{v(t)} + t_0. \tag{2}$$



(a) Time-distance-diagram.

(b) Velocity-distance-diagram.

Figure 2: Comparison between the time-distance-diagram and the velocity-distance-diagram. In the diagrams different modes of motion are shown: acceleration (a), braking (b) and cruising (c). To give a better insight to the representation of the time-distance-diagram, the velocity-distance-diagram is also shown.

$$s(v) = \left. \frac{-\frac{2}{\sqrt{4c_0c_2-c_1^2}} \arctan\left(\frac{c_1+2c_2v}{\sqrt{4c_0c_2-c_1^2}}\right)}{2c_2} \right|_{v(t_0)}^{v(t)} + \left. \frac{\log(c_0 + v(c_1 + c_2v))}{2c_2} \right|_{v(t_0)}^{v(t)} + s(t_0). \tag{3}$$

The cruising mode ist described via the equation for constant-motion:

$$s(t) = v \cdot (t - t_0) + s(t_0). \tag{4}$$

Braking is calculated with the equation of accelerated motion as the braking acceleration is assumed constant.

$$t = \int_{v(t_0)}^{v(t)} \frac{1}{a}\,\mathrm{d}v + t_0 \tag{5}$$

$$s(t) = \int_{t_0}^{t} at\,\mathrm{d}t + s(t_0). \tag{6}$$

3 Calculation

Additionally to the formulae presented in the last section, calculation rules are needed to calculate the running time properly.

- Braking. A train has to obey speed limits, so braking has to be done on time.
- Acceleration. If a train needs not to brake, it accelerates until the speed limit is reached.
- Cruising. If a train does not need to brake or accelerate, it cruises.

4 Conclusion & Outlook

The presented solution for the calculation of running times is fast and precise in terms of prediction. It helps to develop high-efficient railway simulation frameworks. These frameworks are a starting point for new railway operations simulations, capacity calculation methods or tools for energy-efficient driving.

Through the clear structure of the algorithm, additional components can be developed easliy. Generic applications will be shown in future publications.

References

[1] Merlin Becker, Lars Habel, and Michael Schreckenberg. Analytical method for the precise and fast prediction of railway running times. In *Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017 5th IEEE International Conference on*, pages 152–157. IEEE, 2017.

Highly Available LTE Communications through National Roaming

Stefan Monhof
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
stefan.monhof@tu-dortmund.de

Mobile network outages, coverage gaps and fluctuating reception quality are great challenges for critical applications that demand high network availability and predictable Quality of Service (QoS). In this report, results of a long-term measurement series that analyzes the availability and fluctuating quality of commercial Long Term Evolution (LTE) networks are presented. These results show how the use of multiple mobile networks by means of *national roaming* can improve the reliability and quality of communications.

1 Introduction

Commercial mobile network operators advertise with the high and ever-growing coverage of their networks. But in reality, customer often struggle with coverage gaps or bad and dropping connections. Also, reception quality is not static, but fluctuates over time, even at a single location. A solution to handle coverage gaps and fluctuating network conditions, and therefore to improve the communications quality, is to share infrastructure among multiple operators [2]. From a user perspective this can be achieved by means of national roaming [3]. *Roaming* means that a device connects to the Radio Access Network (RAN) of a different operator (visited operator) than the issuer of its Subscriber Identiy Module (SIM) card (home operator). Usually, it is used to get telephony or data services in foreign countries. A user is not required to have a contract with the visited operator, but home and visited operator need to have a *roaming agreement*, which defines the term under which roaming is allowed. Within a country the operators commonly do not have roaming agreements among each other. The term *national roaming* refers to roaming within the home country. Technically it does not differ from roaming in foreign

countries and, due to the missing roaming agreements, it is often realized using foreign SIM cards. The benefit of national roaming is that one SIM card (and therefore modem) is sufficient to use the networks of more than one operator. This may be leveraged to mitigate the effects of outages or coverage gaps.

To evaluate if current, commercial Long Term Evolution (LTE) networks are suited to fulfill the strict requirements of critical applications, a long-term measurement series was performed. In this series the availability and received signal strength of the three commercial German LTE networks were monitored over a period of six months. The results presented in this report show how infrastructure sharing by means of national roaming can increase the availability and quality of LTE communications.

2 Measurement Device and Approach

To evaluate the mobile network coverage and quality a custom measurement device, called *Mobile Network Analyzer (MNA)*, was developed (Figure 1) It is based on a Banana Pi M2+ embedded PC, equipped with a Huawei ME909s-120 LTE module, which allow collecting more detailed information and more precise control than common off-the-shelf smartphones. For localization, the MNA can be equipped with Global Positioning System (GPS) and a barometric pressure sensor. The software is developed in the Python programming language and runs on the Linux distribution armbian.

Figure 1: Mobile Network Analyzer (MNA)

A *network scan* can be performed to obtain knowledge about the mobile networks in range. For LTE a network scan lists all LTE cells in range with details on the operator (Public Land Mobile Network (PLMN) identifier), cell identifier, Physical Cell Identifier (PCI), Tracking Area Code (TAC) and the Reference Signal Received Power (RSRP). The RSRP is the linear average of the power received on resource elements that contain cell-specific reference signals [1]. Basically, it is a path loss measurement and therefore not influenced by noise or intra-cell interference. Measurements showed that a value of about −95 dBm or higher is usually sufficient to establish a stable connection. Below of −110 dBm the connection is expected to be instable and drop. With the used LTE module, available cells with an RSRP equal or greater than −110 dBm can be identified.

The availability of the LTE networks of the three German mobile network operators Telekom, Vodafone and Telefónica was monitored from April to September 2017 in a wind

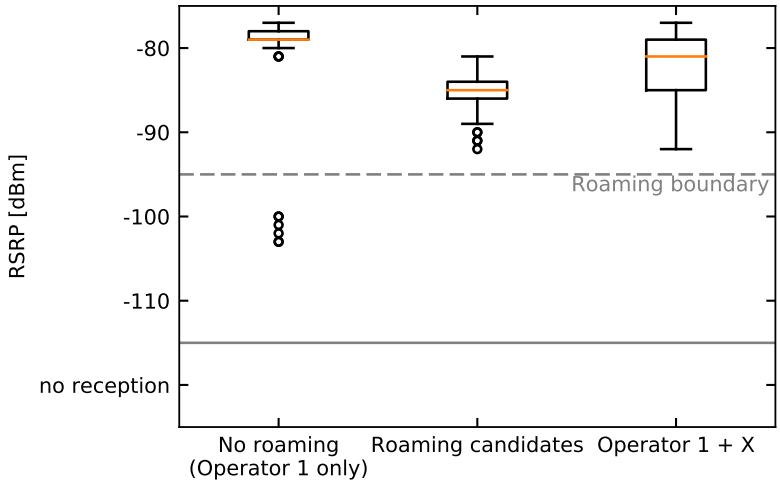


Figure 2: Empirical RSRPs of the strongest cells at location 1 for operator 1, operators 2 and 3 combined and all operators combined

farm in North Rhine-Westphalia, Germany. In this time a network scan was performed each hour at three different (static) locations in the wind farm. For this purpose, MNAs were installed at the foot of three different wind turbine generators spread over the wind farm area (345 ha). Since the wind farm is located on a recultivated area of a lignite surface mine, it is supposed that no detailed network planning was performed for this location. This results in partially harsh cell-edge network conditions.

3 Exemplary Results

Figure 2 shows the RSRP variation of the strongest received cell of operator 1 at location 1 (no roaming). With a median of −79 dBm for the strongest cell, the reception of operator 1 is good in this location. But the outliers show that there were some outages of the strongest cell and the second strongest cell of operator 1 has a RSRP around −101 dBm. To ensure a stable connection, cells with an RSRP of −95 dBm (*roaming boundary*) or higher should be preferred, even though this particular outages could be completely handled within the network of operator 1. The *roaming candidates* are the cells of the other operators that could be used in case of an outage of operator 1. *Operator 1 + X* shows the RSRPs of the cells of all operators that are available through national roaming and lie above the roaming boundary. So, while the overall availability of operator 1 was 100% at location 1, cell outages may have degraded the communication service. The alternative cell of operator 1 has an RSRP value in a range that potentially does not allow a stable connection. In these cases, national roaming could be used to obtain a stable connection with a higher probability.

In Figure 3 the RSRP of the strongest cell of operator 3 is shown. In this case, cell outages lead to RSRPs below −110 dBm (no reception). The roaming candidates (available in

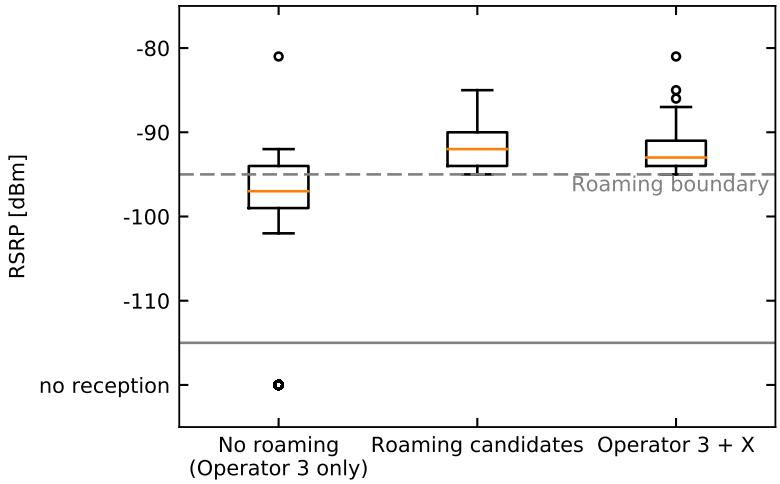


Figure 3: Empirical RSRPs of the strongest cells at location 2 for operator 3, operators 1 and 2 combined and all operators combined

99.6 % of the time) from the other operators can help to close the gap and increase the overall availability of operator 3 from 82 % (or 34 % above the roaming boundary) to 99.8 %.

4 Conclusion

As shown, national roaming could help to improve the availability of LTE communication for critical applications by enabling the use of different mobile network operators. Cell outages (in static applications) or coverage gaps (in mobile scenarios) could be handed this way and therefore the overall availability increases. Also, the quality of connections could be improved by switching to the best available cell from an arbitrary operator.

References

[1] 3GPP. Evolved universal terrestrial radio access (E-UTRA); physical layer; measurements. TS 36.214 version 13.4.0 Release 13, 3rd Generation Partnership Project (3GPP), January 2017.

[2] Jacek Kibilda, Nicholas J Kaminski, and Luiz A DaSilva. Radio access network and spectrum sharing in mobile networks: A stochastic geometry perspective. *IEEE Transactions on Wireless Communications*, 2017.

[3] Roya H Tehrani, Seiamak Vahid, Dionysia Triantafyllopoulou, Haeyoung Lee, and Klaus Moessner. Licensed spectrum sharing schemes for mobile operators: A survey and outlook. *IEEE Communications Surveys & Tutorials*, 18(4):2591–2623, 2016.

Joint Simulation of Vehicular Mobility and Communication

Benjamin Sliwa
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
benjamin.sliwa@tu-dortmund.de

Upcoming Intelligent Transportation Systems (ITSs) are confronted with the converging mobility and communication in the context of automated vehicles and intelligent traffic control. Consequently, novel simulation methods are required for the accurate evaluation of the whole system instead of treating the individual aspects separately. In this report, a lightweight framework for simulating vehicular mobility is presented. In contrast to existing approaches, it can be embedded into a network simulator to enable joint simulation of mobility and communication on a shared-codebase level. With this method, synergies of the different worlds can easily be exploited, e.g. for realizing context-aware vehicular applications.

1 Solution Approach

The proposed simulator Lightweight ICT-centric Mobility Simulation (LIMoSim) consists of the mobility simulation kernel and a graphical user interface. While the kernel is independent on the other part and intended to be integrated into a network simulation framework, the latter can be used for standalone development of mobility models. LIMoSim supports generic scenario modeling as well as real-world map data obtained from OpenStreetMap (OSM). The the architecture of the whole simulation framework is further explained in [1]. For a detailed description about how LIMoSim is embedded into the network simulator Objective Modular Network Testbed in C++ (OMNeT++), including event synchronization, consider [2]. Fig. 1 provides an overview about the relevant components used for the mobility simulation. A hierarchical model is used for

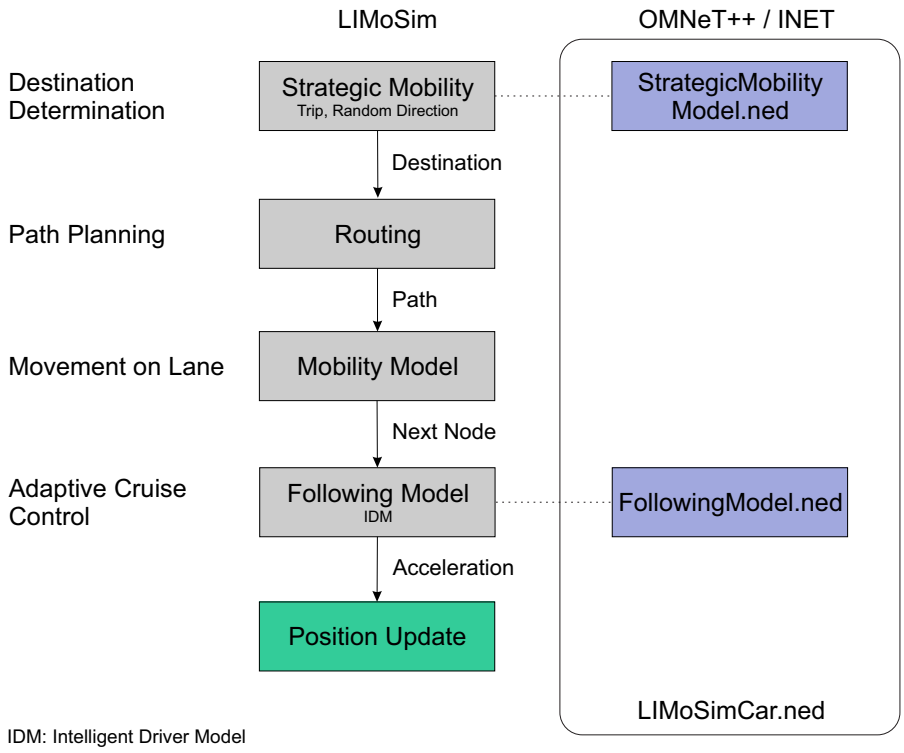


Figure 1: Architecture of the mobility simulation within LIMoSim and OMNeT++

representing the different kinds of logical processes. On the highest layer, the strategic model represents the human decision making and the reason for the movement. Once a destination has been determined, the routing path is computed and the car follows the lanes to reach its destination. While moving, it follows the traffic rules and controls its acceleration with respect to other traffic participants using the Intelligent Driver Model (IDM). Since the simulation kernel of LIMoSim is directly embedded into the source code of the INET framework of OMNeT++, it can be used as a mobility module in a transparent way by all further extension frameworks (e.g. SimuLTE for simulating Long Term Evolution (LTE) networks). In contrast to existing approaches that couple mobility and communication by the use of multiple specialized simulators through means of Interprocess Communication (IPC), here both components have direct access to all available information. The approach aims to support the exploitation of synergies (e.g. integrating trajectory information for intelligent cellular handover mechanisms) in the context of converging mobility and communication in ITS.

2 Simulative Performance Evaluation

For the simulative evaluation, we used LIMoSim with OMNeT++ and the LTE simulation framework SimuLTE. Real-world map data from OSM is used to model the campus area

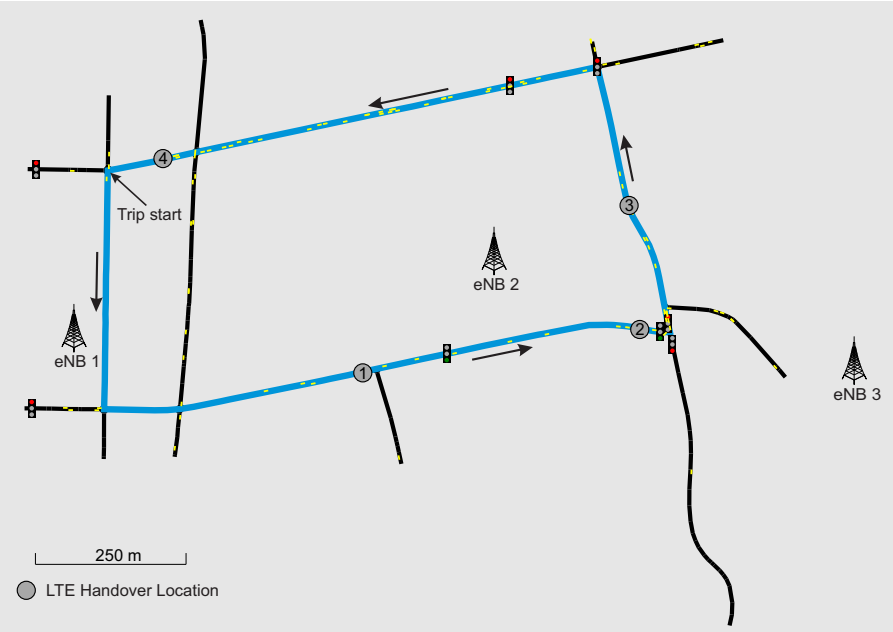


Figure 2: Scenario used for the performance evaluation in the campus area of the TU Dortmund University.

of the TU Dortmund University. A single LTE-enabled car moves along the surrounding streets as shown in Fig. 2. Multiple traffic signals are used to control the traffic flow and 100 other cars act as interference traffic. Three different Evolved Node B (eNB) are positioned according to the position of their real-world equivalent. Fig. 3 shows the temporal behavior of the velocity, acceleration and the Received Signal Strength Indicator (RSSI) for the considered car moving on the campus area of the TU Dortmund University. Due to the presence of other traffic participants, the velocity changes dynamically as braking and acceleration operations are performed often. The resulting behavior is typical for suburban traffic scenarios. Due to the mobility, the quality of the communication channel is frequently changing. In the performed simulation, the User Equipment (UE) performs a handover from eNB2 to eNB3 and shortly afterwards back to eNB2. Considering the handover locations shown in the map, it can be concluded that integrating the trajectory information into the actual handover process could have been beneficial in order to avoid this unnecessary interruption of the data transmission.

3 Conclusion and Further Research

The developed simulation framework will serve as a platform for the analysis of context-aware applications in future work. In the context of cellular and local Vehicle-to-Everything (V2X) communication, this includes mobility-aware handover and packet forwarding mechanisms as well as the predictive steering of pencil beams for Millimeter Wave (mmWave)

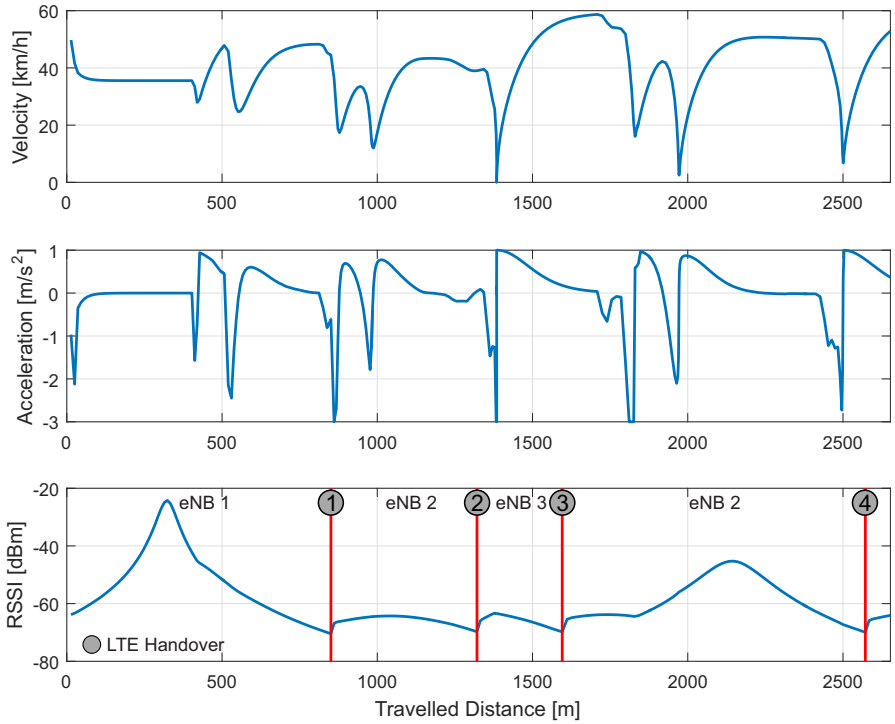


Figure 3: Performance evaluation using real-world map data.

communication scenarios. In contrast to existing approaches, LIMoSim provides the simulation of vehicular mobility as a service for communication networks without binding it to a specific communication technology. Consequently, it is well-prepared for being used in combination with future technologies, once simulation models of the latter are available.

References

[1] Benjamin Sliwa, Johannes Pillmann, Fabian Eckermann, Lars Habel, Michael Schreckenberg, and Christian Wietfeld. Lightweight joint simulation of vehicular mobility and communication with LIMoSim. In *IEEE Vehicular Networking Conference (VNC)*, Torino, Italy, Nov 2017.

[2] Benjamin Sliwa, Johannes Pillmann, Fabian Eckermann, and Christian Wietfeld. LIMoSim: A lightweight and integrated approach for simulating vehicular mobility with OMNeT++. In *4th OMNeT++ Community Summit (OMNeT++ 2017)*, Bremen, Germany, Sep 2017.

Aplication of the breakdown minimization principle to minimize traffic congestions in inner city networks

Tim Vranken
Physik von Transport und Verkehr
Universität Duisburg-Essen
tim.vranken@uni-due.de

This report sums up our recent research on applying the "breakdown minimization principle" [2] in comparison to "Wardrop's user Equilibrium" [4] to avoid traffic congestions in inner city systems. To this end, we created a set of rules based on [6] (an agent based cellular automata model) with which we can simulate traffic in cities on a microscopic scale. We then apply those rules to a traffic system to compare the two routing methods and use these results to create a combination of both.

1 Introduction

The route finding methods for inner city networks represent a more complex problem than the route finding on highways since more restrictions apply (e.g. traffic lights) while there is also a greater degree of freedom (e.g. more crossroads and intersections). In the following work, we will look into two methods, namely the "breakdown minimization principle" [2] (from now on BMP) and "Wardrop's user Equilibrium" [4] (from now on WE). WE is the standard way for modern navigation systems to find the shortest route based on the current traffic situation. Before the BMP can be applied to traffic, information has to be gathered and a preselection of used routes has to be made. We will do these presimulation works in section 2. In section 3 we will then create a traffic system

in which we look into the application of the BMP in inner city networks and compare it to WE. In section 4 we combine WE and BMP to create an improved method.

2 Preinformation gathering

To use the BMP in a network three things have to be done before. Firstly, all "bottlenecks" have to be found. In the following work there are two kinds of bottlenecks. The first are lane ends and the second are traffic lights. Since we are creating the network in our simulation we have a completed understanding of it and know where every bottleneck is. Note that before a traffic light there can be up to three bottlenecks, i.e. a right turn lane, two straight leading lanes and a left turn lane.
In the second step each bottleneck will be assigned a value C^k_{min}in vehicles per hour. If the traffic flow in this bottleneck is greater than C^k_{min}, a breakdown according to the three-phase-theorie [1] can happen. In our tests we found in accordance to [3, 5] that C^k_{min}before a traffic light can be descried by:

$$C^k_{\text{min}} \approx q_{\text{sat}} \cdot \frac{G^{\text{eff}}}{Ap} \tag{1}$$

where G^{eff} is the effective green time, Ap is the Traffic light period and q_{sat} is the average number of vehicles (in $\frac{\text{vehicle}}{\text{h}}$) that can pass over one green traffic light. $G^{\text{eff}} = G - \delta t$ where G is the real green time of the traffic light and $\delta t \approx 3 - 4$ s is a delay that the traffic takes to react to the light chance. In our tests we found that $\delta t = 3.5$ s and $q_{\text{sat}} = 1682\ \frac{\text{vehicle}}{\text{h}}$ for straight leading traffic lights. The average traffic flow is reduced to $q_{\text{sat}} = 1469\ \frac{\text{vehicle}}{\text{h}}$ for right and left turn lanes since the average speed before a turn is lower than in free flow traffic. A special case is represented by right turning lanes that are separated from the intersection and have no traffic light. For these we found that a normal turn lane q_{sat} of 1469 $\frac{\text{Vehicle}}{\text{h}}$ can be used and G^{eff} can be described threw $G^{\text{eff}} = Ap - G_{\text{sum}} - \delta t$ with a δt of 4.8 and G_{sum} being sum of the effective green lights of all intercepting lanes that lead onto the same lane. For the second kind of bottleneck (the lane end) we find that $C^k_{\text{min}}$is approximated by:

$$C^k_{\text{min}} \approx 1000 + 1350 \cdot (N_s - 2)\ \frac{\text{Vehicle}}{\text{h}} \tag{2}$$

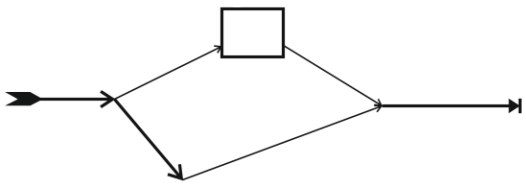
were N_s is the number of lanes before the lane end (e.g. $N_s = 3$ for a lane number chance of $3 \rightarrow 2$). The last thing to do before simulating traffic is to choose the routes that will be used. Since in the BMP routes are chosen so that no breakdowns are created, but the travel times of the routes is not considered, it can happen that routes which are way to long are being used. To prevent this we do a preselection where we only considerwith travel times that do not exceed the shortes way travel time by more than Δt. In section 4 we will see the importants of this preselection.

3 Comparison

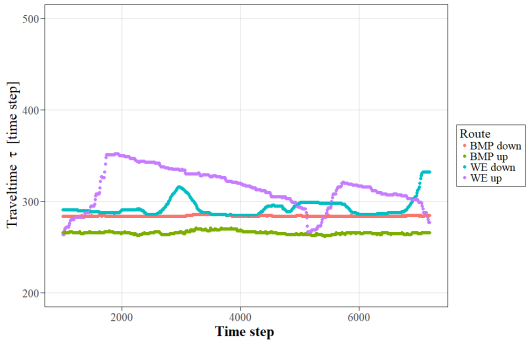
To compare BMP to WE we created multiple systems from which the simpleste can be seen in figure 1(a). Cars drive from left to right and have to decide whether they take the upper route and pass a traffic light or take the lower route over a lane end. According to WE the car will take the shorter route which will result in a congestion. This congestion increases until the other route becomes shorter and all cars choose this route. This will then create congestion on the alternative route until its longer again and everyone chooses the initial route etc.. For the BMP we create an equation system which has to be solved:

$$q^0 + q^1 = q_{00} \tag{3}$$
$$q^0 + s_0 + s = C^0_{\min} \tag{4}$$
$$q^1 + s_1 + s = C^1_{\min} \tag{5}$$
$$\max(s) \tag{6}$$

where q_{00} is the total flow from left to right and s is the difference between $C^k_{\min}$ and the flow q^k over the bottleneck k. In figure 1(b) We see the travel times for $q_{00} = 1800\ \frac{\text{Vehicle}}{\text{h}}$ for the BMP and WE. we clearly see the superiority of the BMP over WE when the traffic volume is high enough.



(a) Sketch of the System. The Arrow with a horizontal line at the end is a sink street. The one with a fletching is a source street. Arrows which are thicker represent streets with more Lanes and the square represents a intersetion with a traffic light.



(b) Traveltimes for WE and BMP for each of the two routes over 7000 time steps. The route passing the traffic light is called "up" and the one over the lane end is called "down"

4 Combination

The BMP shows clear superiority over WE when the traffic volume is high enough. However when the traffic volume is so low that no congestions are created the routes

taken in BMP are longer than the shortest route, which is solely used in WE. Applying BMP in systems with low traffic volume results in worse travel times compared to WE. Additionally, congestions which are created while using BMP are not considered in the equation system and thus they will only grow bigger when the traffic volume is high. To resolve these two and further problems, the BMP and WE are combined into a new method with a set of rules where the vehicle gets assigned a route based on the time of arrival. For every vehicle entering the system, it is checked whether the vehicle could create a flow q^k over the bottleneck k which would be grater than C^k_{min}in $\frac{\text{vehicle}}{\text{h}}$. Like this, each route is checked, starting from the shortest route (based on the current travel time) and the vehicle gets the shortest route assigned on which at no bottleneck $q^k > C^k_{\text{min}}$ holds true. Thus the idea of the BMP is still used while we also still choose the shortest route if the traffic volume is to low to create traffic congestion. We also reduce the time it takes to remove traffic congestion with this method.

References

[1] B. S. Kerner. *Introduction to Modern Traffic Flow Theory and Control: The Long Road to Three-Phase Traffic Theory*. Springer, 2009.

[2] B. S. Kerner. Optimum principle for a vehicular traffic network: minimum probability of congestion. *Journal of Physics A: Mathematical and Theoretical*, 44(9):092001, 2011.

[3] B. S. Kerner. Three-phase theory of city traffic: Moving synchronized flow patterns in under-saturated city traffic at signals. *Physica A: Statistical Mechanics and its Applications*, 397:76–110, 03 2014.

[4] B. S. Kerner. Breakdown minimization principle versus wardrop's equilibria for dynamic traffic assignment and control in traffic and transportation networks: A critical mini-review. *Physica A: Statistical Mechanics and its Applications*, 466, 09 2016.

[5] B. S. Kerner, S. L. Klenov, and M. Schreckenberg. Traffic breakdown at a signal: classical theory versus the three-phase theory of city traffic. *Journal of Statistical Mechanics: Theory and Experiment*, (3):03001, 2014.

[6] W. Knospe, L. Santen, A. Schadschneider, and M. Schreckenberg. Towards a realistic microscopic description of highway traffic. *Journal of Physics A: Mathematical and General*, 33(48):477–485, 2000.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C1

Feature selection in high dimensional data for risk prognosis in oncology

Sangkyun Lee Sven Rahmann
Alexander Schramm

Controlling the False Discovery Rate in Boolean Matrix Factorization

Sibylle Hess
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
sibylle.hess@tu-dortmund.de

Boolean matrix factorization (BMF) is a popular and powerful technique for inferring knowledge from binary data. The mining result is represented via a set of matrices, each determined by an outer product of two binary vectors. The false discovery rate (FDR) represents the expected proportion of erroneously rejected hypotheses in multiple testing problems. While FDR-based methods are successfully applied for supervised learning tasks, no approaches are generally available in the unsupervised setting. We propose and discuss the usage of FDR in Boolean matrix factorization. We prove two bounds on the probability that a found matrix covers only noise. Each bound exploits a specific property of the outer product—using new insights on the results of Boolean matrix factorization in the presence of noise. This leads to improved BMF algorithms by replacing heuristic rank selection techniques with a theoretically well-based approach.

Introduction The problem of finding patterns in a set of data can be addressed in various ways. Each method has its own characteristics, making it suitable for detecting specific types of pattern. A common drawback of most techniques is the absence of quality guarantees on the mining result. Whenever data is collected from an imperfect (noisy) channel—arising from tainted or inaccurate measurements, or transmission errors—the method of choice might be fooled by the noise, resulting in phantom patterns which actually don't exist in the data. The investigation of the *trustworthiness* of data mining techniques is important for their practical applicability. While some approaches for the supervised setting exist, e.g., significant pattern mining, insights for the completely unsupervised case are still missing.

Boolean matrix factorization (BMF) is a popular and powerful technique for inferring knowledge from data. A Boolean product $Y \odot X^\top$ of matrices $X \in \{0,1\}^{n \times r}$ and $Y \in \{0,1\}^{m \times r}$ is the disjunction of r matrices; each matrix is defined by the outer product $Y_{\cdot s} X_{\cdot s}^\top$ of the s-th column vectors. We refer to a pair of such column vectors $(X_{\cdot s}, Y_{\cdot s})$ as a *tile*. We assume that the data matrix $D \in \{0,1\}^{m \times n}$ is composed by a Boolean product and some noise, i.e.,

$$D = Y^* \odot X^{*\top} + N, \tag{1}$$

where $N \in \{-1, 0, 1\}^{m \times n}$ is the noise matrix and Y^* and X^* are the *true* factor matrices of unknown rank r^*. We often state N_+ for the positive part of the noise matrix. While state-of the art BMF techniques [1] employ rather complicated regularization terms to filter the structure from the noise, we want to move to theoretical estimates of trustworthyness.

Bounds on the FDR in BMF The first step towards trustworthy pattern mining is a measure of trustworthiness. The False Discovery Rate (FDR) is a simple yet powerful way to express the probability that something goes wrong:

Definition 1 (FDR)**.** *Given a finite set $\mathcal{H}$ of null hypotheses from which r are rejected. Let v denote the number of erroneously rejected null hypotheses. We say that* the FDR is controlled at level q *if*

$$\mathbb{E}\left(\frac{v}{r}\right) \leq q.$$

In our setting, a null hypothesis H_s^0 indicates that the tile $(X_{\cdot s}, Y_{\cdot s})$ approximates the noise matrix N and not $Y^* \odot X^{*\top}$. In other words, the hypothesis H_s^0 is true when all entries in $Y_{\cdot s} X_{\cdot s}^\top \circ Y^* X^{*\top}$ are actually zero. Bearing this in mind, we see that a BMF of rank r corresponds to a joint rejection of r null hypothesis $\{H_1^0, H_2^0, \ldots, H_r^0\}$. Thus, if the correct rank is $r^* < r$, any rank r factorization could correspond to some erroneously rejected H_s^0, a.k.a. false discoveries.

Now, given factor matrices $X \in \{0,1\}^{n \times r}$ and $Y \in \{0,1\}^{m \times r}$, we define a random variable Z_s with domain $\{0,1\}$, which takes the value 1 if and only if the outer product $Y_{\cdot s} X_{\cdot s}^\top$ covers only noise. The FDR of a BMF is therefore computed via

$$\mathbb{E}\left(\frac{v}{r}\right) = \frac{1}{r} \sum_{s=1}^{r} \mathbb{P}(Z_s = 1)\ . \tag{2}$$

We propose two upper bounds on $\mathbb{P}(Z_s = 1)$, which in turn implies two variants to control the FDR in BMF. However, first we need to employ an independence assumption on positive noise.

Definition 2 (Bernoulli matrix)**.** *Let B be an $m \times n$ binary matrix. If the entries of B are independent Bernoulli variables, which take the value 1 with probability p and zero otherwise, i.e.,*

$$\mathbb{P}(B_{ji} = 1) = p,\ \mathbb{P}(B_{ji} = 0) = 1 - p\ ,$$

then B is a Bernoulli matrix *with parameter p.*

In what follows, we assume that the positive noise matrix N_+ is a Bernoulli matrix. If a tile $(X_{\cdot s}, Y_{\cdot s})$ approximates noise, the outer product $Y_{\cdot s}X_{\cdot s}^\top$ and the the positive noise matrix have some entries in common. The *overlap* is computed by the sum of common 1 entries, e.g., via

$$|Y_{\cdot s}X_{\cdot s}^\top \circ N_+| = Y_{\cdot s}^\top N_+ X_{\cdot s}\ .$$

Accordingly, we define the density of a tile in a binary matrix.

Definition 3 (δ-dense)**.** *Let M be an $m \times n$ binary matrix and $\delta \in [0, 1]$. We say a tile $(X_{\cdot s}, Y_{\cdot s})$ is* δ-dense in M *if*

$$Y_{\cdot s}^\top M X_{\cdot s} \geq \delta |X_{\cdot s}||Y_{\cdot s}|\ .$$

A Boolean matrix product that approximates the data matrix well, assumably covers a high proportion of ones in D. Therefore, the tiles returned by a BMF are expected to be dense in D ($\delta > 0.5$). We explore by the following theorem the probability with that a δ-dense tile of given minimal size exists in a Bernoulli matrix. This gives us an upper bound on the quantity $\mathbb{P}(Z_s = 1)$ from Eq. (2), which in turn allows us to bound the FDR.

Theorem 1. *Suppose B is an $m \times n$ Bernoulli matrix with parameter p, δ is in $[0, 1]$, $a \leq n$, and $b \leq m$. The probability that a δ-dense tile of size $|x| \geq a$ and $|y| \geq b$ exists is at least*

$$\binom{n}{a}\binom{m}{b} \exp(-2ab(\delta - p)^2)\ . \tag{3}$$

The proof of Theorem 1 indicates that the tightness of Bound (3) might suffer from the extensive use of the union bound. This originates from the numerous possibilities to select a set of columns and rows of given size. If we expect that rows and columns which are selected by a tile have proportionately many ones in common, we bypass the requirement to take all possible column and row selections into account. To this end, given an $(m \times n)$-dimensional matrix B, we assess the value of the function

$$\eta(B) = \max_{1 \leq i \neq k \leq n} \langle B_{\cdot i}, B_{\cdot k} \rangle\ .$$

Theorem 2. *Let B be an $(m \times n)$ Bernoulli matrix with parameter p and let $\mu > p^2$. The function value of η satisfies on the normalized matrix $\eta\left((1/\sqrt{m})B\right) \geq \mu$ with probability at least*

$$\frac{n(n-1)}{2} \exp\left(-\frac{3}{2} m \frac{(\mu - p^2)^2}{2p^2 + \mu}\right). \tag{4}$$

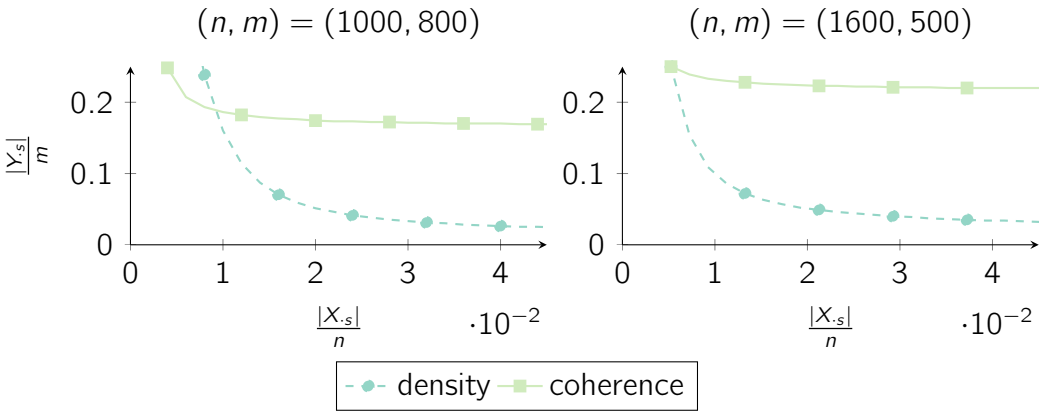


Figure 1: Minimum relative size $|Y_{\cdot s}|/m$, depending on $|X_{\cdot s}|/n$, for which the probability of a false discovery is at most 0.01, bounded by Eq. (3) based on density and Eq. (4) based on coherence.

If the columns of matrix B are normalized, then the function $\eta(B)$ returns the *coherence* of B, which measures how close the column vectors are to an orthogonal system. Thus, we refer to the bound in Eq. (4) as the *coherence bound* and to the bound in Eq. (3) as the *density bound*.

Comparison of the Bounds Theoretical derivations of solutions to the problem

$$\min_{X,Y} L(X,Y) = |D - Y \odot X^\top| \text{ s.t. } X \in \{0,1\}^{n \times r}, Y \in \{0,1\}^{m \times r}$$

enable a comparison of the bounds. Assuming that the positive noise matrix is a Bernoulli matrix with probability $p = 0.1$, we plot the minimum relative size $|Y_{\cdot s}|/m$ against the relative size $|X_{\cdot s}|/n$ such that the probability that the tile $\mathbb{P}(Z_s = 1) \leq 0.01$ in Fig. 1. The comparison indicates that the coherence bound is more loose than the density bound, especially if the dimensions of the data matrix are unbalanced. However, our empirical demonstration shows that both bounds deliver suitable rank selection results in practice.

References

[1] Sibylle Hess, Katharina Morik, and Nico Piatkowski. The PRIMPING routine - tiling through proximal alternating linearized minimization. *Data Min. Knowl. Discov.*, 31(4):1090–1131, 2017.

Functional validation of mutations associated with neuroblastoma progression

Marc Schulte
Molecular Oncology, University Hospital Essen
marc.schulte@uk-essen.de

Neuroblastoma (NB) is one of the most common childhood cancers. In Germany, 150 children are newly diagnosed with NB per year. A special feature of neuroblastoma is that the disease course varies significantly. Some neuroblastomas regress after low-dose chemotherapy or even without any treatment. However, in other cases the disease spreads very aggressively. Even if the tumor shows good response to an initial treatment, difficult to treat metastases and recurrent tumors are often developed. We previously identified several potential key genes involved in NB relapse by analyzing genome sequencing and expression data of paired primary and recurrent NB. In order to validate these data in a biological system, we use CRISPR / Cas9 technology to study the function of those genes in neuroblastoma cell lines. For this purpose, we established CRISPR / Cas9 –mediated overexpression, downregulation or knockout of individual genes. Using these methods we analyzed the role of the potential key genes on proliferation, clonal outgrowth, migratory capacity, invasiveness and drug resistance. Our aim is to gain new insights into mechanisms underlying neuroblastoma recurrence to identify potential targets for a personalized therapy based on individual relapse-driving events.

Neuroblastoma is the most common extra cranial solid tumour in childhood and accounts for 7 -10% of all childhood cancers. As a tumour of the autonomic nervous system, neuroblastoma derives from neural crest tissue and thus usually arises in a paraspinal location in the abdomen or chest [1]. Thanks to improved therapies, neuroblastoma often initially responds very well to the treatment. However, at relapse there is only very little to offer for the patients and hence relapses correlate with poor prognosis and fatal outcome.

The median age at diagnosis of NB is 17 months and the incidence of neuroblastoma is 10.2 cases per million children under 15 years [4,5]. Thus, neuroblastoma develops early in life suggesting that environmental influences are contributing less than the genetic background or mutations in cancer driver genes. Consequently, NB present with significantly lower mutation rate compared to tumours of adulthood including melanoma and lung cancer. Thus, NB is an excellent model for the investigation of the contribution of individual oncogenes to cancer progression.

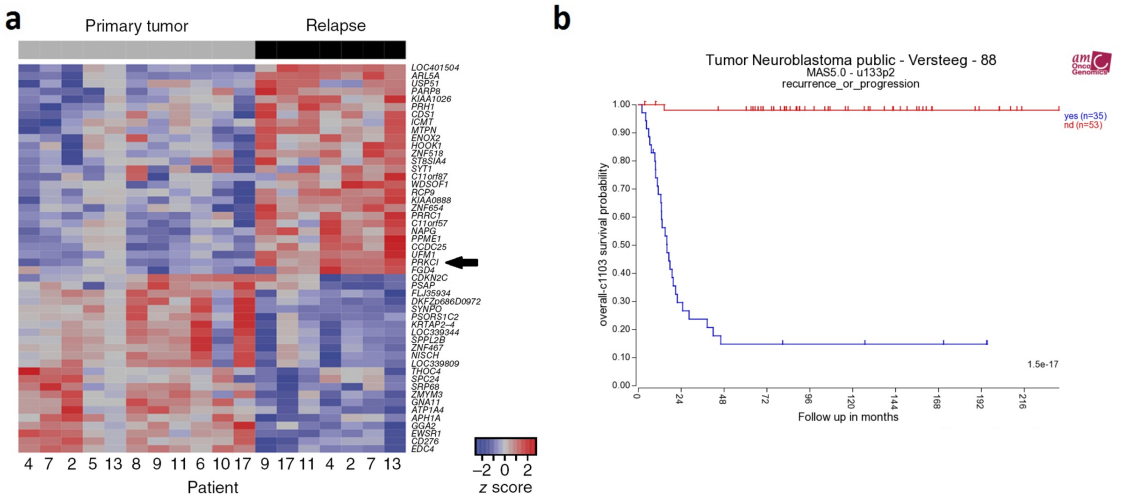


Figure 1: **PRKCI expression is increased in relapsing NBs and high PRKCI levels are correlated with unfavorable clinical outcome (R2: Genomics Analysis and Visualization Platform, AMC Amsterdam).** (a) PRKCI expression level in primary and relapse neuroblastomas (figure from [8]). (b) Kaplan curve of relapse-free survival for patients with PRKCI low and high NBs (data from Valentijn et al., PNAS [9]).

Previously, we used whole-exome sequencing, mRNA expression profiling, array CGH and DNA methylation analyses to characterize 16 paired samples at diagnosis and relapse from individuals with neuroblastoma [6]. We observed that Protein Kinase C Iota (PRKCI) was significantly higher expressed in recurrent neuroblastoma compared to the corresponding primary tumors (Figure1a) [8]. Moreover, an elevated PRKCI expression correlates with NB relapse and poor survival (Figure 1b). As a pro-tumorigenic role for PRKCI has been discussed for other entities, we hypothesized that PRKCI might play a yet unknown role in the development of NB metastases and /or recurrence [2,7,10].

We used different CRISPR / Cas9 based-approaches to knockout (KO) or overexpress PRKCI in NB cells. While the KO was based on a double strand break induced by wild-type Cas9 (Figure 2a)[6], the CRISPR/Cas9 Synergistic Activation Mediator (SAM) system was used to overexpress PRKCI (Figure 2b) [3]. Here, a nuclease inactive Cas9 is fused to a VP64 activation domain and directed to the promoter region of the gene

of interest by the guide RNA. Transcriptional activation is achieved by the integration of two aptamer regions into the guide RNA backbone. Both, knockout and overexpression were confirmed by quantitative RT-PCR (Figure 2c) and Western Blot analysis (Figure 2d).

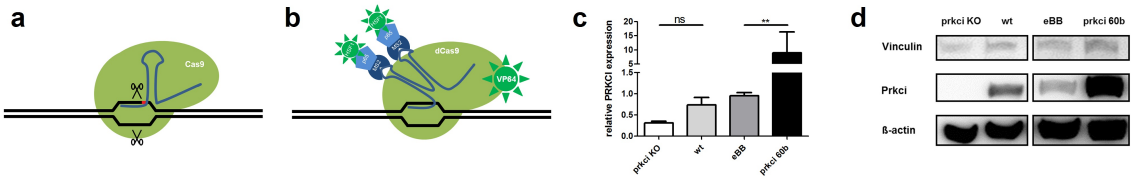


Figure 2: **CRISPR / Cas9 based-approaches to knockout and overexpress PRKCI in NB cells.** (a) Double strand break generation by Cas9 to create a gene knock out (KO). (b) CRISPR SAM [3], which requires incorporation of three activation domains, VP64, P65 and HSF1, into the nuclease inactive Cas9 complex to activate transcription. (c) Confirmation of PRKCI KO and overexpression in SHEP cells (KO: knockout, wt: wild type, eBB: empty backbone control, 60b: overexpression) by real time PCR. (d) Western Blot analysis confirming KO and overexpression, respectively, of PRKCI in SHEP cells.

Colony formation and MTT assays revealed that PRKCI knock out surprisingly increased the rate of clonal outgrowth and proliferation compared to parental cells. Vice versa, overexpression of PRKCI decreased the clonal outgrowth when compared to the vector control cells (Figure 3a-b).Additionally, migratory capacity was reduced for PRKCI KO cells compared to the parental cells (Figure 3c).

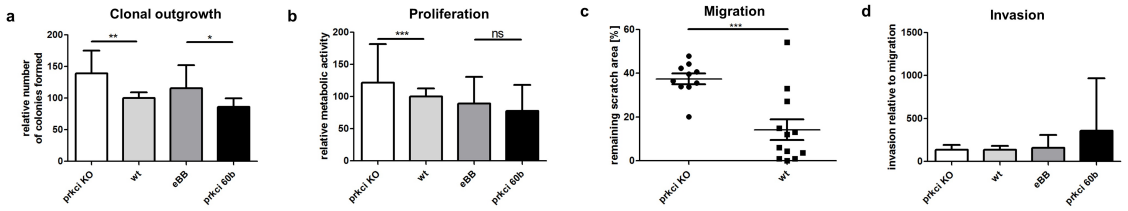


Figure 3: **Clonal outgrowth, proliferation, migratory capacity and invasiveness of SHEP cells with up-or downregulated PRKCI expression in comparison to the parental cells (wt).** (a) Cells were seeded at low density and incubated for 8-10d, colonies (>50 cells) were stained and counted. (b) 10000 cells/well were incubated for 48h in 96-well plates and metabolic activity was determined by addition of MTT dye. (c) Cells at confluency were scratched and closing of the scratch by cell migration was documented at 0h and 24h using Tscratch software (developed by Gebäck and Schulz, ETH Zürich) (d) Serum-deprived cells were seeded in matrigel-coated inserts containing serum-free media, while the lower chamber contained full medium. Cells moving through the inserts were stained and counted.

Based on this observation, we further investigated migratory and invasive capacity of the cells by Boyden chamber assays. As shown in Figure 3d, overexpression of PRKCI results in increased invasiveness of SHEP cells, however, this did not reach statistical significance.

Although PRKCI is found to be elevated in aggressive and recurrent NB tumors, both CRISPR-mediated PRKCI knock-down or overexpression resulted in formation of viable subclones in SHEP cells. While PRKCI knock-down reduced migratory capacity, it surprisingly increased clonal survival and proliferation. In vivo experiments will be necessary to clarify the role, if any, for PRKCI in modulating neuroblastoma aggressiveness.

References

[1] Brodeur, G. (2003). Neuroblastoma: biological insights into a clinical enigma. Nature Reviews Cancer, 3(3), pp.203-216.

[2] Kikuchi, K et al. (2013) Protein kinase C iota as a therapeutic target in alveolar rhabdomyosarcoma. Oncogene, 32(3):286-95.

[3] Konermann, S et al. (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature, 517(7536):583-8.

[4] London, W (2005). Evidence for an Age Cutoff Greater Than 365 Days for Neuroblastoma Risk Group Stratification in the Children's Oncology Group. Journal of Clinical Oncology, 23(27), pp.6459-6465.

[5] Maris, J (2010). Recent Advances in Neuroblastoma. New England Journal of Medicine, 362(23), pp.2202-2211.

[6] Ran, F et al. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc, 8(11), pp.2281-2308.

[7] Regala, RP et al. (2009) Atypical protein kinase C iota is required for bronchioalveolar stem cell expansion and lung tumorigenesis. Cancer Res, 69(19):7603-11.

[8] Schramm, A et al. (2015). Mutational dynamics between primary and relapse neuroblastomas. Nature Genetics, 47(8), pp.872-877.

[9] Valentijn, LJ et al. (2012) Functional MYCN signature predicts outcome of neuroblastoma irrespective of MYCN amplification. Proc Natl Acad Sci USA, 109(47):19190-5.

[10] Wang, Y et al. (2017) PKC iota regulates nuclear YAP1 localization and ovarian cancer tumorigenesis. Oncogene, 36(4):534-545.

Computing Protein Similarity with Recursive Hash Tables

Henning Timm
Genome Informatics, Institute of Human Genetics
University Hospital Essen, University of Duisburg-Essen
henning.timm@tu-dortmund.de

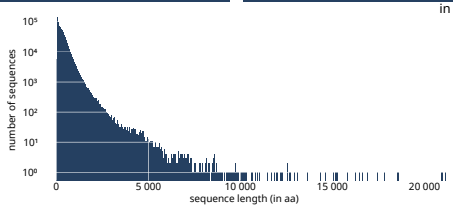
30.11.2017

This report documents our development of an adaptable hashing data structure and its application to solve the protein similarity problem. Finding the most similar proteins to DNA fragments from a metatranscriptomic sample can provide insight into the species present in the sample and their abilities, even if no reference genome for the specific species is available.

Introduction and Problem Definition

Second generation sequencing (SGS) allows collecting very large and heterogeneous sets of DNA data. In metatranscriptomic sequencing, a sample containing a combination of organisms is taken. As opposed to other sequencing approaches, no separation or selection of species is performed so that the sequenced data can originate from an unknown amount of, potentially unknown, species. Since for most species no reference genome is available, annotated protein databases are used to classify the contained species. These databases contain protein sequences that are associated either with a certain function (e.g. photosythesis) or a specific group of organisms (e.g. phototrophic algae). By comparing short DNA fragments (reads) sequenced from the sample using SGS with proteins in such a database, the kinds and abilities of organisms in the sample can be judged. For this task, software for the analysis of metatrascriptomic data, like TaxMapper [2] is required to quickly find query reads in a protein database. As illustrated in Figure 1a, the size of proteins ranges from less than thirty amino acids to several ten thousands. DNA reads on the other hand have a fixed length of about 100 - 250 base pairs (bp),

# sequences	3 143 601	maximal length	21 165
length < 30	371	mean length	351
length ≥ 10^4	52	median length	260



(a) Structure of the reference protein database used by TaxMapper.

(b) Illustration of the protein similarity problem. For each read, the most similar proteins in the database has to be found.

depending on the sequencing technology. As a first step, DNA reads are translated to amino acid sequences (proteins). For this, triplets of DNA bases are translated to one amino acid so that, assuming that the whole read is translated, the protein length of a read is $\ell_{\text{AA}}(r) = \lfloor \frac{\ell_{\text{BP}}(r)}{3} \rfloor$.

The protein similarity problem, as illustrated in Figure 1b, is defined as follows: Given a query set $Q = \{ r_i \mid i = 0, .., n \}$ of reads with a fixed size of $\ell_{\text{AA}}(r_i) = \ell_{\text{AA}}(r) = \lfloor \frac{\ell_{\text{BP}}(r)}{3} \rfloor$ and a reference database of protein sequences $R = \{ p_j \mid j = 0, .., m \}$ with variable lengths $\ell_{\text{AA}}(p_i)$, find the most similar protein(s) in the database for each read in the query set. However, at this point, the question of the similarity measure used is still open.

Other software that solves this problem uses E-Values, i.e. alignment scores computed by BLASTX [1], as similarity measure. This includes RAPSearch [8] which is used by TaxMapper. While reasonably accurate, RAPSearch is comparably slow and is the main runtime sink of TaxMapper. DIAMOND [4] (used by the MEGAN software [5]) uses a dual indexing approach and is faster than RAPSearch, however, its accuracy is lower. DIAMOND, too, computes alignments to report E-Values.

Using our MinHashing approach described in the following section, computing alignment scores might not be necessary. Since MinHashing allows the quick estimation of the similarity of two sequences, this estimated similarity value can potentially be used to avoid the costly computation of alignment scores. We are currently evaluating if the MinHash scores generated by our approach can replace E-Values to judge the similarity of reads and proteins.

Building a MinHash Index

MinHashing is a technique used to approximate the Jaccard similarity of two sequences by using the smallest hash values of each sequence as an identifying feature. The standard approach as described by A. Broder [3], is to compute all k-mers (substrings of length

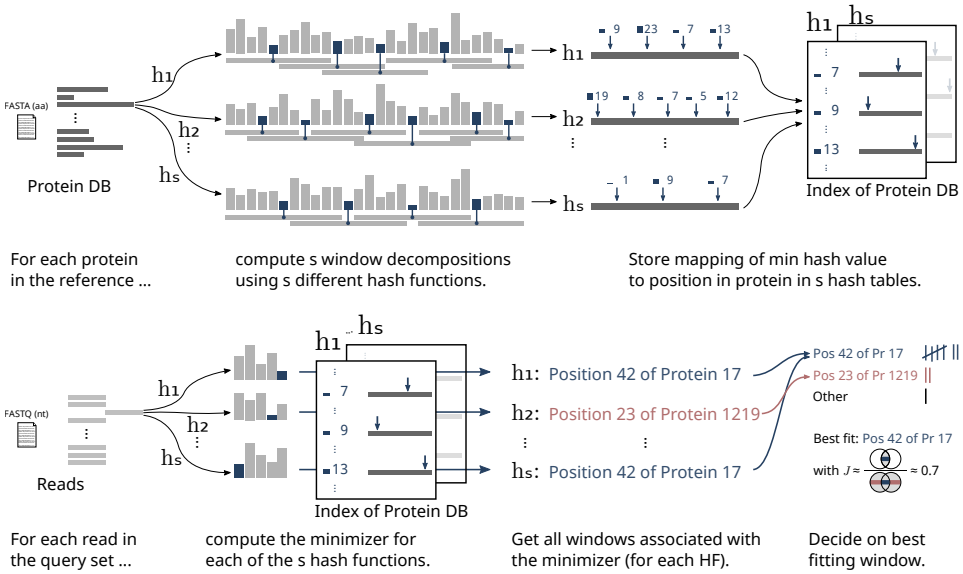


Figure 2: Creation of a MinHash Index(MHI) (a) and application of a MHI to find protein sequences similar to the query read (b). MinHash values for each interval are marked in blue.

k) of the sequences, hash them using s different hash functions, and find the smallest hash value for each hash function. These s different minimal hash values comprise the sketch of the sequence. The hamming distance of two sketches is a robust estimator for the Jaccard similarity of two sequences. However, this approach can only be used if the compared sequences are of comparable sizes.

To make MinHashing usable for our approach we use winnowing [7] (which is almost equivalent to the similar minimizer [6] approach). A sliding window of size w is moved through the hash values of the k-mers of the reference proteins. Each time the smallest hash value in the window is changed, either by the old minimum being push out or by a smaller minimum entering the window, a new reference interval is created. We use winnowing with s different hash functions and a windows size of $w \approx \ell_{\mathrm{AA}}(r)$ to create a mapping of MinHash values to reference protein intervals. This MinHash Index (MHI) is illustrated in Figure 2.

To query the MHI with a read, the minimizer of the read under each of the s hash functions is computed. Then each of the s hash tables T_k for $k = 1, .., s$ is queried with the respective hash value $T_k[h_k]$ to get a set of associated reference intervals. Using these interval sets, the similarity between the read and the proteins is estimated.

However, the performance of the MHI hinges on the performance of the underlying hash data structure. We are currently evaluating several approaches for an efficient implementation, including our own specialized hash table.

Our implementation consists of recursive hash tables using pages with a fixed number of slots for each hash value. Overflowing values from the top level hash table are inserted into the lower level hash table. Several layers of hash table are used, based on the structure of the data. If an entry does not fit into any hash table due to high table load it is put into a stash, which can be implemented as a sorted list or a general purpose hash table. Since colliding entries are saved in the same page, and hence lie adjacent in memory, this approach reduces cache misses. The specific architecture of the hash table, i.e. the number of hash table layers, can be specified by the user. Finding an optimal set of parameters for a specific database is still a topic of research.

References

[1] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[2] Daniela Beisser, Nadine Graupner, Lars Grossmann, Henning Timm, Jens Boenigk, and Sven Rahmann. Taxmapper: an analysis tool, reference database and workflow for metatranscriptome analysis of eukaryotic microorganisms. *BMC genomics*, 18(1):787, 2017.

[3] Andrei Z Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings*, pages 21–29. IEEE, 1997.

[4] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.

[5] Daniel H Huson, Sina Beier, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans-Joachim Ruscheweyh, and Rewati Tappu. Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology*, 12(6):e1004957, 2016.

[6] Michael Roberts, Wayne Hayes, Brian R Hunt, Stephen M Mount, and James A Yorke. Reducing storage requirements for biological sequence comparison. *Bioinformatics*, 20(18):3363–3369, 2004.

[7] Saul Schleimer, Daniel S Wilkerson, and Alex Aiken. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 76–85. ACM, 2003.

[8] Yongan Zhao, Haixu Tang, and Yuzhen Ye. Rapsearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–126, 2011.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C3

Multi-level statistical analysis of high-frequency spatio-temporal process data

Katharina Morik Wolfgang Rhode
Tim Ruhe

Astroteilchenphysik, C3 Project, SFB 876, TU Dortmund

Real-Time Stream Analysis for CTA

Kai Brügge

15.12.2017

The Cherenkov Telescope Array (CTA) is the largest ground based gamma-ray telescope to date. Once completed it will be able to map the gamma-ray sky in a wide energy range from several tens of GeV to some hundreds of TeV and will be more sensitive than previous experiments by an order of magnitude. CTA will try to observe transient phenomena like gamma-ray bursts (GRBs) and flaring active galactic nuclei (AGN). Multi-wavelength observations of these phenomena can only be succesful if other observatories can be alerted as quickly as possible.

The entire CTA array in the southern hemisphere will contain between 80 and 100 telescopes. Up to $20\,000$ images per second will be produced during observation. Noise suppression and feature extraction algorithms are applied to each image in the stream. Previously trained machine learning models are applied to the stream in an online manner. We use the Apache Flink distributed streaming engine to handle the large amount of data coming from the telescopes. Here we present results of our investigation and show a first prototype capable of analyzing CTA data in real-time.

1 The Cherenkov Telecope Array

Once completed, the Cherenkov Telescope Array (CTA) will be able to map the gamma-ray sky in a wide energy range from several tens of GeV to some hundreds of TeV and will be more sensitive than previous experiments by an order of magnitude. Data from the telescopes is sent via ethernet to a central computing facility. Each telescope records an image of the cherenkov light flash in the atmosphere. Different camera types and telescope sizes are part of CTA's array. The cameras record images with a total of ~ 1800 to ~ 11000 pixels depending on camera type.

Cherenkov light produced by cosmic rays is responsible for most of the background in the data recorded by an imaging atmospheric cherenkov telescope (IACT). Filtering

air showers produced by cosmic rays while keeping those produced by gamma rays is the big challenge in IACT data analysis. The real time analysis (RTA) has to apply these filters to supress the cosmic ray background. We use a pre-trained Random-Forest to select gamma-ray events.

2 Real Time Analysis

One of CTA's main goals is monitoring the sky for transient events. These include Gamma-Ray Burst (GRBs) events and variable Active Galactic Nuclei (AGNs). CTA can alert other experiments to trigger observations in other wavelength bands. In the same manner, other observatories can trigger the CTA array to record data from any given sky position in case events are detected by other observatories. The RTA will be supplied with calibrated images of each triggered telescope. While the technical details of the data aquisition system are not known yet, we presume that single telescope images are already merged into one unit in the data stream [3]. To perform supervised machine learning, the images recorded by the telscopes have to be converted to a feature representation. This reduces the size of the events greatly since data from each image gets reduced to ~ 15 numbers per telescope and event. Simulated data is used to train the models using the Python machine learning library `scikit-learn` [5]. These models are then used for filtering of comsic ray showers and estimation of primary particle energy. The trained `scikit-learn` models are converted into the PMML [4] format using the `sklearn2pmml` [6] library. This way the stored model can be shared between programming languages and applied to the data stream from the telescopes. The expected event rate of CTA will be between 10000 and 20000 events per second. CTA's real time processing system will have to be able to handle that data rate with a maximum delay of $30\,\mathrm{s}$

3 Machine Learning Performance

We trained a Random-Forest [2] classifier with 200 trees on simulated data to separate signal events from background events [3]. The trained model is applied to each image from the telescopes in the data stream. Figure 1 presents the results of the model application. The plot on the left shows the performance of background suppression for the three different telescope types. The predictions for each telescope are then averaged to get a combined prediction for the entire event. This improves background rejection significantly as shown in the right image [3].

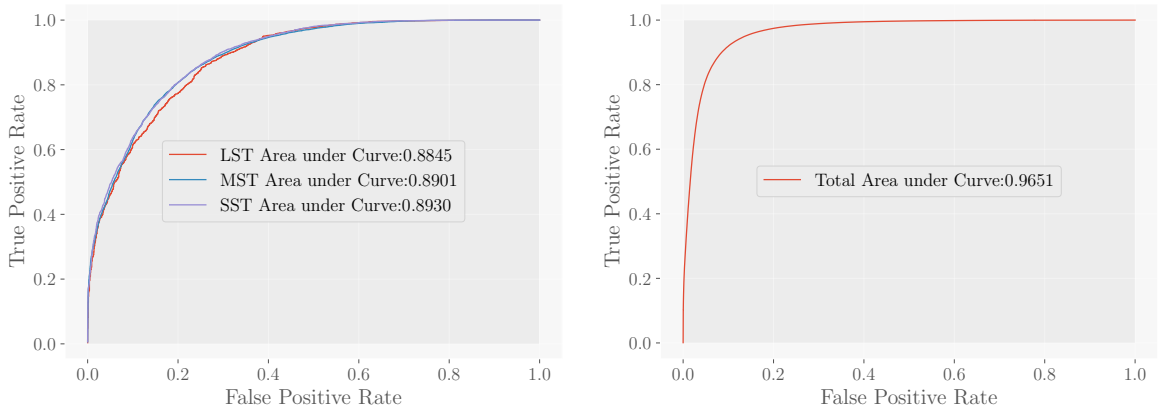


Figure 1: Performance of background suppression for the different telescope types. The small size telescope (SST), the mid size telescope (MST) and the large size telescope (LST). On the right hand site the combined prediction is shown.

4 Runtime Performance

Frameworks for distributed stream processing such as Apache Storm [7] or Apache Flink [1] perform workload distribution in an automated fashion. Features like fault tolerance and high availability mechanisms are built in. We perform our experiments on Apache Flink due to faster processing, easier setup and a more comfortable high-level API compared to Storm [3]. The Figure 2 presents the evaluation of a full CTA pipeline executed on top of Flink. For this test a machine with 24 physical CPU cores was used. The datarate goes up to approximately 14000 events per second on this single machine. To reach higher data rates we simply add another machine with 24 cores. We use a simple self hosted Flink cluster on these two machines. This way event rates of more than 20000 events per second are achieved. The right hand side of figure 2 shows a screenshot of Flink's web interface.

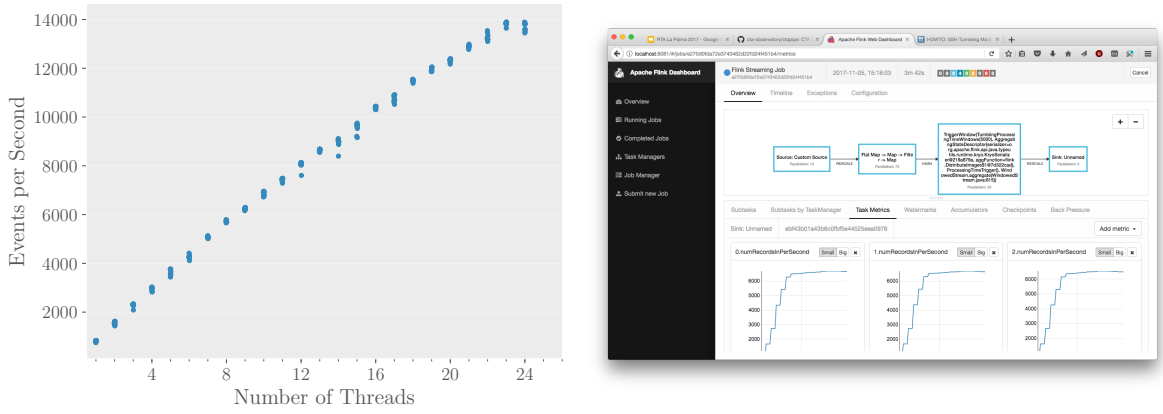


Figure 2: On the the throughput of CTA analysis pipeline executed on multiple cores is plotted. On the right hand side three parallel processes on two machines write results to a CSV file. As seen in the display, data rates of 20000 events per second can be achieved with only two computers

References

[1] A. Alexandrov et al. “The Stratosphere platform for big data analytics”. In: *The VLDB Journal* 23.6 (2014-12), pp. 939–964.

[2] L. Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[3] K. Brügge et al. “Distributed Real-Time Data Stream Analysis for CTA”. In: *Proceedings of the 2017 ADASS Conference*. ADASS 2017. Santiago, Chile, 2017.

[4] A. Guazzelli et al. “PMML: An open standard for sharing models”. In: *The R Journal* (2009).

[5] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[6] V. Ruusmann. *Python library for converting Scikit-Learn models to PMML*. 2015. URL: `https://github.com/jpmml/sklearn2pmml` (visited on 10/10/2017).

[7] A. Toshniwal et al. “Storm@Twitter”. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’14. Snowbird, Utah, USA: ACM, 2014, pp. 147–156. ISBN: 978-1-4503-2376-5. DOI: 10.1145/2588555.2595641. URL: `http://doi.acm.org/10.1145/2588555.2595641`.

Long-term performance of FACT's SiPMs

Evaluation of aging effects due to intensive light yield in an IACT with SiPMs

Jens Björn Buß
Experimentelle Physik 5
Technische Universität Dortmund
jens.buss@tu-dortmund.de

The First G-APD Cherenkov Telescope (FACT) is the first application of silicon photmultipliers (SiPMs) in an imaging atmospheric Cherenkov telescope (IACT). A major advantage of these kind of sensors is, that it is possible to operate them under more severe light conditions than commonly used photomultiplier tubes without damaging them. In addition, SiPMs promise to provide a larger duty cycle due to a lack of light yield related aging. During the past 6 years FACT has been operated under a variety of light conditions from dark night to nights with direct illumination of the cameras by the full moon. So far neither a sign of aging nor a light yield related damage to the SiPMs has been detected. This work's approach is to investigate signs of aging or damages of FACT's SiPMs more deeply by use of routinely measured dark count spectra aka. single p.e. spectrum of each pixel.

Introduction FACT's camera holds 1440 silicon photo multipliers (SiPM) as pixels. By use of these devices, the FACT collaboration proved the application of semi-conductor based photon detectors for the imaging atmospheric Cherenkov technique [1].

Since FACT is able to operate during bright moon light without the use of UV filters or a reduced gain, the detectors are exposed to a larger amount of light than those of other IACTs. Over the past six years, FACT's SiPMs were operated almost daily with a variety of light conditions. Figure 1 shows the distribution of the mean camera current per pixel and measurement interval. This current is proportional to the mean light yield

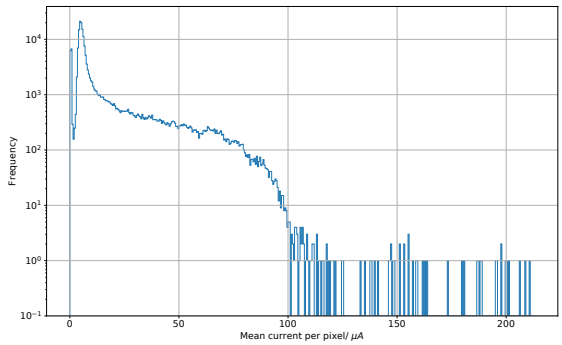


Figure 1: Distribution of the mean current over all pixels per measurement interval during observations.

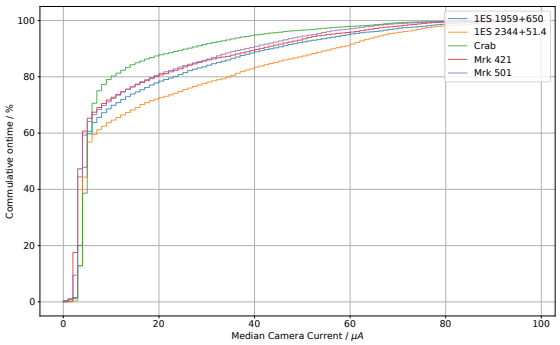


Figure 2: Relative gain of observation time depending on light conditions.

of the pixels. It is visible that most of the observations are done with low and moderate light conditions at about 5 µA. A smaller fraction of observations were undertaken at severe light conditions corresponding to currents above 50 µA and direct moon light with currents above 200 µA. With this, FACT is able to increase its duty cycle and fulfill its goal of unbiased monitoring of bright Blazars in the northern hemisphere.

The relative gain of observation time with respect to the underlying light conditions is presented in figure 2. Going above 50 µA generates a gain of about 10 % observation time, which is not only a general increase, but also supports gap-less monitoring without the need to interrupt for severe light conditions.

Even though, FACT's SiPMs promise to be very robust for such conditions and did not show indications for aging, so far, their quality and performance after six years of operation need to be investigated more deeply. Thus, the long-term behavior of each SiPM's dark count spectrum [2] is investigated in order to monitor the SiPM's core properties, i.e. gain and crosstalk, over time. Details to the utilization of single p.e. spectra for determining SiPM properties are described in [2]. A nightly performed measurement of such spectra allows to investigate those properties behavior over time. Unexpected changes are considered to indicate aging and malfunctions.

Behavior of the SiPM' properties over time Figure 3 and 4 present the course of the gain and the crosstalk probability per night averaged (blue dots) over all pixels between 2013 and 2017. Furthermore, the distance between the 10 and 90 percent quartile (orange), as well as the distance between the according values of the pixels with the larges and the smallest value (blue) are depicted.

The mean gain per night (fig 3) is stable for large periods of time and shows only small seasonal variation. The variation of individual pixel gains, indicated by the inter-percentile distance, stays stable for most of the time. However, several periods with emerging

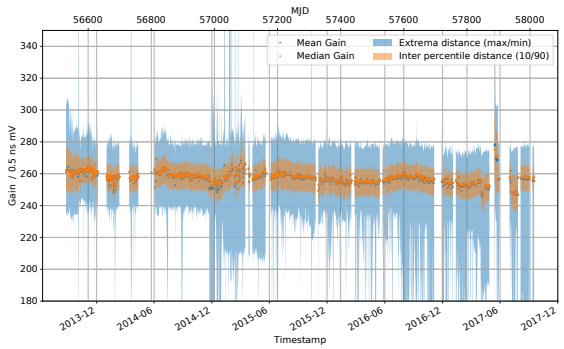


Figure 3: Distribution of the SiPMs' gain over time.

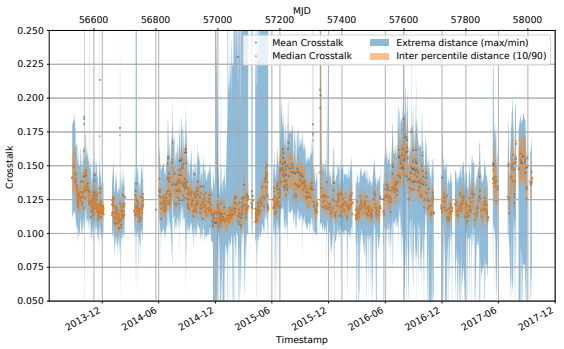


Figure 4: Distribution of the SiPMs' crosstalk probability over time.

quantities of outliers are visible, e.g. in the first half of 2015 and the summer of 2017. While the maximum gain seems to stay stable over time, the minimum gain drifts to smaller values. Furthermore, a gain drift of the whole camera is visible in January 2015 and June 2017.

Also the course of the crosstalk (fig 4) indicates a changed behavior of some pixels in the first half of 2015, as well as an increase of lower value outliers beginning with 2016. In addition, a clear seasonal variation of the crosstalk is visible. Simulation studies have shown that the method, which was chosen to measure the crosstalk, is prone to variations of the dark count rate (DCR). In fact, the DCR and the measured crosstalk probability show a proportionality. Correspondingly, these variations appear related to the seasonal variations of the DCR, which has a direct temperature dependency, according to [3]

The crosstalk probability itself, alike also the gain, is only temperature dependent via the breakdown voltage of the SiPM. However, this dependency is corrected by adjusting the voltage applied to the SiPMs, to keep a constant over-voltage [1]. As visible in figure 3, this is generally the case. Nevertheless, the influence of the DCR to the crosstalk measurement has not been corrected yet but will be implemented in a continuation of this study in order to determine a possible drift of the crosstalk and evaluate possible aging effects.

Course of faulty pixels The course of the gain, allows to determine the trend of the general performance of the SiPMs. A mostly homogeneous response of all pixels is required at all times. The gain is a measure for the pixels response. Thus, the number of pixels diverging from the median of all camera pixels is an indication for malfunctioning pixels. In figure 5, the number of pixels without a valid signal (green) and the number of pixels diverging from the median gain of the camera (orange and blue) is depicted.

At beginning of FACT's operation, twelve pixels were broken already, possibly due to flaws during the camera's assembly. Two additional groups of four pixels failed each at the

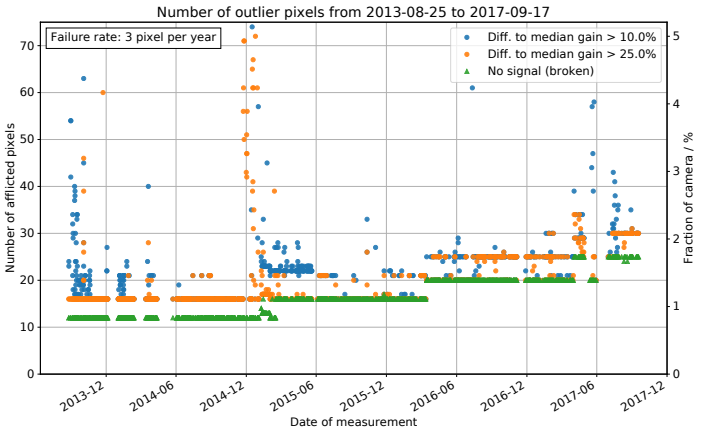


Figure 5: Course of the quantity of pixels without a valid signal (green) and pixels diverging from the median camera gain (orange and blue).

beginning of the years 2015 and 2016. In April 2017 a third group of five pixels stopped working. Investigations of the hardware showed that at that time the according data acquisition boards got broken. So far there are no indications that the malfunctioning of these pixels is related to the SiPM itself, but appears to be more likely due to the broken data acquisition hardware. In total, this indicates an estimated failure rate of FACTs pixels of about three pixels per year.

Additionally, it is visible that also the number of outliers in the gain is increasing at the same dates. This may be caused by correlations of the individual pixels due to shared voltage supply channels. However, the actual cause of the increase of gain outliers is not yet determined and requires further investigations.

Nevertheless, it is also visible that, after six years of operation, the number of pixels with a difference of more than 10 % is still below 5 % for the whole camera. This indicates that FACT still has a homogeneous flat-fielding and provides a good performance.

References

[1] H. Anderhub et al. Design and operation of FACT – the first g-apd cherenkov telescope. *JINST*, 8(06):P06008, 2013.

[2] Jens Björn Buß. *FACT - Signal Calibration: Gain Calibration and Development of a Single Photon Pulse Template for the FACT Camera*. Diploma thesis, TU Dortmund, 2013.

[3] T Krähenbuehl. *G-APD arrays and their use in axial PET modules*. Diploma thesis, ETH Zürich, 2008.

Unfolding of the Muon Neutrino Energy Spectrum with Multiple Years of IceCube Data

Mathis Börner
Experimentelle Physik 5b
Technische Universität Dortmund
mathis.boerner@tu-dortmund.de

The goal of my work is to measure the atmospheric muon neutrino energy spectrum with IceCube data. The covered energy ranges from 125 GeV to multiple 2.5 PeV. The analysis can be split into two major steps. In the first step a sample of muon neutrino event candidate is created. In the second step a model independent unfolding is conducted. This report presents the results achieved in the year 2017. It covers the final muon neutrino sample and gives an overview of the used unfolding approach.

1 Introduction

The analysis I am conducting was successfully applied to previous years of IceCube data [1, 3, 2]. My analysis is the first analysis using multiple years of data. A detailed physical motivation and a broader overview of the analysis can be found in [6].

To obtain the energy spectrum for muon neutrinos first a sample of muon neutrinos has to be generated. The result of the separation is presented in section 2. This sample then can be unfolded. The approach for the unfolding is introduced in section 3.

2 Final Muon Neutrino Sample

The analysis starts with a signal (upgoing muon neutrinos) to background (mostly atmospheric muons) ratio of 1:1 000 000. After the separation the purity of the sample is approximately 99.7 % (sig/bkg-ratio: ∼1003:1). This is achieved with a Random Forest [8] classification followed by an energy dependent classification score cut. To utilize the fact that the background contamination is not only energy dependent but also depends on the zenith, two cuts for two different zenith regions are used. A more detailed description of the separation can be found in [7].

The event rate of the final sample is 3.22 mHz. In comparison to the previous best muon neutrino sample in IceCube [4] the rate is increased by 32.5 % with both sample having the same purity of 99.7 %. The classification score distribution for data and the different components from simulations is shown in Figure 1.

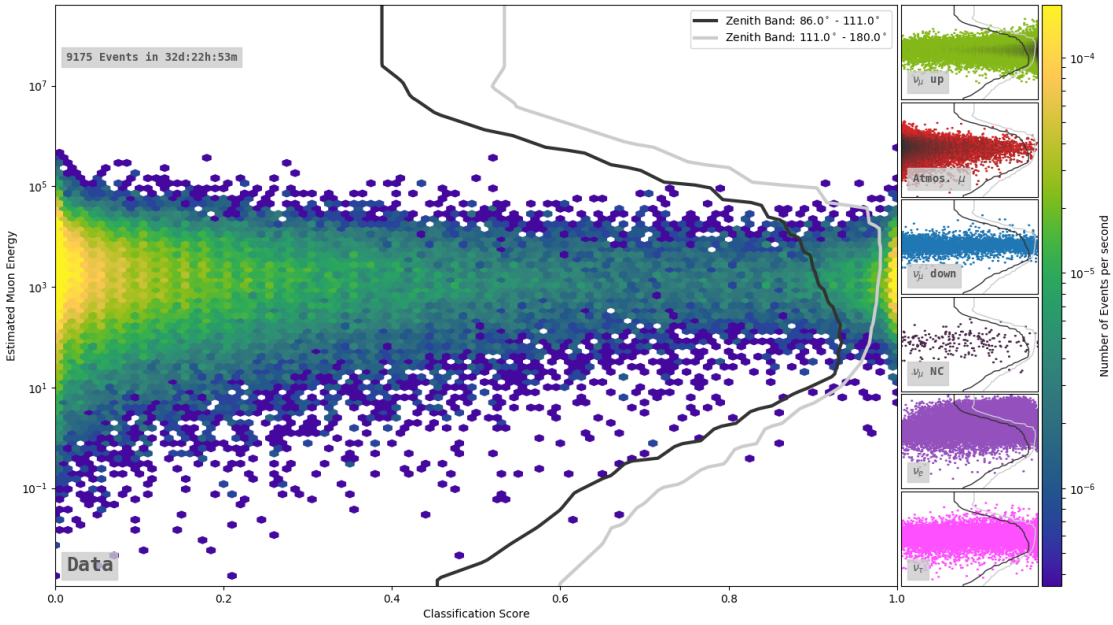


Figure 1: The histogram on the left shows the energy dependent distribution of the classification score for the measured events. On the left side the contribution of the signal component ("ν_μ up") and all the different background components are shown. The two curves illustrates the used cuts.

3 Unfolding

The used unfolding is based on the approach described in [5] and e.g. implemented in TRUEE [10]. In the approach the likelihood

$$\mathcal{L}\left(\vec{g}\middle|\vec{f}\right) = \prod_{i=1}^{m}\left(\frac{\hat{g}_i^{g_i}}{g_i!}\exp\left(-\hat{g}_i\right)\right)\cdot \underbrace{\underbrace{\frac{1}{\sqrt{(2\pi)^n\det(\tau\mathbb{1})}}}_{k}\exp\left(-\frac{1}{2}\tau\vec{f}^{\intercal}\mathbf{C}^{\intercal}\mathbb{1}\mathbf{C}\vec{f}\right)}_{\text{Tikhonov Regularization: } R} \tag{1}$$

is maximized. In (1) $\vec{f}$ is the sought-after neutrino energy spectrum. The entries of $\vec{f}$ are the number of events in the different energy bins. In my analysis I use 14 equidistant bins in between $\log_{10}(125\,\text{GeV})$ and $\log_{10}(2.5\,\text{PeV})$. The idea of (1) is to use the linear model $\mathbf{A}\vec{f}$ to obtain an expected number of events $\hat{\vec{g}}$ in the binned observable space. The matrix $\mathbf{A}$ (response matrix) consists of the conditional probabilities to measure an event in observable bin i given it is in true energy bin j.

The number of expected events ist compared with the actual measured number of events $\vec{g}$ assuming a poissonian distribution for every bin. As shown in [5] trying to unfold without using a regularization in most cases leads to unphysically fluctuating solutions. To suppress those solution a gaussian prior on the second derivative (Tikhonov Regularization) of the solution vector is used. The second derivative is the product $\mathbf{C}\vec{f}$. $\mathbf{C}$ is the so called regularization matrix and consists of factors to calculate the derivative via the finite difference method. The strength of the regularization is set by the factor τ (regularization strength).

The regularization is motivated by the assumption that the correct solution is relatively smooth i.e. has a small second derivative. The muon neutrino energy spectrum is expected to be close to a power law E^{γ}. This spectrum has a no/small second derivative in logarithmic energy and logarithmic flux. To justify the assumption of a smooth solution the logarithm of the solution has to be used in the regularization. Furthermore, the result of the unfolding provides the event spectrum at detector. The measured spectrum at the detector is the original spectrum folded with the detector acceptance. The detector acceptance has not to be smooth. This leads to a modified regularization term:

$$R_{\text{modified}} = k\cdot\exp\left(-\frac{1}{2}\tau\log_{10}\left(\left(\vec{f}+1\right)\mathbf{a}\right)^{\intercal}\mathbf{C}^{\intercal}\mathbb{1}\mathbf{C}\log_{10}\left(\left(\vec{f}+1\right)\mathbf{a}\right)\right). \tag{2}$$

Matrix $\mathbf{a}$ is a diagonal matrix with the inverse effective area for each energy bin on the diagonal. The product $\mathbf{a}\vec{f}$ is the energy spectrum on production level. The regularization term in (2) uses the logarithmic, acceptance corrected spectrum. For this spectrum the it is justified to assumed a smooth solution.

The approach of the former analyses used a parabolic approximation for the likelihood to calculate the uncertainties. The highest energy bins in my analysis can be expected

to be close to zero. For bins close to zero a parabolic approximation can be problematic because solutions below zero are forbidden and therefore the likelihood space become asymmetric.

The solution is obtained via a Markov Chain Monte Carlo (MCMC) [9]. The MCMC provides a sample of solution vectors $\vec{f}$ with the corresponding likelihood values. This sample can be used to get the uncertainties of the solution and to calculate the agreement between the unfolding and a given event spectrum that for example comes from model calculations. Deriving the uncertainties from the sampled posterior pdf is a significant improvement to the former used approach.

References

[1] M. G. Aartsen et al. "Development of a general analysis and unfolding scheme and its application to measure the energy spectrum of atmospheric neutrinos with IceCube". In: *The European Physical Journal C* 75.3 (2015).

[2] M. G. Aartsen et al. "Measurement of the $\nu_\mu$energy spectrum with IceCube-79". In: *The European Physical Journal C* 77.10 (Oct. 2017).

[3] M. G. Aartsen et al. "Unfolding measurement of the Atmospheric Muon Neutrino Spectrum using IceCube-79/86". In: *Proceedings, 34th International Cosmic Ray Conference (ICRC 2015)*. 2015.

[4] M.G. Aartsen et al. "Observation and Characterization of a Cosmic Muon Neutrino Flux from the Northern Hemisphere Using Six Years of IceCube Data". In: *The Astrophysical Journal* 833.1 (2016).

[5] Volker Blobel. "An unfolding method for high energy physics experiments". In: *arXiv preprint hep-ex/0208022* (2002).

[6] M. Börner. *Measurement of the Muon Neutrino Energy Spectrum with Multiple Years of IceCube Data*. Tech. rep. Technischer Report für das SFB 876 - Graduiertenkolleg. TU Dortmund University, Dec. 2015.

[7] M. Börner. *Sample Generation for the Measurement of the Muon Neutrino Energy Spectrum with Multiple Years of IceCube Data*. Tech. rep. Technischer Report für das SFB 876 - Graduiertenkolleg. TU Dortmund University, Dec. 2016.

[8] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001).

[9] Daniel Foreman-Mackey et al. "emcee: the MCMC hammer". In: *Publications of the Astronomical Society of the Pacific* 125.925 (2013).

[10] N. Milke et al. "Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 697 (2013).

Cascade Selection for a Search for High-Energy Astrophysical Tau-Neutrinos in IceCube Data

Maximilian Meier
Lehrstuhl für Experimentelle Physik V
Technische Universität Dortmund
maximilian.meier@tu-dortmund.de

In this work a potential astrophysical flux of tau neutrinos is investigated. The measurement of tau neutrinos would be a clear astrophysical signal since no tau neutrino flux from the atmosphere is expected. Tau neutrinos can be detected with the IceCube detector by their unique double-cascade signature at high-energies. But the signal is buried in a large amount of background events with a signal-to-background ratio of about $\mathcal{O}(1 : 10^{10})$ at trigger level. To remove these background events methods from the field of machine learning are applied. A detection of tau neutrinos would require a very pure sample to achieve a high significance and a very efficient analysis to obtain sufficient statistics in limited time of measurement. This report presents the results achieved in the year 2017.

1 Introduction

A physical motivation, a detailed introduction to the IceCube detector and a description of all event signatures relevant for this analysis can be found in [5].

To analyse the astrophysical flux of tau neutrinos a sample with a high expected fraction of tau neutrinos has to be obtained at first. Such a sample can be created by selecting events with a double pulse signature in at least one of IceCubes Digital Optical Modules (DOMs) and by removing track-like events at the same time.

2 Double Pulse Selection

A Random Forest classifier [4] is used to discriminate double pulse waveforms created by a charged current tau neutrino interaction and a subsequent hadronic tau decay in the detector from waveforms created by single cascade events such as charged current electron neutrino interaction and all flavour neutral current neutrino interactions. For the identification of double pulse waveforms each waveform gets characterised by features describing amplitudes of rising and falling edges detected on its derivative and its statistical properties. All events containing at least one waveform with a classification score of 0.2 or higher are kept for the further steps of the analysis.

This results in an expected event rate of 0.486 per year with a purity of 97 % with respect to single cascade events assuming an unbroken power law astrophysical neutrino flux with a spectral index of -2.13 [1]. The remaining expected event rates for signal and background as well as the purity as a function of the classification score cut are shown in Figure 1.

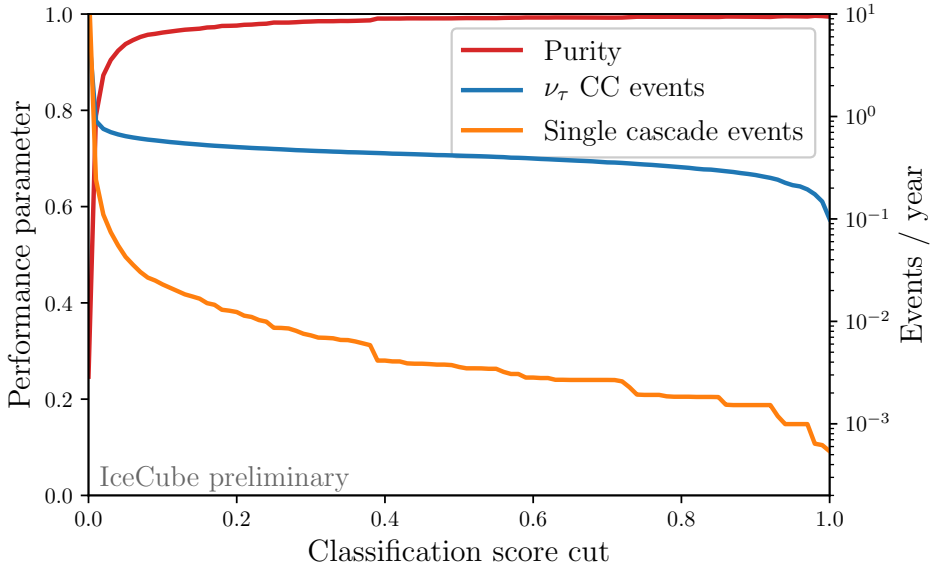


Figure 1: Result of the double pulse selection. Purity and remaining expected event rates as a function of the classification score cut. [2]

3 Cascade Selection

However the signal is still buried by a background of high energy atmospheric muons, muon bundles and muon neutrino events. These types of track-like events can also produce significant double pulses in the IceCube waveforms, usually due to subsequent high energy stochastic losses near a DOM. These are removed in a second classification stage.

The double pulse selection is run on low level data, so before performing any further classification more detailed event reconstructions have to be done. The reconstructions are e.g. maximum likelihood reconstructions based on the hypothesis of an infinite track or a point like cascade. After all reconstructions are performed both simulation and data contain a set of 714 features. One of the important challenges for multivariate analyses where Machine Learning models are trained or multidimensional cuts are developed is to also check the agreement between the simulation and actual measured data. Many previous analyses were comparing one dimensional distributions for all features with the intention to detect problematic mismatches. This analysis uses a new approach which is explained in a bit more detail in [3].

The first step is to remove redundant features, which includes completely constant features as well as features with a Pearson correlation $\rho \geq 0.98$ to any other feature. This preselection reduces the number of features to 281. With the remaining ones a Random Forest is trained to distinguish between measured data and simulation. In the ideal case of a good match between data and simulation the Random Forest should not be capable of separating the events from each other. However features with significant mismatches can lead to a good separation, which is indicated by an ROC-AUC above 0.5. These features can be identified as outliers in the feature importance distribution of the Random Forest. This procedure removes another 21 features.

The final feature set of 86 features is selected via the mRMR feature selection [6]. With this set the Random Forest is trained again, it shows a reduction of the AUC from 0.77 ± 0.01 to 0.56 ± 0.01 for the final feature set. The ROC curves for both feature sets can be compared in Figure 2.

4 Outlook

The actual classification of cascade like events is currently work in progress. The most limiting factor in this classification seems to be simulation statistics for atmospheric muons.

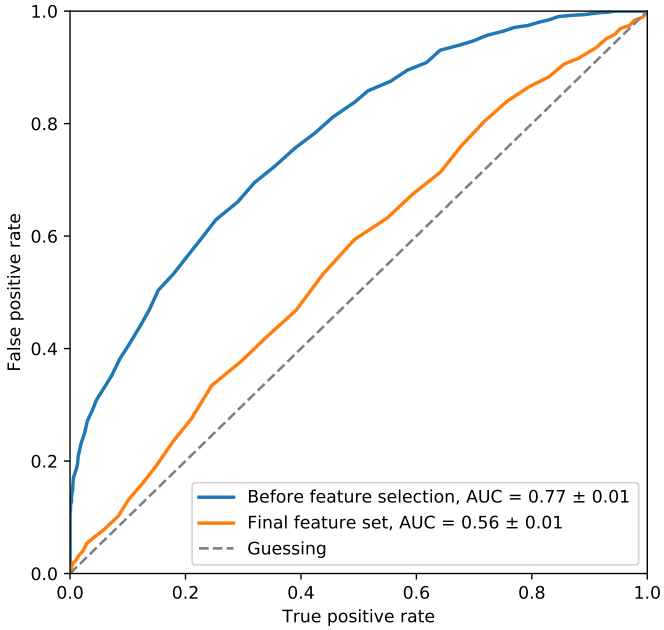


Figure 2: Comparison of the ROC Curve from a Random Forest trying to distinguish measured data from simulation using all features after removing redundant features with the ROC Curve for the final feature set.

References

[1] M. G. Aartsen et al. "Observation and Characterization of a Cosmic Muon Neutrino Flux from the Northern Hemisphere Using Six Years of IceCube Data". In: *Astrophys. J.* 833.1 (2016).

[2] M. G. Aartsen et al. "The IceCube Neutrino Observatory - Contributions to ICRC 2017 Part II: Properties of the Atmospheric and Astrophysical Neutrino Flux". In: (2017), pp. 86–93. arXiv: 1710.01191 [astro-ph.HE].

[3] M. Börner et al. "Measurement/Simulation Mismatches and Multivariate Data Discretization in the Machine Learning Era". In: *ADASS XXVII*. Ed. by TBD. Vol. TBD. ASP Conf. Ser. San Francisco: ASP, 2018, TBD.

[4] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[5] M. Meier. *Search for high-energy astrophysical Tau-Neutrinos in IceCube Data*. Technischer Report für das SFB 876 Gradiuertenkolleg. TU Dortmund University, 2016.

[6] H. Peng, F. Long, and C. Ding. "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (2005), pp. 1226–1238.

Stacking point source search of a lower energy neutrino contribution at the HESE track positions

Thorben Menne
Experimentelle Physik 5b
Technische Universität Dortmund
thorben.menne@tu-dortmund.de

With increasing connections between multiple sophisticated astronomical instruments in different messenger particles and energy ranges it is more promising than ever to find a consistent picture and mechanisms of astrophysical sources. With these increasing efforts many searches aim to combine these different information to find common sources of different astrophysical messenger particles.

Despite the discovery of multiple neutrinos of astrophysical origin with IceCube [1] no significant source of these high energy events has been found yet. Also no significant clustering of lower energy neutrinos at a single point has been found in an all sky search with 7 years of IceCube data [2]. Nevertheless recently found correlations between a high energy IceCube neutrino and a flaring Blazar [7] makes correlation searches in different messenger particles and energies even more promising.

This analysis aims to find a signal from lower energy neutrinos originating from the positions of high energy starting track events measured in IceCube. A stacking approach is used to collectively search for multiple weak emissions from the proposed source class. Both a time dependent and steady flux scenario are investigated using multiple years of all sky IceCube neutrino data.

1 Introduction

The IceCube detector is a cubic kilometer sized neutrino detector located at the south pole. It consists of 5160 digital optical modules (DOMs) mounted on 86 strings in depths between 1450 m to 2450 m directly in the antarctic ice. Additionally 81 instrumented ice tanks, the IceTop array, are installed on the surface to detect air showers [3]. Neutrinos can be measured indirectly via secondary particles produced in charged or neutral current interactions of the neutrinos with the surrounding matter. Those secondary particles then create tertiary particles in further interactions with the ice. All produced charged particles that are faster than the speed of light in the surrounding medium emit Cerenkov light which is measured and used to reconstruct the original neutrino properties.

With large neutrino detectors like IceCube the physics of astrophysical sources can be studied by measuring neutrinos originating from them. Currently the neutrino signal for single source positions on the sky is too low to be seen against the large background of atmospheric neutrinos. With the stacking method multiple sources of the same type are be bundled into one catalog. The combined signals have a better signal over background ratio than a single source so the time needed to measure a significant signal from a class of sources can be reduced. [4] To further reduce background contamination time information can be utilized in a time dependent likelihood approach.

2 Time independent search

The general likelihood approach in IceCube for a point source search is described in [6]. The most basic likelihood function $\mathcal{L}$ can be expressed as

$$\mathcal{L}(n_S) = \prod_{i=1}^{N} \left[\frac{n_S}{N} \, S_i + \left(1 - \frac{n_S}{N}\right) B_i \right] \tag{1}$$

where n_S is the number of assumed signal events, S_i, B_i the expected signal and background probabilities per event i and N the number of total events. The background is usually estimated by scrambling real data events in the azimuth which is justified by the assumption of only having a small amount of signal events.

For a stacked analysis only a small modification to the likelihood function (1) is needed, where the signal hypotheses is replaced by a sum over M weighted hypotheses for each source in the catalog

$$S_i \to S_i^{\text{tot}} = \frac{\sum_{j=1}^{M} W^j R^j S_i^j}{\sum_{j=1}^{M} W^j R^j} \tag{2}$$

Here W^j is the relative theoretical weight, R^j the relative detector acceptance for each source and S_i^j the signal probability for each event i regarding source j.

In this work the predefined source positions are given by the positions of measured high energy starting events, which are known to be of astrophysical origin [1]. It is then searched for spatial clustering of lower energy events in a larger sample at those source locations.

3 Time dependent search

For the time dependent search an extended Likelihood formalism [5] is used together with per event and source time information:

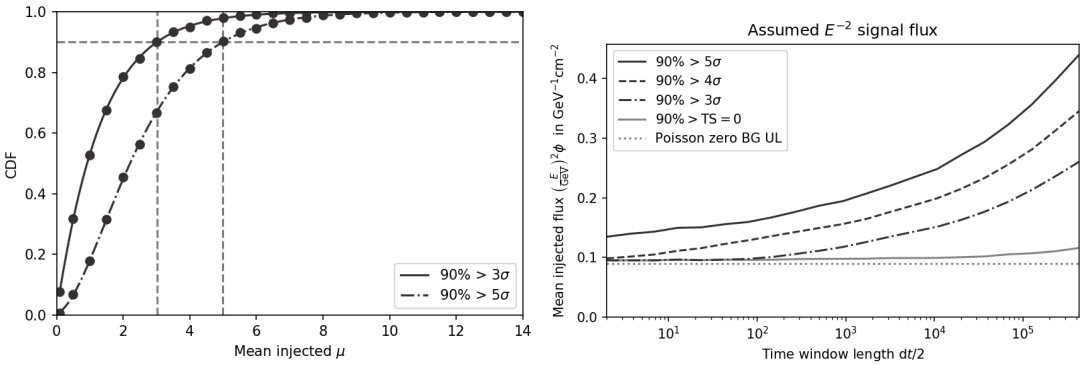
$$\mathcal{L}(n_S|\gamma) = \frac{(n_S + \langle n_B \rangle)^N \, e^{-(n_S + \langle n_B \rangle)}}{N!} \cdot \prod_{i=1}^{N} \left[\frac{n_S \sum_{j=1}^{N_{\text{srcs}}} w_j S_{ij}(\vec{x}_i | \vec{x}_{\text{src},j}, t_{\text{src},j}, E_i, \gamma) + \langle n_B \rangle B_i(\delta_i | E_i, \gamma)}{n_S + \langle n_B \rangle} \right] \quad (3)$$

where the sum over the sources is the stacking term as explained above. Here the only fit parameter is the number of expected signal events n_S from all source positions at once. The expected background events $\langle n_B \rangle$ are fixed for the fit and estimated from data measured outside the tested ranges. To include timing information in the Likelihood, the signal and background PDFs are time dependent only allowing events around a given time around the source locations to contribute to the signal or background PDF. The analysis tests multiple time windows ranging from 2 s to 5 d.

Again we can test if the measured data has a significant signal contribution by doing background trials. Here we need to inject background under consideration of poisson fluctuations which becomes more important in small time windows compared to the time-independent search. To test the analysis performance a-priori simulated signal events are injected until a predefined type-I and type-II error rate is matched. To better control fluctuations due to limited sample points of signal injected test statistics the mean number of events is systematically increased and a fixed number of trials with injected signal events are done. For each of these trials the percentile above the desired background only test statistic is calculated and then a χ^2 CDF is fitted to all points. Then the performance can be read of the fitted CDF. The plots show both the CDF fitting procedure and the resulting performance for all tested time windows.

4 Outlook

Currently systematic tests are done to estimate the impact on unknown effects on the analysis. On a short timescale this should be done and the analysis will be done on real



(a) Fitting a χ^2 CDF to the test statistic thresholds.

(b) Performance of this analysis regarding the flux that needs to be injected to state a detection with the depicted confidence in 90% of the experiments.

data. For the time independent analysis we still try to fit the source position with respect to the lower energy neutrinos.

References

[1] M. G. Aartsen et al. The IceCube Neutrino Observatory - Contributions to ICRC 2015 Part II: Atmospheric and Astrophysical Diffuse Neutrino Searches of All Flavors. In *Proceedings, 34th International Cosmic Ray Conference (ICRC 2015): The Hague, The Netherlands, July 30-August 6, 2015*, 2015.

[2] M. G. Aartsen et al. All-sky Search for Time-integrated Neutrino Emission from Astrophysical Sources with 7 yr of IceCube Data. *Astrophys. J.*, 835(2):151, 2017.

[3] M. G. Aartsen et al. The IceCube Neutrino Observatory: Instrumentation and Online Systems. *JINST*, 12(03):P03012, 2017.

[4] A. Achterberg et al. On the selection of AGN neutrino source candidates for a source stacking analysis with neutrino telescopes. *Astropart. Phys.*, 26:282–300, 2006.

[5] R. J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series)*. WileyBlackwell, reprint edition, 1989.

[6] Jim Braun, Jon Dumm, Francesco De Palma, Chad Finley, Albrecht Karle, and Teresa Montaruli. Methods for point source analysis in high energy neutrino telescopes. *Astropart. Phys.*, 29:299–305, 2008.

[7] R. Mirzoyan. First-time detection of VHE gamma rays by MAGIC from a direction consistent with the recent EHE neutrino event IceCube-170922A. *The Astronomer's Telegram*, 10817, October 2017.

Improving the Angular Resolution of FACT Using Machine Learning

Maximilian Nöthe
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
maximilian.noethe@tu-dortmund.de

The First G-APD Cherenkov Telescope (FACT) is pioneering the usage of solid state photo sensors (G-APDs aka SiPMs) for groundbased gamma-ray astronomy. FACT is located on the Roque de los Muchachos on the canary Island of La Palma. Since October 2011, the FACT collaboration has successfully been showing the reliability of SiPM for the IACT technique and blazar monitoring.

An important task of Imaging Cherenkov Gamma-ray Astronomy is the reconstruction of the direction of the measured gamma rays. Based on the recored Image of the Cherenkov photons, the origin of the gamma ray has to be estimated. This is an especially hard task for single telescopes.

The improvements of the angular resolution and sensitivity of FACT are presented, which were gained by using a Machine Learning based implemention of the so called disp method. For this approach, two Random Forests have been trained. First, a Random Forest Regressor to estimate the absolute value and second a Random Forest Classifier to perform the Head/Tail-Disambiguation of the shower.

1 Image Parameterization for FACT data

The FACT telescope records the fast flashes of Cherenkov light produced by extensive air showers in the atmosphere. Those flashes have a duration of roughly 20–60 nanoseconds. The FACT camera records a timeseries of 300 values for each of its 1440 pixels at a sample rate of 2 GSample/s, corresponding to 150 ns. [1]

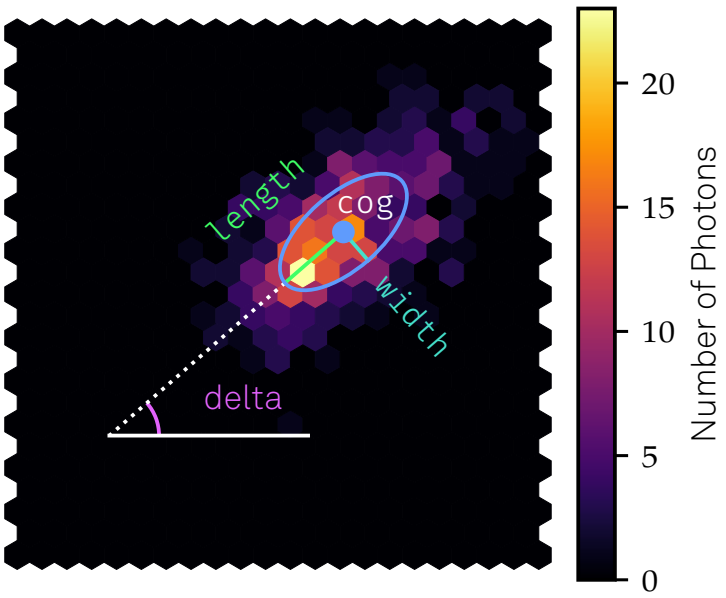


Figure 1: A sample FACT event with a visualisation of the Hillas parameters.

After the necessary calibration steps, that convert adc-values to voltages and remove artifacts, a first step of data reduction is performed. For each pixel, the total number of photons and their mean arrival time are estimated. In the following step, the resulting image is cleaned by removing low signal pixels, which probably do not contain any Cherenkov signal. The resulting image is now parameterized. The most common parameterization of cherenkov images are the Hillas parameters [6], show for a sample event in Figure 1. Additional features are also calculated and used to estimate the three most important target values in gamma-ray astronomy:

1. The particle type, where gammas form the signal and hadrons form the background class.
2. The particle energy
3. The particle's origin in the sky

The first task is a binary classification task, while the second is a one-dimensional regression. The third task is what will be dealt with in this report and is a two-dimensional regression.

2 Estimation of Gamma Ray Origin

One method to estimate the origin of a particle is the so called *disp method*, where the origin is estimated as a point on the main shower axis at a certain distance from the center of gravity of the light distribution on the main shower axis. This distance is called

`disp`. The task is now reduced to perform a regression of `disp`. The problem can be solved more easily by splitting it into the task of estimating the absolute value of `disp` and performing a binary classification for the sign of `disp`. Until now, this was done using simple formulas which were fit to simulated gamma-ray events.

To improve the angular resolution of FACT for reconstructing the origin of gamma-rays, the formulas are replaced by a random forest regressor for the absolute value of `disp` and a random forest classifier for its sign.

The two forests are trained on simulated gamma-ray events that are randomly scattered over the complete field of view of FACT. The extensive air showers and the Cherenkov light are simulated using the software CORSIKA [5] and the detector response, including ray-tracing and electronics simulation is done using CERES [4]. Raw data calibration and image parameterization is done using FACT-Tools [3], an extension of the `streams`-framework [2].

3 Results

The random forest regression for the absolute value of disk reaches a r^2-score of 0.6, the classifier for the sign of disp has an accuracy of 0.74. Figure 2 shows the angular resolution of the old analysis vs. the results of the machine learning approach. Figure 3 shows the detection significance of the Crab Nebula for both analysis versions. The significance was greatly increased by the new analysis.

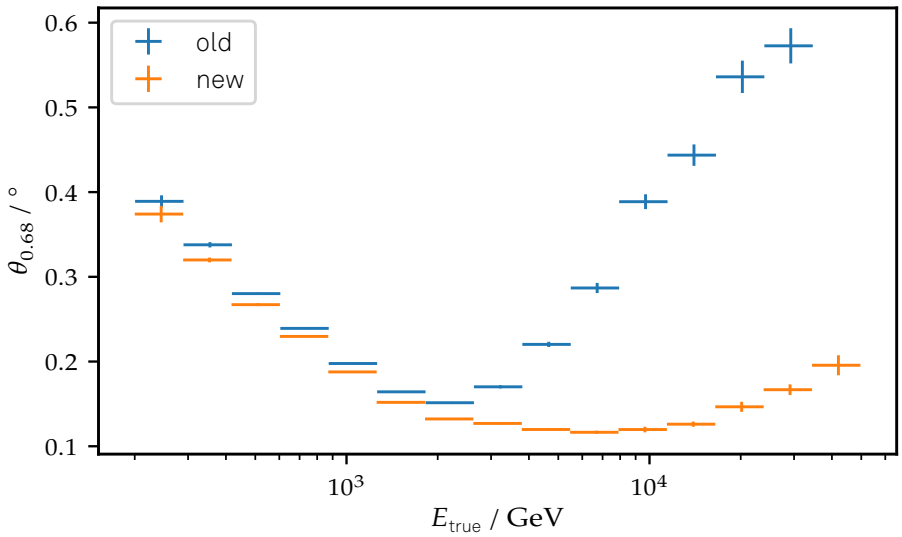
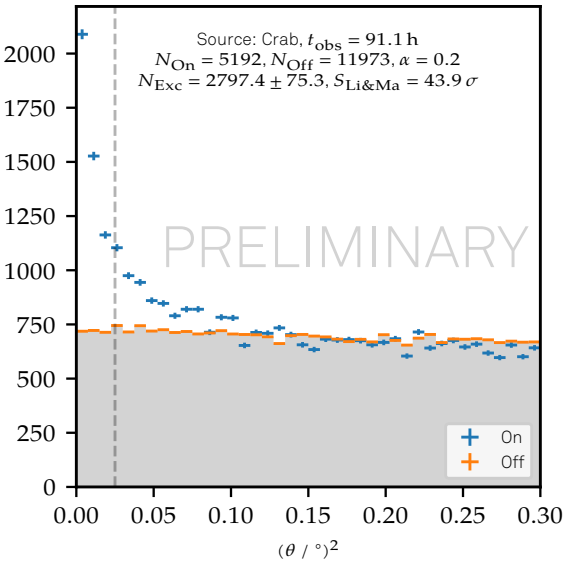
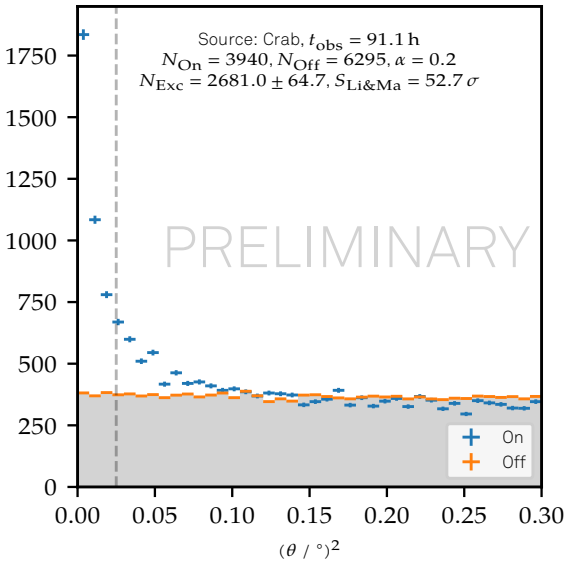


Figure 2: 68 % containment radius of reconstructed gamma-ray origin for different energy bins.



(a) Theta-Squared plot for the old analysis



(b) Theta-Squared plot for the new analysis

Figure 3: Source detection for both analysis versions

References

[1] H. Anderhub et al. Design and operation of fact–the first g-apd cherenkov telescope. *Journal of Instrumentation*, 8(06):P06008, 2013.

[2] Christian Bockermann and Hendrik Blom. The streams framework. Technical Report 5, TU Dortmund, 12 2012.

[3] Christian Bockermann, Kai Brügge, Maximilian Nöthe, et al. Fact-tools github-repository.

[4] Thomas Bretz and Daniela Dorner. Mars - cheobs goes monte carlo. In *Proceedings of the 31st ICRC*, 2009.

[5] D. Heck, G. Schatz, T. Thouw, J. Knapp, and J. N. Capdevielle. Corsika: A monte carlo code to simulate extensive air showers. Technical Report FZKA 6019, Forschungszentrum Karlsruhe, 1998.

[6] A Michael Hillas. Cerenkov light images of eas produced by primary gamma. In *Proceedings of the 19th International Cosmic Ray Conference*, volume 3, pages 445–448, 1985.

Improvements of the Lepton Propagator PROPOSAL

Jan Soedingrekso
Experimentelle Physik 5b
Technische Universität Dortmund
jan.soedingrekso@tu-dortmund.de

The simulation of high energy charged leptons propagating through the cubic kilometer scaled neutrino detector IceCube is done with the lepton propagator PROPOSAL. My task is to further improve the simulation library which splits into two topics: On the one hand physical aspects were enhanced to increase the precision of the propagation and reduce the systematic uncertainties. On the other hand programming aspects were improved to increase the performance and simplify the usage and maintenance.

1 Introduction

The IceCube Detector [1] detects Cherenkov light emitted by high energy charged leptons ($> 10\,\mathrm{GeV}$) with a detector volume of $1\,\mathrm{km}^3$ of glacial ice at the South Pole. These charged leptons are mainly produced as secondary products of cosmic rays interacting with the earth's atmosphere. The aim of IceCube is to detect neutrinos originating from astrophysical sources, which also produce charge leptons. The energy reconstruction of the neutrinos is directly correlated to the reconstructed energy of the charged leptons based on Monte Carlo simulations. To increase the performance of the energy reconstruction, the systematic uncertainties of the simulation have to be improved

The lepton propagator PROPOSAL[1] propagates charged leptons through a medium, while calculating the energy losses and the scattering. PROPOSAL (PRopagator with Optimal Precision and Optimized Speed for All Leptons) [6] implemented in C++ is the

[1] The code is available under `https://github.com/tudo-astroparticlephysics/PROPOSAL`.

successor of MMC (Muon Monte Carlo) [3] implemented in java. To further optimize the precision, physical improvements have been done, which are described in section 2. To further optimize the speed, a number of code improvements have been performed, which are described in section 3.

2 Physical challenges

Charged leptons lose their energy via the four processes ionization, bremsstrahlung, pair production and inelastic nuclear interactions. The average energy loss of these processes can be parametrized in a quasi-linear parametrization

$$-\langle\frac{\mathrm{d}E}{\mathrm{d}x}\rangle = a(E) + b(E)E \tag{1}$$

where a and b depend logarithmically on the energy. The Ionization, described by a, is nearly constant and dominates the energy losses up to 100 GeV. The other three energy loss processes, summarized by b, scale linear with the energy and dominate above around a TeV, where the nuclear interaction contributes only around 10 % to the energy loss compared to pair production and bremsstrahlung.

The bremsstrahlung cross section diverges for small energy losses, but not for the average energy loss. Therefore the energy losses are estimated *continuously* from the minimum energy loss to a specific energy cut with the average energy loss. From the energy cut to the maximum energy loss, the energy losses are sampled *stochastically*. This energy cut distort the propagation, which can be seen e.g. in the energy distribution of the energy losses or the propagated leptons [5]. One future target of this work is to quantify this distortion to an adjustable error to runtime ratio.

The uncertainties of the cross sections lie around 3 % exept of the nuclear interaction, which has an uncertainty of around 10 % [7]. One source of the uncertainties for all processes arise due to the effective description of the interaction with a nuclear target, which is difficult to improve. In addition to that, these cross sections only concern tree-level processes. An effort of calculating higher order corrections have been done for bremsstrahlung [8], where a 2 % difference in the average energy loss is obtained. One future goal of this work is to calculate higher order cross sections for the pair production cross sections.

Additional improvements in the decay routine and in the scattering routine were also implemented. Two new scattering parametrizations, introduced in [4], were (now correctly) implemented. In the old version a decay was always treated as a two body decay, which is a good aproximation for leptonic decays, where one charged lepton and two nearly massless neutrinos are produced. But for hadronic decays, where multiple massive particles can be produced, this is resulting in a step function in the energy distribution

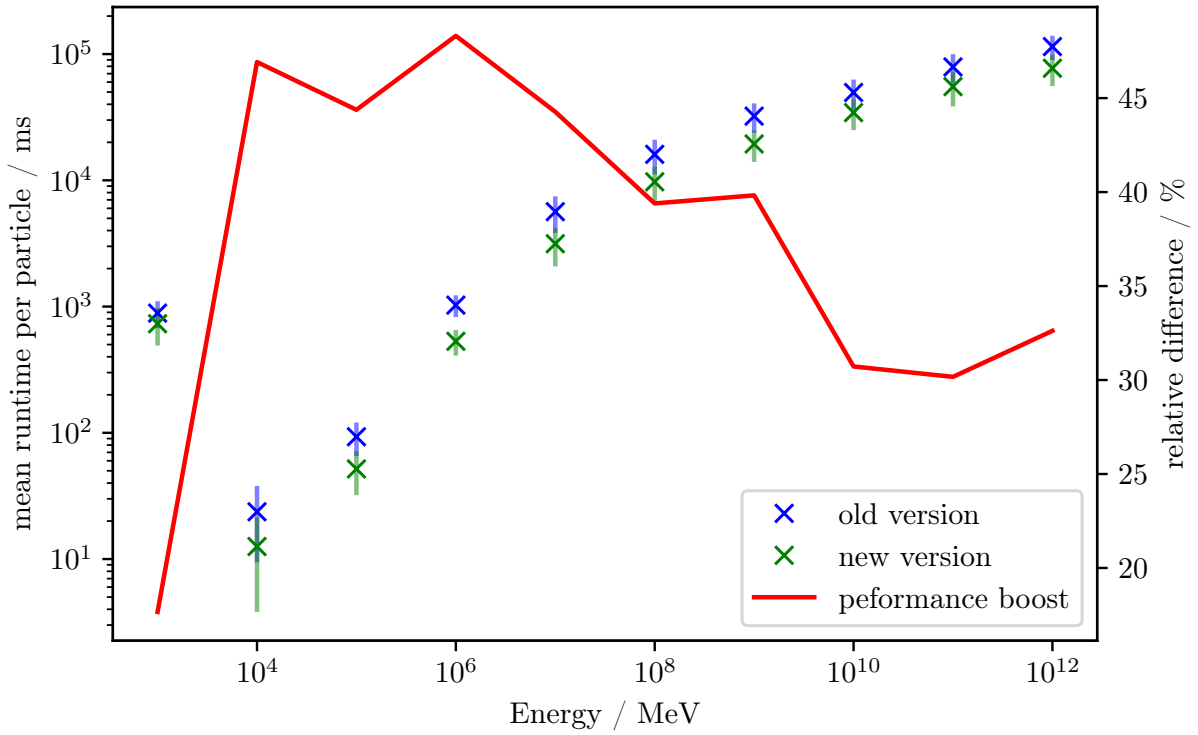


Figure 1: For each energy bin 100 Muons are propagated until they decay with the old and the new version. The runtime is plotted to the left axis. The mean values of the new version in relation to the old version, indicating the performance boost, are plotted to the right axis.

of the hadronic decay products. This artefact was removed by using the Raubold-Lynch algorithm [2] to estimate a multi-body decay.

3 Code improvements

The structure of the code was already improved to get rid of java structures [5]. Now, the Code was further transposed to a more C++ like programming style with the main difference, that the runtime type information was removed by introducing polymorphism. This makes it also easier to extend the library with new parametrizations for crosssections or scattering, particles, mediums and geometries. The unnecessary need of pointers was reduced, which reduces the heap allocation of memory and therefore increases the performance and reduces the danger of memory leaks. This is all done not only to increase the performance. This also makes the whole project much easier to maintain and for outstanding users easier to understand what happens.

The runtimes, propagating Muons of different energies with the old and the new version,

are shown in Figure 1. A significant performance boost of around 30 % is obtained. However, this is a conservative measurement, because for other scattering parametrizations, the performance further improves.

For a simplified usage, a python wrapper was added, so PROPOSAL is not only a C++ library, but also a python library. The surface can be controled in python, while there is a C++ library working behind it. Next to a python script, the user can set the detector configuration (media, geometries and energy cuts) in a configuration file together with some general settings like the scattering parametrization. The old selfwritten configuration file was changed to a json format, which is easy to understand and to read out, so it should be now very simple to use PROPOSAL.

References

[1] Abraham Achterberg et al., The IceCube Collaboration. "First year performance of the IceCube neutrino telescope". In: *Astroparticle Physics* 26.3 (2006), pp. 155–173. DOI: 10.1016/j.astropartphys.2006.06.007. arXiv: astro-ph/0604450.

[2] Eero Byckling and Keijo Kajantie. *Particle Kinematics*. Wiley, 1973.

[3] Dmitry Chirkin and Wolfgang Rhode. *Propagating leptons through matter with Muon Monte Carlo (MMC)*. 2004. arXiv: hep-ph/0407075.

[4] Malte Geisel-Brinck. "Revision of the multiple scattering algorithms in PROPOSAL". Bachelor thesis. TU Dortmund, 2013.

[5] Jan-Hendrick Koehne. "Der Leptonpropagator PROPOSAL". PhD thesis. TU Dortmund, 2013. DOI: 10.17877/DE290R-13191.

[6] Jan-Hendrick Koehne et al. "PROPOSAL: A tool for propagation of charged leptons". In: *Computer Physics Communications* 184 (2013), pp. 2070–2090. DOI: 10.1016/j.cpc.2013.04.001.

[7] Rostislav Pavlovich Kokoulin. "Uncertainties in underground muon flux calculations". In: *Nuclear Physics B (Proc. Suppl.)* 70.3 (1999), pp. 475–479. DOI: 10.1016/S0920-5632(98)00475-7.

[8] Alexander Sandrock et al. "Radiative corrections to the average bremsstrahlung energy loss of high-energy muons". In: *Physics Letters B* (2017). DOI: 10.1016/j.physletb.2017.11.047. arXiv: 1706.07242.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C4

Regression approaches for large-scale high-dimensional data

Katja Ickstadt Christian Sohler

On transferring the Merge & Reduce technique to regression models

Leo Geppert
Lehrstuhl Mathematische Statistik und biometrische Anwendungen
Fakultät Statistik, TU Dortmund
geppert@statistik.uni-dortmund.de

This technical reports introduces the versatile technique of Merge & Reduce, which so far has been primarily used on clustering algorithms, and presents ideas on transferring it to regression analysis. Some results and conclusions are discussed for both frequentist and Bayesian regression approaches. The approximation provided is good if the number of observations per variable is large enough.

Setting

The Merge & Reduce technique is a very general approach that can be employed to handle data sets that are too large to be processed in one go. It was introduced by [2] as a general method and has been adapted in Computer Science for clustering algorithms by [1] and [4]. More recently, Merge & Reduce is employed to efficiently analyse very large data sets [3] and in the design of streaming approximations with improved error guarantees [6].

The fundamental idea behind Merge & Reduce, given a data set of dimension n by p where k is large, is to perform the desired analysis sequentially on blocks of size n_b by p. On each of the blocks, the parameters or summary statistics of a statistical model are calculated and stored. The results from different blocks are combined in an efficient way, ensuring that the additional memory needed for these operations is bounded by $O(\log n)$. To that end, $L+1$ buckets $B_0, B_1, \ldots, B_L$ are introduced, where $L = \lceil \ln n \rceil$. In the first step, the first n_b observations are read, analysed and the results of the analysis stored in the *working bucket* B_0. The other buckets represent a hierarchy of results, where B_1

contains results based on one block, B_2 results based on two blocks and so on. After analysing a new block, the results are copied from B_0 to B_1 if B_1 is empty. Otherwise, the results in B_0 and B_1 are merged and saved in B_0, while B_1 is emptied. In this case, we continue the same process up the hierarchy until block B_i $(i = 1, \ldots, L)$ is empty. This ensures that only models from the same level of hierarchy are merged.

After handling the current n_b observations as described, the process is repeated iteratively with the following blocks of observations until the end of the data set or data stream is reached. Now, one final pass through all levels of hierarchy is required to obtain a final result. During the final pass, it may be necessary to merge models from different hierarchical levels.

The merged models need to be reduced to make sure that the complexity does not increase with the amount of data the models are based on, which would in turn increase the additional memory required.

Merge & Reduce is a versatile technique that can be used for different statistical models. In Computer Science, there is a focus on clustering methods. In the present report, we aim at transferring the method to regression models. To that end, it is necessary to fill the basic principle with life, i.e. find ways of representing regression models as well as merge these representations and reduce the complexity.

Approaches

We propose two approaches that differ for the type of statistical model we consider. Frequentist models consist of a parameter estimate and the estimated standard error as a measure of the estimate's variation. Bayesian models on the other hand usually consist of a distribution that can be characterised by measures of location and dispersion as well as by quantiles.

Approach 1, which is suitable in the frequentist case, consists of the estimated parameter $\hat{\beta}$ and its estimated standard error $\widehat{se}_\beta$. These two vectors represent the regression model well, they can be used to assess the influence of variables as well as predict new values $\hat{Y}$. For the merge-operation, we additionally require the number of observations the model is based on, giving a total of $2(p+1)+1$ parameters for a model with an intercept term where p is the number of variables.

Approach 2 needs to represent the posterior distribution, which makes it both more flexible and more dependent on the model in question. As a balanced solution that covers many cases, we suggest employing the posterior distribution's mean, median, and standard deviation as well as the posterior 2.5%, 25%, 75%, and 97.5% quantiles. Including the number of observations the model is based on, this results in $7(p+1)+1$ parameters.

For both approaches, the basic merge-operator and the reduce-operator are the same. Having obtained the above-mentioned summary values that represent the model, we merge two models by calculating the weighted mean for each of the summary values. We employ the number of observations the model is based on as weights. Towards the end of the data set or data stream, there may be blocks that represent fewer observations than expected, i.e. blocks with less than n_b observations. Weighting the results according to the number of observations downweighs such blocks, thus giving all observations the same importance. A separate merge-operator is not required as the number of parameters stays constant for every level by construction.

Results

We conduct a simulation study with a focus on linear regression and Poisson regression. It confirms that Merge & Reduce is able to recover the results of the original regression models, both for the frequentist case (using Approach 1) and for the Bayesian case (using Approach 2). For the measures of location ($\hat{\beta}$ for Approach 1, posterior mean and median for Approach 2), the only prerequisite is that there are enough observations per variable in the blocks, i.e. $\frac{n_b}{p}$ is large enough. For standard regression models, [5] suggests a ratio of $\frac{n}{p} \geq 10$ or preferably $\frac{n}{p} \geq 20$. Our simulation study indicates that this fraction needs to be slightly higher ($\frac{n_b}{p} \geq 40$) to achieve a good approximation with Merge & Reduce.

For the estimated standard error in the frequentist case and all quantiles except the median in the Bayesian case, after obtaining the final model, a correction factor that will be introduced in a forthcoming publication needs to be applied. With this factor, all of these measures behave in a similar way to the measures of location, i.e. when $\frac{n_b}{p}$ is high enough, the approximation using Merge & Reduce is good.

The posterior standard deviation in the Bayesian case is not well-approximated by Merge & Reduce. A useful correction could also not be found. For that reason, it seems advisable to employ suitable quantiles and not the standard deviation as a measure of the posterior distribution's dispersion.

All in all, the simulation study indicates that Merge & Reduce can be applied to a variety of regression models – for linear and Poisson regression as well as for frequentist and Bayesian models. While measures of location are well-recovered by both approaches, measures of dispersion (quantiles for Bayesian models) require a correction factor after obtaining the final model. In both cases, at least 40 to 50 observations per variable in regular blocks is required, a slight increase compared to classical algorithms.

Conclusion

Both approaches are able to recover the results of the original model well. For the parameter estimates and the measures of location, this works directly. For the estimated standard errors and other quantiles, suitable correction factors need to be applied. The only very mild restriction is that the ratio of observations per variable in regular blocks has to be at least 40 to 50.

Open questions are whether the approximation is good for all quantiles, e.g. for extreme quantiles like the posterior 99.9% quantile that may be of interest when considering outliers or extreme values.

The simulation study indicates that Merge & Reduce works well for both linear and Poisson regression models. It is not clear whether the technique can be applied to other regression models easily or whether some adaptations would be necessary. However, it seems plausible that other regression models can also be well-recovered provided, β plays a role similar to that in linear and Poisson regression.

References

[1] Mihai Bādoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proc. of STOC*, pages 250–257, 2002.

[2] Jon Louis Bentley and James B. Saxe. Decomposable searching problems I: Static-to-dynamic transformation. *J. Algorithms*, 1(4):301–358, 1980.

[3] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Proc. of SODA*, pages 1434–1453, 2013.

[4] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proc. of STOC*, pages 291–300, 2004.

[5] Frank E. Harrell, Jr. *Regression Modeling Strategies*. Springer-Verlag, 2001.

[6] Marc Heinrich, Alexander Munteanu, and Christian Sohler. Asymptotically exact streaming algorithms. *CoRR*, abs/1408.1847, 2014.

On Sampling and Sketching Methods for approximating Generalized Linear Regression Models

Alexander Munteanu
Efficient algorithms and complexity theory
TU Dortmund, CS department
alexander.munteanu@tu-dortmund.de

Our current research aims towards obtaining a common understanding of sampling and sketching methods for generalized linear regression models. The general approach is to reduce the size of the data via subsampling and reweighting the data or computing linear combinations of data points via sparse and fast linear sketches. The subsequent calculations can be carried out more efficiently on the data summary of small size. In some cases the existence of such summaries with rigorous approximation guarantees has been shown. In other cases, however, a worst case perspective establishes the impossibility of such summaries. We survey recent results and discuss the possible remedy of going beyond the worst case perspective.

Introduction to coresets and recent developments

Coresets and sketches are succinct data summaries which act as substitute for the original large data set or from which the required information can be derived efficiently in a machine learning or more generally optimization task. This leads us to the *sketch and solve paradigm* which is a general approach to reduce the data first in an offline or data stream setting to a small coreset or sketch and then to apply the machine learning algorithm on this succinct data representation.

Coresets usually depend on the considered objective function or at least a class of such functions. The first definitions were only implicitly given or were restricted to specific

problems like shape fitting problems [2], clustering [2, 8] and regression [3–5]. We give a more general Definition.

Definition 1 ([11, 12])**.** Let X be a set of points from a universe U and let Γ be a set of candidate solutions. Let $f : U \times \Gamma \to \mathbb{R}^{\geq 0}$ be a non-negative measurable function. Then a set $C \subset X$ is an η-coreset of X for f and some $\eta \geq 1$, if

$$\forall \gamma \in \Gamma : |f(X, \gamma) - f(C, \gamma)| \leq (\eta - 1) \cdot f(X, \gamma).$$

Note that the original point set is a 1-coreset but has linear size. Usually it is required to have at least sublinear, better polylogarithmic or constant size in the number of inputs and a small polynomial in the points' dimension to be a useful data reduction. Coresets that are exponential in the dimension have also been studied and are sometimes unavoidable but they are usually useful only in the context of constant dimension [1].

In a recent work we have extended the study of coresets to modern statistics and applications to machine learning methods like *dependency networks* [9], i.e., possibly cyclic probabilistic graphical models to study dependencies between random variables. Given our positive results on Normal distributions it is highly desirable to extend these methods to generalized linear models.

Coresets for Generalized Linear Models

A *generalized linear regression model* takes the form $\mathbf{E}[Y|x] = h(\beta^T x)$ for a non-linear link function h. The ordinary least squares linear regression model is only a simple special case in which h is the identity function. Encouraged by our positive sketching results on Bayesian linear regression based on Normal distributions [7], we have focused on extending our work to more general classes of distributions.

Among several possible generalizations we studied variants of generalized linear models based on the exponential family of probability distributions. For example, the Poisson regression model is frequently used for predicting non-negative integer valued count variables Y from independent real variables X. Another prominent example is the logistic regression model, which can be used to classify the examples two classes. While mere classification can be dealt with more efficiently and flexibly by the support vector machine classifier, logistic regression has a deeper statistical and probabilistic interpretation. Namely it is useful in estimating the probability of an event to happen, based on the observed data. For both, the logistic regression as well as the Poisson regression model we have shown impossibility results to hold regarding subsampling and sketching with rigorous approximation guarantees.

Theorem 2 ([12]). *For any $\delta > 0$ and any non-negative integer n there exists a set of n points P, such that any strong $(1 \pm \varepsilon)$-coreset of P for logistic regression must consist of $\Omega(n^{1-\delta})$ points.*

Theorem 3 ([11]). *Every η-coreset for Poisson regression with approximation factor $\eta < \frac{\exp(\frac{n}{4})}{2n^2}$ must consist of at least $\Omega(\frac{n}{\log n})$ points.*

Beyond worst-case analysis

Despite these worst-case impossibility results, coreset methods have proven useful in practical data mining and machine learning applications such as dependency networks to speed up the computations and lower their memory requirements. Coresets are also known to introduce an implicit regularization effect, making the models obtained from coresets not only close to the respective *optimal* models but even more accurate in their predictive performance on new data than the *optimal*.

This has been observed before by [10] and was confirmed in the empirical evaluation of [11] via a 10-fold cross-validation performed on traffic data. The benefits observed in these practical applications motivate the study of going *beyond* worst case analysis.

The first approach when confronted with a new model is to adapt the algorithm to deal with the changes. This approach has proved successful when dealing with generalized Normal distributions defined over ℓ_p-spaces [6]. This may not be possible for other generalizations as we have proved in the above impossibility results for GLMs.

We have identified two possible ways to tackle the information theoretic limitations encountered with these models:

1. Relaxing and exploring the underlying statistical model may provide insights into tractable subproblems [11]
2. Imposing natural assumptions on the input to exclude the hard instances but still include all practically relevant instances [13]

Conclusion and future research

Our research encounters methodological barriers. We have shown how to deal with these limitations and develop algorithms that have rigorous guarantees for practically relevant data settings. We want to further explore the possibilities given by algorithm design, statistical modeling and relaxation techniques. This pushes our knowledge towards a unified view on sublinear techniques and unified algorithmic blueprints for regression problems, one of the main targets of the third phase of SFB876.

References

[1] Pankaj K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. *Algorithmica*, 72(1):83–98, 2015.

[2] Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings of STOC*, pages 250–257, 2002.

[3] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM Journal on Computing*, 38(5):2060–2078, 2009.

[4] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proc. of SODA*, pages 1127–1136, 2006.

[5] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

[6] Leo N. Geppert, Katja Ickstadt, Steffen Müller, Alexander Munteanu, and Jonathan Rathjens. Random projections for various Bayesian regression generalizations. *Unpublished manuscript (to appear)*, 2017.

[7] Leo N. Geppert, Katja Ickstadt, Alexander Munteanu, Jens Quedenfeld, and Christian Sohler. Random projections for Bayesian regression. *Statistics and Computing*, 27(1):79–101, 2017.

[8] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In László Babai, editor, *STOC*, pages 291–300. ACM, 2004.

[9] D. Heckerman, D. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for density estimation, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–76, 2000.

[10] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[11] Alejandro Molina, Alexander Munteanu, and Kristian Kersting. Core dependency networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2018. (to appear).

[12] Alexander Munteanu and Chris Schwiegelshohn. Coresets - methods and history. *Invited article in KI, special issue on Algorithmic Challenges and Opportunities of Big Data*, (to appear).

[13] Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David Woodruff. Sketching and sampling algorithms for logistic regression. *Unpublished manuscript (to appear)*, 2017.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C5

Real-Time Analysis and Storage of High-Volume Data in Particle Physics

Bernhard Spaan Jens Teubner

Measurement of *CP* violation in $B_s^0 \to D_s^{\mp} K^{\pm}$ decays with the LHCb experiment

Ulrich Eitschberger
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
ulrich.eitschberger@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP* violation in decays of beauty and charm hadrons. In order to maximize the sensitivity with respect to these specialized targets the LHCb detector is built as a forward single arm spectrometer.

In the collisions of protons inside the vertex locator (VELO), new particles are created and decay until they finally leave traces in the various subcomponents of the detector. These traces are hits in the tracking systems (VELO, TT, T1-T3, M1-M6), clusters in the calorimeters (ECAL, HCAL) and Cherenkov radiation in the Ring Imaging Cherenkov Detectors (RICH1, RICH2). To conclude on the presence of particles the information of these subdetectors needs to be reconstructed, e.g. by fits of trajectories to ensembles of hits and pattern recognition algorithms looking for clusters of energy deposition. Finally, particle candidates need to be combined to heavier particles in order to perform physics measurements on the same. The endeavour to find the particles of interest is hindered for two main reasons. Firstly, hundreds of particles are produced inside the angular acceptance which allows for a large number of combinations to be made in the reconstruction as well as the combination of particles. Secondly, the interaction rate of 50ns / 25ns together with the limitation on the bandwidth that can be written to disk enforces a fast reconstruction that leads to the selection of interesting events and the rejection of physically uninteresting ones. Both these points together set the frame for investigating these tasks in the context of resource limitation: The reconstruction and combination tasks can be parallelized and therefore performed faster.

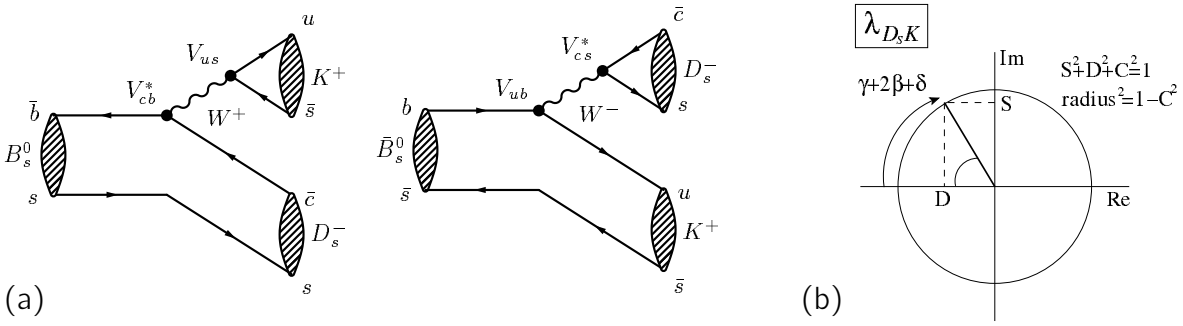


Figure 1: (a) Feynman diagrams of B_s^0 and $\overline{B}_s^0$ mesons decaying into the final state $D_s^- K^+$. (b) Illustration of the parameter λ_{D_sK} in the complex plane, together with the *CP*-asymmetry observables S, C and D mentioned in the text.

In order to demonstrate the need for an improved processing of the data collected by the LHCb detector, a new physics analysis is being performed that studies time-dependent charge-parity (*CP*) violation. *CP* violation is one of the keys for understanding the matter-antimatter-asymmetry observed in our universe. and precision measurements of *CP* violation allow to test the CKM sector of the Standard Model (SM). The tree-level decay $B_s^0 \to D_s^{\mp} K^{\pm}$ provides sensitivity to the CKM angle $\gamma = \arg\left[-\frac{V_{ud}\, V_{ub}^*}{V_{cd}\, V_{cb}^*}\right]$ without theoretical complications of penguin contributions or other hadronic uncertainties [4, 5]. The sensitivity to the CKM angle γ arises from the interference of $b \to u$ and $b \to c$ amplitudes (see Fig. 1a). A time-dependent analysis of $B_s^0 \to D_s^{\mp} K^{\pm}$ decays provides a complementary measurement of γ with respect to the time-integrated analyses using the $B \to DK$ decays, where the charged and neutral B modes need different auxiliary inputs in the final determination of γ. In $B_s^0 \to D_s^{\mp} K^{\pm}$ decays the mixing of the neutral B_s^0 provides the interfering amplitudes and gives access to $\gamma - 2\beta_s$ (see Fig. 1b), where β_s is the mixing phase. The mixing phase was measured accurately and is used as an external input to the determination of the CKM angle γ. The presented new measurement uses a data set corresponding to 3.0 fb^{-1} of pp collisions recorded with the LHCb detector at $\sqrt{s} = 7$ and 8 TeV in 2011 and 2012, respectively.

The analysis is performed by fitting the decay-time-dependent decay rates of the initially produced flavour eigenstates $|B_s^0(t=0)\rangle$ and $|\overline{B}_s^0(t=0)\rangle$, which are proportional to

$$\frac{\mathrm{d}\Gamma_{B_s^0 \to f}(t)}{\mathrm{d}t} \propto e^{-\Gamma t}\left[\cosh\left(\frac{\Delta\Gamma_s t}{2}\right) + D_f \sinh\left(\frac{\Delta\Gamma_s t}{2}\right) + C\cos\left(\Delta m_s t\right) - S_f \sin\left(\Delta m_s t\right)\right],$$
$$\frac{\mathrm{d}\Gamma_{\overline{B}_s^0 \to f}(t)}{\mathrm{d}t} \propto e^{-\Gamma t}\left[\cosh\left(\frac{\Delta\Gamma_s t}{2}\right) + D_f \sinh\left(\frac{\Delta\Gamma_s t}{2}\right) - C\cos\left(\Delta m_s t\right) + S_f \sin\left(\Delta m_s t\right)\right].$$

Here, Γ is the average B_s^0 decay width and $\Delta\Gamma_s$ is the positive [1] decay-width difference between the heavy and light mass eigenstates in the B_s^0 system. Similar equations can be written for the *CP* conjugate decays replacing S_f by $S_{\bar{f}}$, and D_f by $D_{\bar{f}}$, while $C = -C_{\bar{f}}$

under the assumption of no *CP* violation in either the decay or mixing amplitudes. The sinusoidal *CP* observables can only be measured using signal candidates in which the initial flavour of the B_s^0 meson is determined, a process known as "flavour tagging". By contrast, all signal candidates provide sensitivity to the hyperbolic *CP* observables. The six *CP* observables are related to the magnitude of the amplitude ratio $r_{D_sK} \equiv |\lambda_{D_sK}| = |A(\overline{B}_s^0 \to D_s^- K^+)/A(B_s^0 \to D_s^- K^+)|$, the strong phase difference δ, and the weak phase difference $\gamma - 2\beta_s$ by the following equations:

$$
\begin{aligned}
C =& \frac{1 - r_{D_sK}^2}{1 + r_{D_sK}^2}\,, \\
D_f = \frac{-2r_{D_sK}\cos(\delta - (\gamma - 2\beta_s))}{1 + r_{D_sK}^2}\,, &\quad D_{\bar{f}} = \frac{-2r_{D_sK}\cos(\delta + (\gamma - 2\beta_s))}{1 + r_{D_sK}^2}\,, \\
S_f = \frac{2r_{D_sK}\sin(\delta - (\gamma - 2\beta_s))}{1 + r_{D_sK}^2}\,, &\quad S_{\bar{f}} = \frac{-2r_{D_sK}\sin(\delta + (\gamma - 2\beta_s))}{1 + r_{D_sK}^2}\,.
\end{aligned}
\tag{1}
$$

The analysis strategy largely follows that described in Ref. [2]. The fit to the decay-time distribution of the B_s^0 candidates is performed in a statistically background subtracted way using *sWeights* [6]. The *sWeights* are computed in a multidimensional fit to the invariant B_s^0 and D_s^- mass distributions as well as the particle identification distribution of the bachelor particle. Furthermore, the kinematically similar, but flavour-specific decay $B_s^0 \to D_s^- \pi^+$ is used to determine the decay-time-dependent efficiency and the flavour-tagging calibration. Prompt D_s^- decays combined with a random primary vertex track are used to calibrate the decay-time resolution. Figure 2 shows the fitted decay-time distribution of the $B_s^0 \to D_s^\mp K^\pm$ signal candidates. The analysis finds 5955 ± 90 signal candidates and the following *CP* parameters

$$
\begin{aligned}
C &= 0.73 \pm 0.14 \pm 0.05, \\
D_f &= 0.39 \pm 0.28 \pm 0.15, \qquad & D_{\bar{f}} &= 0.31 \pm 0.28 \pm 0.15\,, \\
S_f &= -0.52 \pm 0.20 \pm 0.07, & S_{\bar{f}} &= -0.49 \pm 0.20 \pm 0.07\,.
\end{aligned}
$$

From the measured *CP* parameters the CKM angle $\gamma = (128\,^{+17}_{-22})^\circ$ is determined (see Fig. 2), which corresponds to $3.8\,\sigma$ evidence for *CP* violation in the interference between decay and decay after mixing. The analysis has been submitted to the Journal of High Energy Physics (JHEP) and has already been published on arXiv [3].

The LHCb experiment will multiply the amount of collected data during the upcoming LHC Runs. Currently, the available LHCb dataset is already two times larger than that used in the presented analysis. In order to fully exploit the potential of these large datasets, analysis tools need to be scalable in terms of both storing and processing. In the presented analysis advantages could arise especially in the selection steps, both the signal and the flavour tagging particles. Improvements in the handling of large amounts of data (for example using Apache Hadoop and Apache Flink) would make looser preselections possible and thus retain more potential signal candidates. Necessary to achieve this

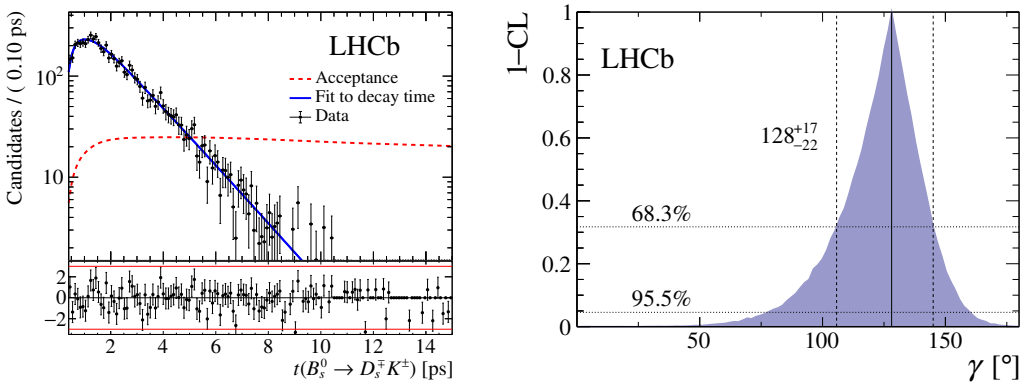


Figure 2: On the left the decay-time distribution of the $B_s^0 \to D_s^\mp K^\pm$ candidates is shown. The graph on the right illustrates the $1-\mathrm{CL}$ for the angle γ, together with the central value and the 68.3% confidence level (CL) interval.

is the transformation of the experiments data from existing data formats into modern versatile data formats like HDF5. A cut-based offline selection could then be replaced by a selection based on more sophisticated machine learning techniques. Network bottlenecks due to heavy data access can be overcome by exploiting the individual storage of cluster nodes. Making use of the MapReduce programming model or even more advanced ideas at the same time should lead to reasonable processing times.

References

[1] R. Aaij et al. Determination of the sign of the decay width difference in the B_s^0 system. *Phys. Rev. Lett.*, 108:241801, 2012.

[2] R. Aaij et al. Measurement of *CP* asymmetry in $B_s^0 \to D_s^\mp K^\pm$ decays. *JHEP*, 11:060, 2014.

[3] Roel Aaij et al. Measurement of *CP* asymmetry in $B_s^0 \to D_s^\mp K^\pm$ decays. 2017. `https://arxiv.org/abs/1712.07428`.

[4] Isard Dunietz and Robert G. Sachs. Asymmetry Between Inclusive Charmed and Anticharmed Modes in B^0, Anti-B^0 Decay as a Measure of CP Violation. *Phys.Rev.*, D37:3186, 1988.

[5] Robert Fleischer. New strategies to obtain insights into CP violation through $B_{(s)} \to D_{(s)}^\pm K^\mp$, $D_{(s)}^{*\pm} K^\mp$, ... and $B_{(d)} \to D^\pm \pi^\mp$, $D^{*\pm}\pi^\mp$, ... decays. *Nucl.Phys.*, B671:459–482, 2003.

[6] Muriel Pivk and Francois R. Le Diberder. sPlot: A statistical tool to unfold data distributions. *Nucl.Instrum.Meth.*, A555:356–369, 2005.

Reoptimisation of LHCb Flavour-Tagging algorithms

Kevin Heinicke
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
kevin.heinicke@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP* violation in decays of beauty and charm hadrons. In order to maximize the sensitivity with respect to these specialized targets the LHCb detector is built as a forward single arm spectrometer (see Figure 1).

In the collisions of protons inside the vertex locator (VELO), new particles are created and decay until they finally leave traces in the various subcomponents of the detector. These traces are hits in the tracking systems (VELO, TT, T1-T3, M1-M6), clusters in the calorimeters (ECAL, HCAL) and Cherenkov radiation in the Ring Imaging Cherenkov Detectors (RICH1, RICH2). To conclude on the presence of particles the information of these subdetectors needs to be reconstructed, e.g. by fits of trajectories to ensembles of hits and pattern recognition algorithms looking for clusters of energy deposition. Finally, particle candidates need to be combined to heavier particles in order to perform physics measurements on the same.

The endeavour to find the particles of interest is hindered for two main reasons. Firstly, hundreds of particles are produced inside the angular acceptance which allows for a large number of combinations to be made in the reconstruction as well as the combination of particles. Secondly, the interaction rate of 50 ns / 25 ns together with the limitation on the bandwidth that can be written to disk enforces a fast reconstruction that leads to the selection of interesting events and the rejection of physically uninteresting ones. Both these points together set the frame for investigating these tasks in the context of resource limitation: The reconstruction and combination tasks can be parallelized and therefore performed faster.

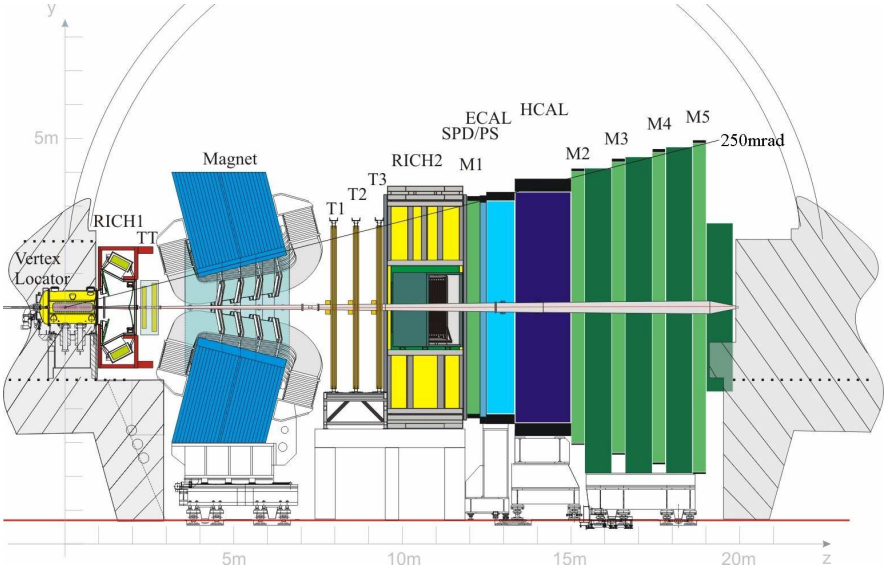


Figure 1: The LHCb detector with the various subdetectors for the identification of particles and reconstruction of their tracks [2].

A major subset of the LHCb analyses involve studies of *CP* violation, which can be used to test the Standard Model of particles physics (SM). These measurements include time-dependent decay rate measurements, many of which are subject to mixing of neutral B meson states. The knowledge of the initial flavour of these mesons is therefore crucial. It is being extracted with several Flavour Tagging algorithms, which are executed after the particle-decays have been fully reconstructed. The algorithms are designed to reconstruct particles on the same-side (SS) and opposite-side (OS) of the signal B candidate. The charge information of these particles is correlated with the initial flavour of the B candidate via different weak transitions. The different types of algorithms are depicted in figure 2.

In general each algorithm aims to identify a decay product either from the OS non-signal B meson, or from a SS hadron which hadronised with the signal B meson by applying different selection criteria. The selected particles are referred to as tagging particles. The algorithms differ in their specific decay product and selection strategies. Due to the large number of tracks which do not necessarily originate from the primary vertex, this process is error prone. The quality of the identification is therefore evaluated by calculating an overall tagging efficiency

$$\varepsilon_{\text{tag}} = \frac{N_{\text{tagged}}}{N_{\text{tagged}} + N_{\text{untagged}}}\,, \tag{1}$$

describing the ratio of B candidates for which a tagging particle has been found within

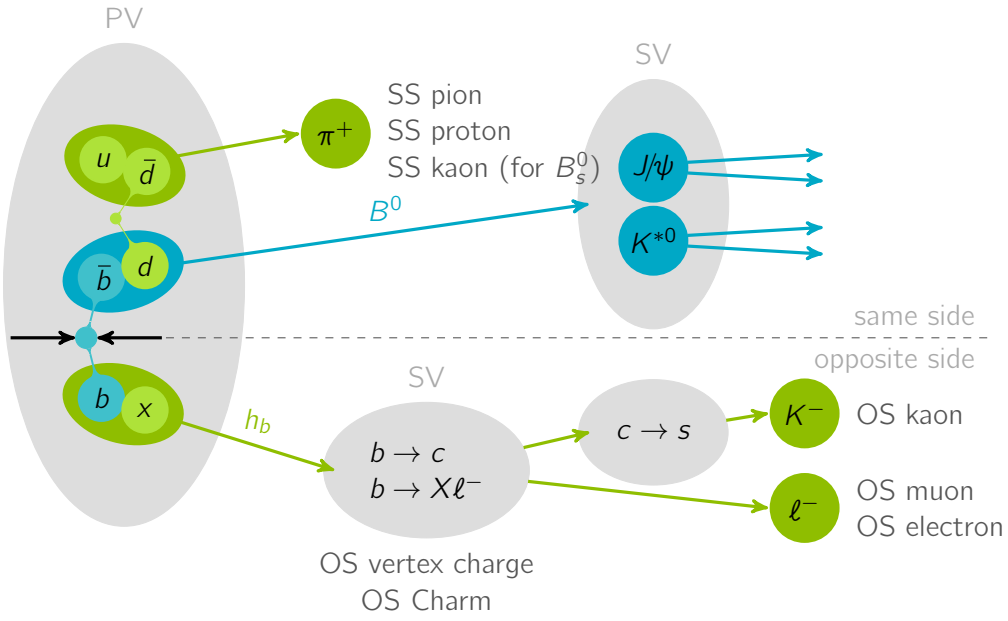


Figure 2: Schematic description of different Flavour Tagging algorithms. The same side algorithms infer the initial b flavour from pions and protons that hadronise alongside with the B meson. The opposite side algorithms infer this information from the non-signal B partner and its decay into leptons or $b \to c \to s$ transitions.

all B candidates, and the mistag rate

$$\omega = \frac{N_{\text{incorrect}}}{N_{\text{tagged}}}\,, \tag{2}$$

which is the rate of incorrectly tagged events or false-positive rate. The overall performance of these algorithms is usually described in terms of the tagging power

$$\overline{\varepsilon}_{\text{eff}} = \varepsilon_{\text{tag}}\,(1-2\omega)^2\,. \tag{3}$$

To estimate the tagging power on untagged data, the mistag rate ω_i is predicted on a per-event basis by different multivariate analysis tools for each individual tagging algorithm, leading to the final figure of merit for Flavour Tagging, the per-event tagging power

$$\varepsilon_{\text{eff}} = \frac{1}{N_{\text{tagged}} + N_{\text{untagged}}} \sum (1-2\omega_i)^2\,. \tag{4}$$

After the centre-of-mass energy has been upgraded to $\sqrt{s} = 13\,\text{TeV}$ for the LHC Run 2 in 2015, these algorithms were re-factored and re-optimised to adopt for the new physics environment. Given the fact that the full event reconstruction, including the Flavour Tagging algorithms might be moved into a full-software trigger in an upcoming LHCb upgrade, the re-factoring is especially valuable. With a High Level Trigger event rate of

several kHz and an increased pile-up with multiple primary vertices per event [1] a highly efficient execution of Flavour Tagging algorithms is necessary.

One major part of the re-factoring included the extraction of common read operations from the transient event store (TES) which is implemented in the Gaudi framework [5] into a common interface. While very specific implementation features can still reside in the algorithms code, common calls are abstracted into this interface. Moreover the implementation of different multivariate analysis frameworks has been simplified and the reliance on the TMVA framework (part of the ROOT framework [3]) has been reduced. Especially the computational expensive evaluation of large boosted decision trees has been reduced by using smaller BDTs which were trained with XGBoost and offer a similar performance. [4].

Furthermore a re-optimisation strategy has been defined for all single-track tagging algorithms which use a cut-based tagging particle selection (OS muon, OS electron, OS kaon). These algorithms suffered from the increased track multiplicity after the energy bump in LHC Run 2 and the tagging power of these algorithms decreased by up to 50 %. Therefore the cut selection as well as the mistag prediction have been re-trained which recovered the tagging power loss up to 5 %. Future re-optimisation iterations might even surpass the Run 1 performance. With the fixed strategy, this process can now be carried out in a fraction of the time previously needed.

References

[1] LHCb Trigger and Online Upgrade Technical Design Report. Technical Report CERN-LHCC-2014-016. LHCB-TDR-016, May 2014.

[2] A. A. Alves, Jr. et al. The LHCb detector at the LHC. *JINST*, 3:S08005, 2008.

[3] Rene Brun and Fons Rademakers. Root - an object oriented data analysis framework. In *AIHENP'96 Workshop, Lausane*, volume 389, pages 81–86, 1996.

[4] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[5] Marco Clemencic, Hubert Degaudenzi, Pere Mato, Sebastien Binet, Wim Lavrijsen, Charles Leggett, and Ivan Belyaev. Recent developments in the lhcb software framework gaudi. *Journal of Physics: Conference Series*, 219(4):042006, 2010.

Distributed Physical Data Analysis using Apache Drill

Michael Kußmann
Lehrstuhl für Datenbanken und Informationssysteme
Technische Universität Dortmund
michael.kussmann@cs.tu-dortmund.de

This report reports on advancements in *DeLorean*, a data analysis framework for particle physics. In cooperation with physicist from the LHCb project at CERN, modern database technology is being deployed in the field of physical data analysis. *DeLorean* now supports a distributed execution mode. The experiments show a linear scalability in both node and CPU core count.

Introduction

Today, particle physics is a data heavy science, where petabytes of data are being analysed to find a "needle in a haystack". In times of the Big Data revolution, that task does not sound too difficult, but standard database processors lack the expressiveness that would be needed to support the complex query patterns that arise in particle physics. As a joint effort with physicists from the LHCb experiment [3], *DeLorean* [4] is designed to bridge the gap between complex analyses and query capabilities.

In previous work [4], our group already reported on *DeLorean* and showed how *DeLorean* can improve data access by several factors already in single-node configurations. This report shows how *DeLorean* scales to a distributed cluster setup, leveraging the advantages of Hadoop-style data processing.

1 System Overview

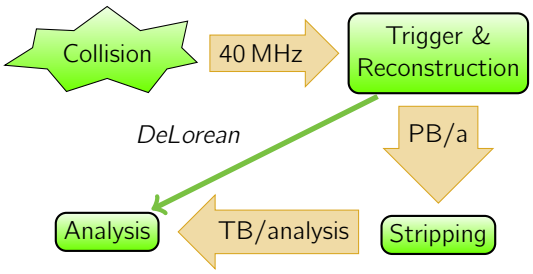


Figure 1: Processing pipeline at CERN.

DeLorean is a framework to analyse large datasets in particle physics at scale. It builds upon the distributed Apache Drill [1] and Hadoop Distributed File System (HDFS) platforms. Apache Drill is a storage agnostic SQL query engine for Hadoop and the open source counterpart to Google's Dremel system [5]. It enables relational-style data processing at massive scale. To optimize scan volume, *DeLorean* employs a columnar storage implemented by Apache Parquet [2]: Typical queries only filter on few columns of the data. A columnar storage allows to just read the necessary columns and to discard events without reading them entirely. Additionally, a columnar storage allows for better compression efficiency and data locality, which can be crucial facing petabytes of data.

1.1 DeLorean Characteristics

factor	1	2	4	8	16	24	32
ROOT (GB)	4.8	9.6	19.2	38.5	76.9	115.2	153.9
Parquet (GB)	4.0	8.1	16.1	32.3	64.6	96.6	129.2

Table 1: Data size in GB **for different scale factors.**

To showcase the characteristics of *DeLorean*, experiments with data sets of various sizes have been implemented on a test environment in our lab. Table 1 lists the characteristics of the subsets that were extracted from a real-world dataset from the LHCb experiment. As discussed above, the LHCb collaboration uses ROOT as their native storage format. Data is being converted to Parquet to enable further processing by *DeLorean*.

DeLorean uses a columnar storage format. As can be seen in the table, this results in on-disk data sizes that are about 15 % smaller than those of ROOT, the base format. This is remarkable as the ROOT format uses heavy-weight LZMA compression while Parquet uses a light-weight approach.[1]

1.2 Hardware

The test platform is a virtualised Hadoop cluster consisting of seven nodes, each having access to 128 GB of main memory and four CPU cores of an Intel Xeon E5-2697v2

[1] In this case: Run-length encoding, Bit packing and Snappy

system. Each node further has access to 7 TB of local hard drive space and to a 10 Gbit/s network connection.

The Apache Drill deployment contains a dedicated master node that runs Apache ZooKeeper to coordinate the Drill nodes and a name node for HDFS. That leaves six worker nodes to be used by Drill in distributed mode and HDFS data nodes. The Drill instances on the nodes cannot allocate the system memory exclusively: Drill is limited to a Java heap size of 16 GB and direct memory allocation of 24 GB.

2 Scalability with Data Sizes

DeLorean is designed to scale well with very large data volumes. Figure 2, evaluates the throughput that can be achieved for the data sets of Table 1. Throughput thereby refers to the I/O bandwidth that the physicists' current solution would have to reach when processing ROOT files.

One observation from Figure 2 is that *DeLorean* achieves throughput rates beyond 1 GB/s.

A second observation is that *DeLorean* requires sufficiently large data sets to excel. For small scale factors, individual work units (per default drill is allowed to start up to 1000 worker threads) become too small to hide latency overheads of the magnetic drives. In practice, *DeLorean* will operate on even much larger data volumes, that is, right in the system's sweet spot.

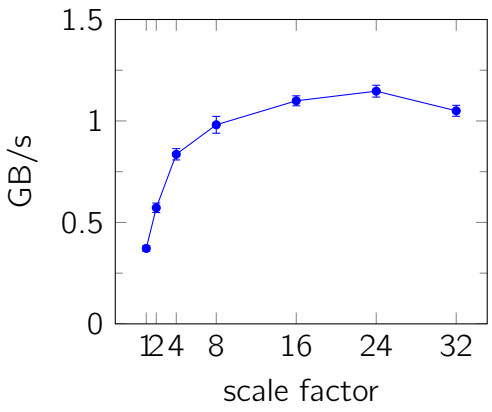


Figure 2: Influence of job size on the scan bandwidth using Drill's default setting planner.width.max_per_query=1000 (cf. Section 2).

3 Scale Out

The decisive aspect of *DeLorean*, however, is its readiness to scale horizontally (by adding more nodes/cores to the installation). The physical size of the cluster is fixed to six/seven nodes. For this experiment Drill's planner.width.max_per_query configuration option is being varied. This option limits the number of worker threads Drill is allowed to start globally per query. The test cluster has six worker nodes with up to four CPU cores each, which suggest a performance peak at 24 threads.

Figure 3 shows the results of the experiment using a scale factor of 16. For a value of n = planner.width.max_per_query less than six, queries can only run on n nodes concurrently.

During the lifetime of a query, those are not necessarily always the same nodes. Values of n greater six allow Drill to spawn more scanner threads than there are nodes. This makes sense, as each node has four CPU cores at disposal. Figure 3 shows a linear growth in scan bandwidth for $1 \leq n \leq 24$. This conforms to the cluster having access to a total of 24 CPU cores, suggesting linear scalability of the approach.

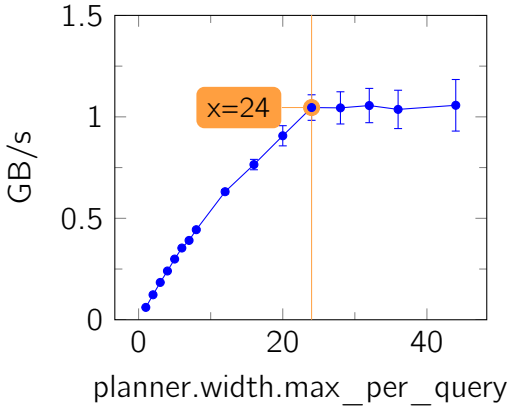


Figure 3: Influence of planner.width.max_per_query on the scan bandwidth.

Conclusion and Future Work

DeLorean is a framework that leverages database and MapReduce technologies to support analyses as they arise in particle physics. *DeLorean* provides an SQL-like query interface on top of a relational data representation and achieves linear scalability on cluster hardware.

Next steps include mechanisms to automatically extract queries from formula-style analysis representations. This will help to integrate *DeLorean* seamlessly into the analysis pipelines at CERN.

References

[1] Apache drill - schema-free sql for hadoop, nosql and cloud storage, 2016. `https://drill.apache.org`, 15.09.16.

[2] Apache parquet, 2016. `https://parquet.apache.org/`, 09.09.16.

[3] Lhcb collaboration, 2017. `http://lhcb.web.cern.ch/lhcb/`, 24.11.17.

[4] Michael Kußmann, Maximilian Berens, Ulrich Eitschberger, Ayse Kilic, Thomas Lindemann, Frank Meier, Ramon Niet, Margarete Schellenberg, Holger Stevens, Julian Wishahi, Bernhard Spaan, and Jens Teubner. Delorean: A storage layer to analyze physical data at scale. In *Datenbanksysteme für Business, Technologie und Web (BTW 2017)*, 2017.

[5] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. Dremel: Interactive analysis of web-scale datasets. *Proc. VLDB Endow.*, 3(1-2):330–339, September 2010.

Efficient Parallel Processing of High-Volume Scientific Data Streams on Low Power Hardware Devices

Thomas Lindemann
Lehrstuhl für Datenbanken und Informationssysteme (DBIS)
Technische Universität Dortmund
thomas.lindemann@cs.tu-dortmund.de

In the LHCb Project, a continuous stream of hits is produced by the several stages of the LHCb detector, which have to be processed in real time, since there are no capabilities to store all collision events permanently with the current storage technology. Thus, the High Level Trigger (HLT) needs to select events that have to be stored for further analysis. At the moment the average time budget for storing or one event is 50 ms, assuming a event frequency of 20 MHz, but it will be increased to 40 MHz. In our research, we are evaluating different techniques to handle with these restrictions. [2]

Since the last report, we extracted two different Algorithms from the LHCb software framework. The current focus is on the HybridSeeding Tracking Algorithm which reconstructs particle tracks in the High Level Trigger. We have adapted this algorithm on small modern low power devices and we have run several efficiency tests in comparison to state of the art high performance server systems. We achieved good results in energy efficiency and execution time compared to our reference system.

In addition, we went further on in our research on low power GPU utilisation in Trigger Algorithms.

1 Introduction

The LHCb project is a large and complex research project grown over the last decades. Named after the b-quark, LHCb is one of the four big experiments at CERN. The general scope is to explain the matter/anti-matter asymmetry. The main focus is the study of

particle decays involving beauty and charm quarks. [1] At the time of writing this report, we working on this project for two years. Our specific research topic is to improve the High Level Trigger (HLT) decision time.

The HLT has to process all the experiment data in hard time constraints and sample it down to the maximum load which the storage can handle. (Figure 1) [2] [3]

Figure 1: Trigger system setup

The challenge is to find new solutions for processing this big amounts of data with limited resources much faster than it has been performed in the first run of the LHCb project and allow the physicists to make experiments with more precise decisions. Currently our focus is one the HybridSeeding Algorithm, which reconstructs particle tracks in the High Level Trigger.

Our approach is to use modern low power hardware devices, because we expect this hardware to have more powerful compute capabilities at the same level of energy consumption. We reimplemented some algorithms of the LHCb Trigger successfully on a low power cluster and could show that using modern low power hardware improves the energy efficiency drastically while the event processing time is not increased significantly.

2 Evaluation of Event Processing on Low-Power Compute Units

At the time of writing this report, the High Level Triggers HLT1 and HLT2 concept used by the LHCb Project is a large computing grid of state of the art Intel-based server machines. The LHCb Hardware Trigger is placed before the High Level Software Trigger and prevents it from being overran by too many events, but it is not be used anymore, since the goal is to process all events by the Software Triggers. [2]

The student project group PG603 planned and constructed a 160-core cluster of low-power ARM Cortex-A53 compute devices, which has given us the opportunity to test our approach under more realistic conditions than with a single ARM Board. (Figure 2)

The tested HybridSeeding Algorithm has been extracted from the LHCb Software Framework and all dependencies to the underlying Gaudi- and Root-Framework have been removed due to the fact that these Frameworks offer a lot of functionality in analysis and

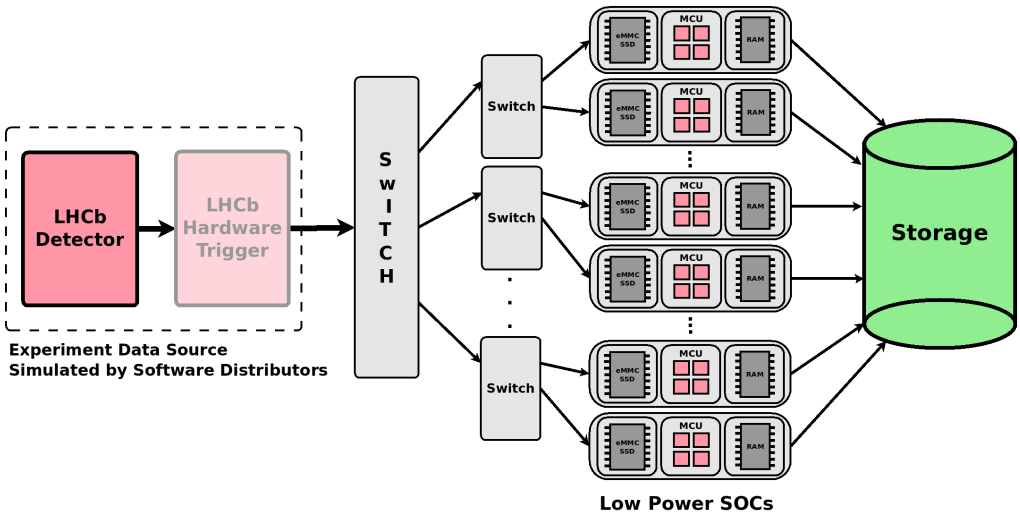


Figure 2: New Tested Event Processing Model

storage but also need a big amount of storage and system resources, what is not suitable for low power compute devices. Thus, all used analysis functions from the Gaudi Framework have been replaced. The storage functionality from the Root framework has been eliminated by using Google's Protocol Buffer Format instead of Root-files. [4]

We ran several experiments on our modified HybridSeeding Algorithm and visualized some of the results in (Figure 3). This extracted algorithm includes the whole processing path from receiving an event in form of detector hits as protobuf binary, the calculation to particle tracks with the HybridSeeding Algorithm and the writing of the output.

For this evaluation, we ran the the algorithm in two different configurations, the first is a local variant which loads events from a protobuf file on the local drive, in Figure 3 referred to as *local*. Due to the fact that the ARM Cluster has a distributed file system, the input data for the 48-thread dual socket Xeon E5-2695 reference server has been stored in a RAM drive. Otherwise the file system access could be a bottleneck and distort the measurement. The other implemented variant is even more a realistic simulation of the the LHCb Detector as we implemented Software Distributors, which job is to be the seeding source of detector hits to the processing system in the same way the real detector hardware does, in Figure 3 referred to as *8d-all* since it is a test with eight distributors and all worker nodes in the cluster active.

There are various experiment parameters we are interested in, on the one hand the execution time and on the other hand the power consumption during the event processing.

The experiments have shown that we can process events 1.34 times faster on the reference Xeon Dual Socket Server but we measured a factor of 2.28 in energy efficiency, which is scalable for any number of maschines. We still have to keep in mind that there

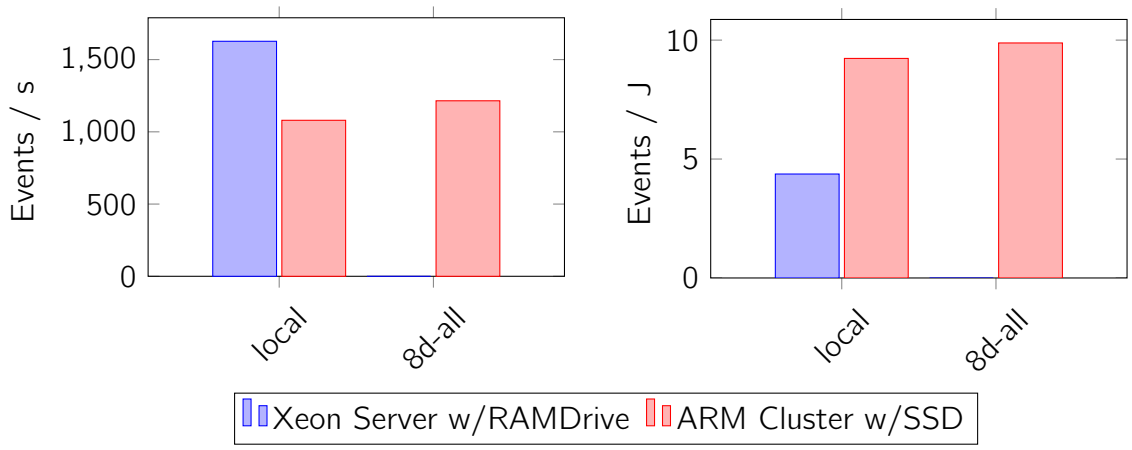


Figure 3: HybridSeeding Execution of 56000 events on ARM Cluster and Reference Xeon E5-2695 Dual Socket Server

is an extra need of network devices and we have to consider that the xeon servers energy consumption was measures on the mains side.

3 Conclusion and Future Work

Since the last report we did a lot of experiments on our new concepts and got a proof of concept of the approaches made.

Furthermore, we are working on optimizing the HybridSeeding Tracking Algorithm for the utilization of low-power ARM Mali GPUs.

In addition, our research went further on to intelligent storage solutions with filter capabilities on stored data, which are suitable for both, analysis jobs on stored experiment data on the one hand but also for trigger applications on the other hand.

References

[1] C. Langenbruch et. al., The LHCb collaboration. Angular analysis of the $B^0 \to K^{*0}\mu^+\mu^-$ decay. *LHCb-CONF-2015-002*, 2015.

[2] The LHCb collaboration. Technical Report LHCb Tracker Upgrade Technical Design Report. *CERN-LHCC-2014-001. LHCB-TDR-015*, Feb 2014.

[3] The LHCb collaboration. Technical Report CERN-LHCC-2014-016. LHCB-TDR-016. *CERN-LHCC-2014-001. LHCB-TDR-015*, May 2014.

[4] Google Inc. Protocol buffers. `https://github.com/google/protobuf`.

Measurement of *CP* violation in $B^0 \to J/\psi(e^+e^-)K^0_S$ and $B^0 \to \psi(2S)K^0_S$ decays with the LHCb experiment

Ramon Niet
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
ramon.niet@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP* violation in decays of beauty and charm hadrons. In order to maximize the sensitivity with respect to these specialized targets the LHCb detector is built as a forward single arm spectrometer, whereas the other three experiments ATLAS, ALICE and CMS are so called General Purpose Detectors, covering a symmetrical region around the proton-proton interaction point.

In the collisions of protons inside the vertex locator (VELO), new particles are created and decay until they finally leave traces in the various subcomponents of the detector. These traces are hits in the tracking systems (VELO, TT, T1-T3, M1-M6), clusters in the calorimeters (ECAL, HCAL) and Cherenkov radiation in the Ring Imaging Cherenkov Detectors (RICH1, RICH2). To conclude on the presence of particles the information of these subdetectors needs to be reconstructed, e.g. by fits of trajectories to ensembles of hits and pattern recognition algorithms looking for clusters of energy deposition. Finally, particle candidates need to be combined to heavier particles in order to perform physics measurements on the same. The endeavour to find the particles of interest is hindered for two main reasons. Firstly, hundreds of particles are produced inside the angular acceptance which allows for a large number of combinations to be made in the reconstruction as well as the combination of particles. Secondly, the interaction rate of 50ns / 25ns together with the limitation on the bandwidth that can be written to disk enforces a fast reconstruction that leads to the selection of interesting events and the

rejection of physically uninteresting ones. Both these points together set the frame for investigating these tasks in the context of resource limitation: The reconstruction and combination tasks can be parallelized and therefore performed faster.

Decays of neutral B mesons involving $b \to c\overline{c}s$ transitions are referred to as "golden modes" for measuring CP violation in the interference of direct decay and decay after B–$\overline{B}$ mixing. These decay channels are theoretically clean, as higher-order contributions that could introduce additional strong and weak phases in the decay amplitudes are expected to be small [1–3].

In the B^0 system, the decay modes $B^0 \to c\overline{c}K^0_S$ and $B^0 \to c\overline{c}K^0_L$ belong to this class of decays, where $c\overline{c}$ denotes a charmonium resonance like $J/\psi\,, \psi(2S), \eta_c$, *etc.* Here, '$K^0_S$' and '$K^0_L$' do not denote the undecayed K^0 mass eigenstates, but rather their $\pi\pi$ (CP even) and $\pi\pi\pi$ (CP odd) final states, respectively. At LHCb, only the K^0_S with the $\pi^+\pi^-$ final state is considered, as it does not contain neutrals. As the decay width difference $\Delta\Gamma$ is very small, $|\Delta\Gamma_d/\Gamma| = (0.1 \pm 1.0) \cdot 10^{-2}$ [4], and CP violation in the mixing is negligible, $|q/p| = 1$, the time-dependent decay rate asymmetry can be written as

$$\begin{aligned}\mathcal{A}_{J/\psi K^0_S}(t) &\equiv \frac{\Gamma(\overline{B}^0(t) \to c\overline{c}K^0_S) - \Gamma(B^0(t) \to c\overline{c}K^0_S)}{\Gamma(\overline{B}^0(t) \to c\overline{c}K^0_S) + \Gamma(B^0(t) \to c\overline{c}K^0_S)} \\ &= S\sin(\Delta m\, t) - C\cos(\Delta m\, t).\end{aligned} \tag{1}$$

The states $\overline{B}^0(t)$ and $B^0(t)$ represent evolving B^0 mesons decaying at decay time t after being produced as $\overline{B}^0$ and B^0 at $t = 0$, respectively. The parameter Δm represents the mass difference between the two B^0 mass eigenstates. While the sine term accounts for CP violation in the interference of decay and mixing, a non-vanishing cosine term results from CP violation in the decay. As higher order penguins are suppressed, direct CP violation is expected to be negligible at the current level of experimental precision, hence $C \approx 0$ and $S \approx \sin 2\beta$, where β is one of the angles of the unitarity triangle of the CKM-matrix.

The present work is on the LHCb measurements of S and C using $B^0 \to J/\psi\, K^0_S$ candidates reconstructed in the di-electron final state of the J/ψ and the $B^0 \to \psi(2S)K^0_S$ mode with the charmonium state $\psi(2S)$ reconstructed in the di-muon final state. The former will be referred to as the J/ψ mode, the latter as the $\psi(2S)$ mode. In both decays, only the charged $\pi^+\pi^-$ final state of the K^0_S meson is considered. The analysed data sample corresponds to the same 3 fb^{-1} of pp collisions used in the analysis of $B^0 \to J/\psi(\mu^+\mu^-)K^0_S$ decays. While the analysis of the $\psi(2S)$ mode will allow to increase the precision of the $\sin 2\beta$ measurements by LHCb while being very similar to the $B^0 \to J/\psi(\mu^+\mu^-)K^0_S$ mode, the J/ψ mode with the di-electron final state additionally represents a benchmark measurement for flavour tagged, decay time dependent CP analyses in di-electron modes at LHCb.

Fits to the reconstructed mass distributions are used to determine signal weights to statistically unfold the signal B^0 components (see Fig. 1), resulting in 10630 ± 140

$B^0 \to J/\psi(e^+e^-)K^0_S$ and 7970 ± 100 $B^0 \to \psi(2S)K^0_S$ observed decays. Probability density

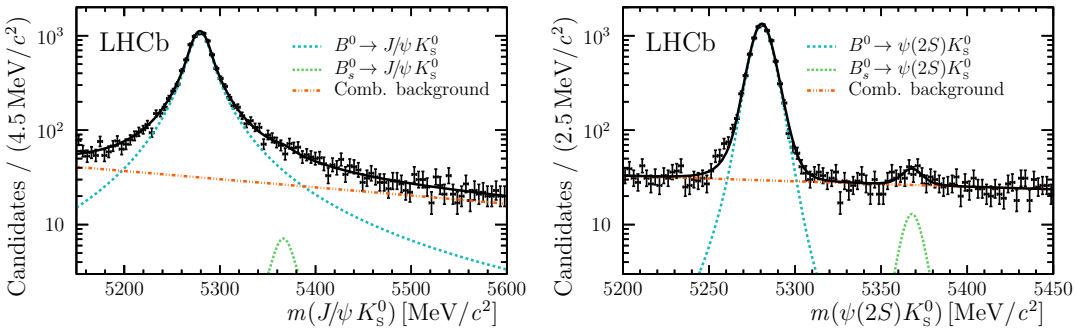


Figure 1: Nominal massfit to extract sWeights for the *CP* fit to the decay time to $B^0 \to J/\psi(e^+e^-)K^0_S$ (left) and $B^0 \to \psi(2S)K^0_S$ (right) candidates.

functions to describe the decay time and tagging (flavour identification) information are developed and validated for the extraction of the *CP* parameters. The parameters are finally extracted through a maximum likelihood fit to the signal-weighted candidates. The *CP* fit is performed simultaneously to both decay modes, allowing for different amounts of *CP* violation. The obtained decay-time-dependent *CP* asymmetries are given in Fig. 2, resulting in measurements of the *CP* parameters as

$$\begin{aligned} C\left(B^0 \to J/\psi K^0_S\right) &= 0.12 \pm 0.07 \pm 0.02\,, \\ S\left(B^0 \to J/\psi K^0_S\right) &= 0.83 \pm 0.08 \pm 0.01\,, \\ C\left(B^0 \to \psi(2S) K^0_S\right) &= -0.05 \pm 0.10 \pm 0.01\,, \\ S\left(B^0 \to \psi(2S) K^0_S\right) &= 0.84 \pm 0.10 \pm 0.01\,. \end{aligned}$$

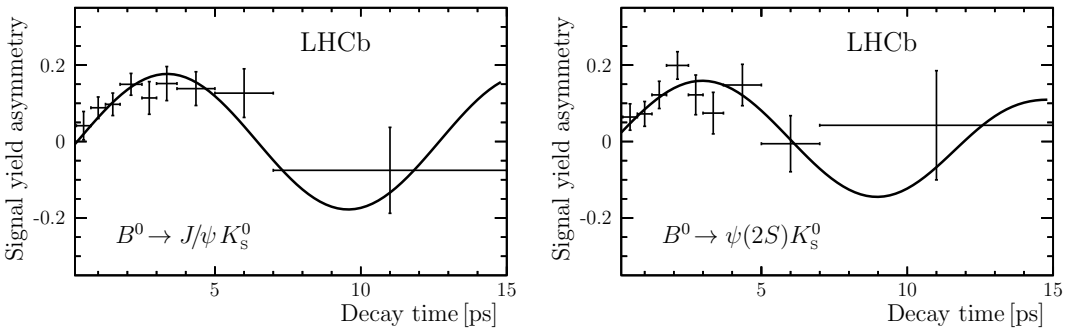


Figure 2: Decay-time-dependent *CP* asymmetries for the $B^0 \to J/\psi(e^+e^-)K^0_S$ (left) and $B^0 \to \psi(2S)K^0_S$ (right) modes.

Combinations with the previous measurements of LHCb in $B^0 \to J/\psi(\mu^+\mu^-)K^0_S$ [5] have been performed through likelihood scans (see Fig. 3), resulting in

$$
\begin{aligned}
C(B^0 \to J/\psi\, K^0_S) &= -0.014 \pm 0.030\,, \\
S(B^0 \to J/\psi\, K^0_S) &= \quad 0.75 \quad \pm 0.04\,, \\
C(B^0 \to [c\overline{c}] K^0_S) &= -0.017 \pm 0.029\,, \\
S(B^0 \to [c\overline{c}] K^0_S) &= \quad 0.760 \pm 0.034\,,
\end{aligned}
$$

for modes involving the J/ψ in the final states and for all charmonium modes, respectively. The results have been published in the Journal of High Energy Physics (JHEP) [6].

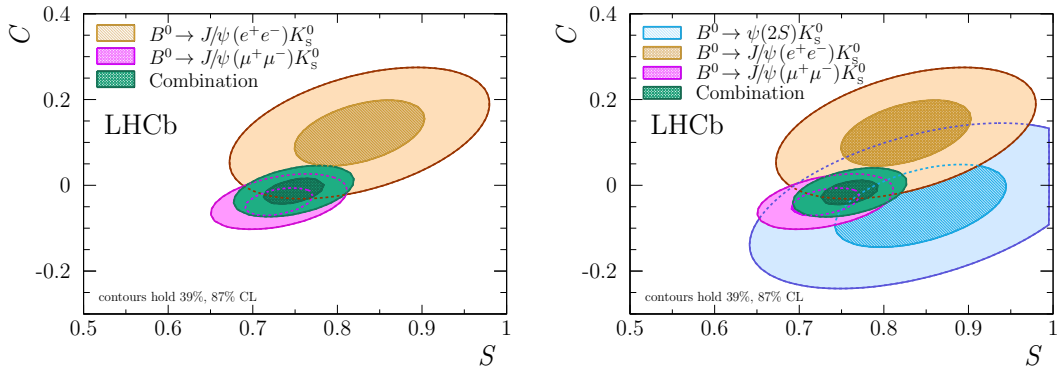


Figure 3: Likelihood contours for combinations $\sin 2\beta$ measurements in modes including the J/ψ meson in the final state (left) and all charmonium modes at LHCb.

References

[1] S. Faller, R. Fleischer, M. Jung, and T. Mannel, "The golden modes $B^0 \to J/\psi K^0_{S,L}$ in the era of precision flavour physics," *arXiv.org*, vol. D79, p. 014030. 4 p, Sept. 2008.

[2] M. Jung, "Determining weak phases from $B \to J/\psi\, P$ decays," *Phys.Rev.*, vol. D86, p. 053008, 2012.

[3] R. Fleischer, "Penguin effects in $\phi_{d,s}$ determinations," 2012.

[4] Y. Amhis *et al.*, "Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2014," 2014. updated results and plots available at: `http://www.slac.stanford.edu/xorg/hfag/`.

[5] R. Aaij *et al.*, "Measurement of CP violation in $B^0 \to J/\psi\, K^0_S$ decays," *Phys. Rev. Lett.*, vol. 115, p. 031601, 2015.

[6] R. Aaij *et al.*, "Measurement of CP violation in $B^0 \to J/\psi K^0_S$ and $B^0 \to \psi(2S)K^0_S$ decays," *JHEP*, vol. 11, p. 170, 2017.

Improved precision measurement of *CP* violation

Margarete Schellenberg
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
margarete.schellenberg@tu-dortmund.de

One of the four big experiments at the Large Hadron Collider (LHC) near Geneva is the LHCb experiment [1]. Its main focus lies on the research of the asymmetry of matter and anti-matter in the observable universe. During the Big Bang matter and anti matter should have been produced in equal parts. Today we observe a large asymmetry between the two, so that it is assumable, that physical laws influence matter and antimatter in different ways. Physicist at the LHCb experiment are investigating the charge-parity (*CP*) violation in decays of beauty and charm hadrons as one possible cause for this asymmetry. Due to the focus on hadrons containing *b* and *c* quarks, the LHCb detector is designed as a single-arm forward spectrometer (see Figure 1). At the interaction point

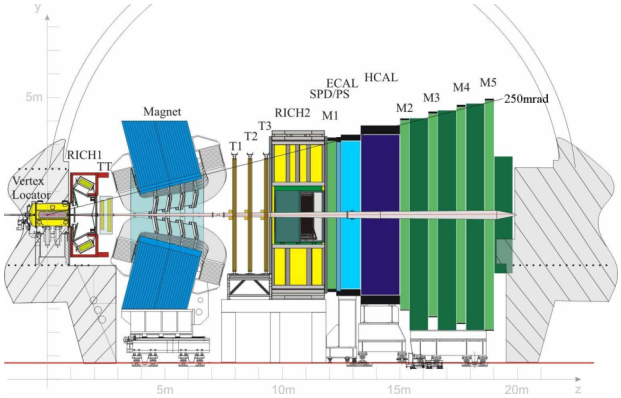


Figure 1: Scheme of the LHCb detector, illustrating the various subdetectors for identification and reconstruction of particles and their tracks. [2]

of the proton beams, which lies inside the vertex locator (VELO), a large amount of particles is produced by many different physical processes. These particles decay into new particles, which fly through the detector and interact with the detector material. Different subdetectors are responsible for the reconstruction and identification of these particles. The flight path reconstruction is performed by the tracking system consisting of the VELO, TT, IT and OT. The particle identification system is composed of the two RICH detectors, the two calorimeters ECAL and HCAL, as well as the Muon chambers. Information from both systems is used to reconstruct the complete decay chain, by combining the tracks and tracing them back to the heavier mother particles.

At the LHC, collisions are produced at a rate of 40 million collisions per second. Considering down times of the collider, the experiment has to handle nearly 40×10^{14} collisions per year. It is not possible to store every collision, as each one of them creates about 100 kB of data. Thus, an online trigger system is utilised that only a few per mille of the events pass, resulting in an amount of data that is savable on a large storage cluster. After a centralised loose preselection of the complete dataset, the data is used by physicists, who store the data that corresponds to their analysis conditions in form of ROOT [4] nTuple structures. These nTuples are much smaller than the full recorded dataset, but can still reach a size of several hundreds of Gigabytes.

To show the necessity of handling the data in an efficient way, a measurement of time-dependent *CP* violation in the decay $B^0 \to D^{*\pm}D^{\mp}$ is being performed. The reconstruction is done with $D^- \to K^+\pi^-\pi^-$ and $D^{*+} \to D^0\pi^+$, where the D^0 decays into $K^-\pi^+$.

Due to the fact that freely propagating B^0 mesons can mix into their anti particle state ($\overline{B}^0$) and vice versa and because the charge-conjugated final states $D^{*+}D^-$ and $D^{*-}D^+$ are reachable for B^0 and $\overline{B}^0$ mesons, a decay-time-dependent *CP* asymmetry can be measured. It results from the interference between the amplitudes of the direct decay and decay after B^0-$\overline{B}^0$ mixing:

$$A_f(t) = \frac{\Gamma(\overline{B}^0(t) \to f) - \Gamma(B^0(t) \to f)}{\Gamma(\overline{B}^0(t) \to f) + \Gamma(B^0(t) \to f)} = \frac{S_f \sin(\Delta m t) - C_f \cos(\Delta m t)}{\cosh\left(\frac{\Delta\Gamma t}{2}\right) + D_f \sinh\left(\frac{\Delta\Gamma t}{2}\right)}. \qquad (1)$$

The decay-time-dependent asymmetry is given by the difference between the time-dependent-decay widths of B^0 and $\overline{B}^0$ mesons decaying into the final state f, normalised to the sum. An analogous asymmetry exists for the final state $\overline{f}$. $B^0(t)$ and $\overline{B}^0(t)$ denote the initial B flavour and the parameters Δm and $\Delta\Gamma$ are the differences of the masses and decay widths between the heavy and light mass eigenstates concerning the B^0-$\overline{B}^0$ system [5]. With some assumptions, the time-dependent asymmetries become

$$A_f(t) = S_f \sin(\Delta m t) - C_f \cos(\Delta m t), \qquad A_{\overline{f}}(t) = S_{\overline{f}} \sin(\Delta m t) - C_{\overline{f}} \cos(\Delta m t), \quad (2)$$

where S_f, $S_{\overline{f}}$, C_f and $C_{\overline{f}}$ are the *CP* observables.

The first analysis step is the separation of signal and background decays. Background decays comprise particles, that are mistaken with the signal due to wrong reconstruction or misidentification. In order to extract information from such large data samples, it is necessary to achieve a good control over the different backgrounds. The selection that is applied consists of different steps. It starts with rectangular cuts on kinematical and geometrical requirements. Afterwards, a multivariate analysis is performed. At the end of the selection the data sample still includes background contributions. Thus, a statistical background subtraction is achieved by fitting the invariant B^0 mass. In Figure 2 the probability density function used to parametrise the mass distribution is shown.

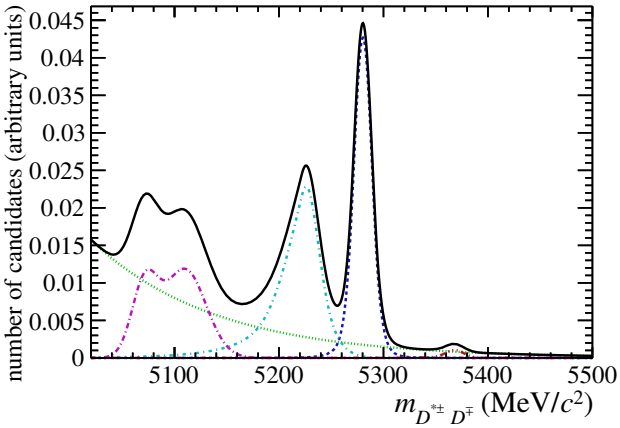


Figure 2: Parametrisation of the invariant B^0 mass distribution. The highest peak (blue) describes the signal. Left to the signal in magenta and cyan exclusive background can be seen. The green line corresponds to combinatorial background. To the right of the $B^0 \to D^{*\pm}D^{\mp}$ signal the background decay $B_s^0 \to D^{*\pm}D^{\mp}$ is present (red). The solid black curve is the sum of all components.

The next analysis step is a maximum-likelihood fit of the B^0 meson decay-time distribution that measures the *CP* parameters. Figure 3 shows a representative fit on simulated data. Results of the data fit cannot be shown, because the analysis is not finished and the central values of the *CP* parameters are still blinded. However, preliminary statistical uncertainties were already obtained. For the parameter S this uncertainty is about twice as large as in the respective analyses of the Belle and BaBar experiments [3, 6]. Thus, to become competitive with the B-factories, the statistics needs to be increased.
Since the amount of measured data is limited, the only way to increase the statistics is to improve the selection. One possibility to optimise the selection in the analysis of $B^0 \to D^{*\pm}D^{\mp}$ is to use more elaborated cut points. Therefore, a scan of the cut points on a variable needs to be performed by determining the cut efficiency at every step, *i.e.* considering the background rejection and signal efficiency. The optimal cut point is the one with the best preservation of the signal while rejecting as much background as possible. Executing cuts on an nTuple that is several hundreds of GB large is very time consuming. To increase the performance, a programming model like MapReduce for

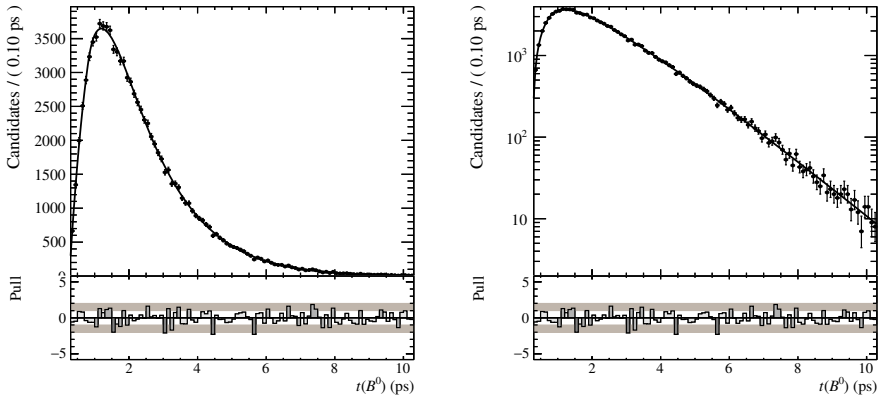


Figure 3: Decay time distribution of the decay $B^0 \to D^{*\pm}D^{\mp}$ of simulated data.

parallelising the cut process is considered, which results in a massively reduced processing time, especially compared to nonparallelised processes on a single multi-core machine. The Apache Hadoop framework [8] fulfils the conditions with the Hadoop Distributed File System (HDFS) [7] that allows for scalable distributed storing abilities, while the MapReduce programming model can be used for the processing task.

References

[1] A. A. Alves, Jr. et al. The LHCb detector at the LHC. *JINST*, 3:S08005, 2008.

[2] R Aaij et al. Letter of Intent for the LHCb Upgrade. Technical Report CERN-LHCC-2011-001. LHCC-I-018, CERN, Geneva, Mar 2011.

[3] Bernard Aubert et al. Measurements of time-dependent *CP* asymmetries in $B^0 \to D^{(*)+}D^{(*)-}$ decays. *Phys. Rev.*, D79:032002, 2009.

[4] R. Brun and F. Rademakers. ROOT: An object oriented data analysis framework. *Nucl. Instrum. Meth.*, A389:81, 1997.

[5] K. A. Olive et al. Review of particle physics. *Chin. Phys.*, C38:090001, 2014.

[6] M. Röhrken et al. Measurements of Branching Fractions and Time-dependent *CP* Violating Asymmetries in $B^0 \to D^{(*)\pm}D^{\mp}$ Decays. *Phys. Rev.*, D85:091106, 2012.

[7] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, MSST '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[8] Tom White. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2009.

Use of GPUs in the LHCb Online Farm for the track reconstruction in the SciFi tracker of the LHCb upgrade detector

Holger Stevens
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
holger.stevens@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider (LHC) near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP*-violation in decays of beauty and charm hadrons [1]. Due to some physical constraints in the production of b and c quarks through proton proton collisions the LHCb detector is designed as a single-arm forward spectrometer. Over the past years the understanding of the detector and its systematical effects has reached an almost perfect level. At the moment, the most limiting factor for analyses is the statistical uncertainty. The only way to improve this is to massively increase the dataset. For this reason, an upgrade of the experiment is foreseen in 2019 [4]. The upgrade LHCb detector is shown in figure 1. There are smaller changes to the existing detector, but the general structure of the components will remain the same. In the Vertex Locator (Velo) the position of the primary interaction is detected. The Upstream Tracker (UT) and the SciFi Tracker also belong to the tracking system. Other components like the Ring Imaging Cherenkov Detectors (RICH), the Electronic Calorimeter (ECAL), the Hadronic Calorimeter (HCAL) and the Muon Chambers (M2-M5) are used for the particle identification.
Induced by an already implemented upgrade of the LHC the rate of proton-proton collision is doubled from 20 MHz to 40 MHz. There is no major problem for the old detector to handle this higher event rate. But due to the trigger system, which will be explained later, the output of data is not doubled. After the next upgrade of the LHC the performance of the old LHCb detector would be reduced significantly. The center-of-mass energy will be increased to 14 TeV and the luminosity to $2\times10^{33}\,\mathrm{cm}^{-2}s^{-1}$. This causes a lot more hits in the detector. Most components are designed for a single hit resolution and can not

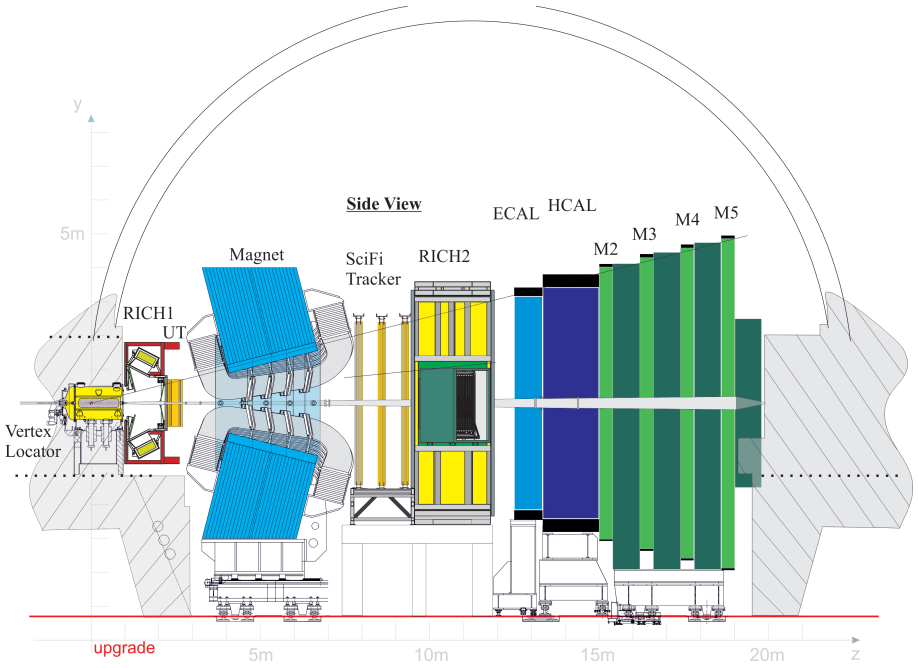


Figure 1: The LHCb upgrade detector with the various subdetectors for the identification of particles and reconstruction of their tracks [3].

distinguish multiple hits in one event. The most meaningful example are the drift tubes in the existing tracking stations. When a particle passes a tube, a signal is produced. This tube is not operative for a certain time the so called dead time, which is in this case more than 25 ns. Because there is no detector technology without dead time, a new tracking station with a higher granularity was developed, the so called SciFi Tracker. In the SciFi Tracker scintillating fibre with a diameter of 250 μm are used. This is tiny compared to the 5 cm thick drift tubes. The Tracker is composed of 3 stations. Each has 4 layer of detector material. Every layer has only a 2D resolution, the first and last, per station, in x-z direction. Because they middle ones are rotated through $\pm$ 5 degree the resolution is in y-z, these are the so called stereo layer. The hight of the layers is 5 meters, but they are split in the middle of the detector.

As mentioned before, the trigger system is a bottleneck in the LHCb data acquiring system. The trigger is needed, because the amount of data would be too high if every event would be stored. The trigger system of the present LHCb detector consists of two stages: the Level-0 (L0) hardware trigger and a software trigger, the so called High Level Trigger (HLT). The aim of the L0 is to reduce the rate to 1 MHz. The HLT reduces the rate to 12.5 kHz which can be stored. Not all of the dismissed events are uninteresting, so it would be good to be able to store more. Therefore, not only the LHCb detector will be upgraded but also the trigger system.

The new trigger is a full software trigger, as a result the whole electronics has to handle a

trigger-less readout frequency of 40 MHz. The Online Farm will receive about 32 Tbit/s. These are more or less just raw information like channel IDs from the tracking system where a hit was and factors for energy deposit in the calorimeter systems. Because of the asynchronous arrival of data from different subdetectors for one event the data needs to be buffered and then are collated. This takes place in the Eventbuilder farm and the assembled event is sent to the Eventfilter farm (EF). In the EF the raw information is decoded. This requires a detailed model of the detector. Then, the reconstruction is performed. Several algorithms are trying to build tracks from the hits in the detector. There are two types of algorithms, on the one hand the independent and on the other hand the dependent. The independent algorithms only get information from one tracking component to find tracks and the dependent ones get input tracks from other algorithms to extend them.

The aim of the project is to build an independent GPU-based algorithm to reconstruct tracks in the SciFi tracker, so called T tracks. But the framework for the LHCb-Software (GAUDI) is only supporting CPUs. For this a CoProssesor(CP)-manager was developed in the community. A task in the last year was to maintain this CP-manager to keep it compatible with the newest GAUDI versions [2].
A first version of a GPU-algorithm is executable. The basic concept is similar to the Hybrid-Seeding-Algorithm, which is the state-of-the-art CPU-Algorithm for T tracks [5]. This code conversion from CPU to GPU was just a proof of principle. No performance speedup was expected due to the special case and loop structure of the original code. But through this work the major problems for the usage of GPUs are discovered. All this insights lead to a new concept of track-finding which will be described later on. At the moment the tracking procedure is to pick a hit in the first layer and one in the very last, the connecting line has to point to the origin of the collisions. Then other x-z hits in the middle layer are added to the track candidate to build a triplet. Through the bending of the magnet in x-direction a parabolic fit is needed during this hits from the remaining x-z layers are added. After this hits in the expected regions of the stereo layers are collected. With these a 3 dimensional fit is performed. One result of some performance studies was, that it is faster to compute directly all the triplets instead of building a doublet and add a third hit from the region of interest. Normally the amount of possible combinatorics should decrease during the track-finding-process, but through the tilt of the stereo layer an uncertainty of 21 cm on the corresponding x position exists. Therefore the collection of the stereo hits and the reduction of useless combinatorics is quite slow. Most time consuming are the fits, especially the 3D one. As the fit procedure is fixed and a GPU is rather slow compared to a CPU it is important to delete as many candidates as possible before the fit. Therefor a new procedure of track-finding is developing. Thereby so called micro-tracks are used. A micro-track is a subelement of a real track but only located in one of the three stations. Then micro-tracks from different stations with a similar slope are connected. That is because the LHCb experiment is mostly interested in decays of particles with a high momentum. The slope of a track, only looking at x-z hits, is calculated for every station. In figure 2 the difference from station 1 to 2 (top) and 2

to 3 (bottom) is shown. The assumption of an equal slope is more precise for the outer station, because the magnet-field is weaker there.

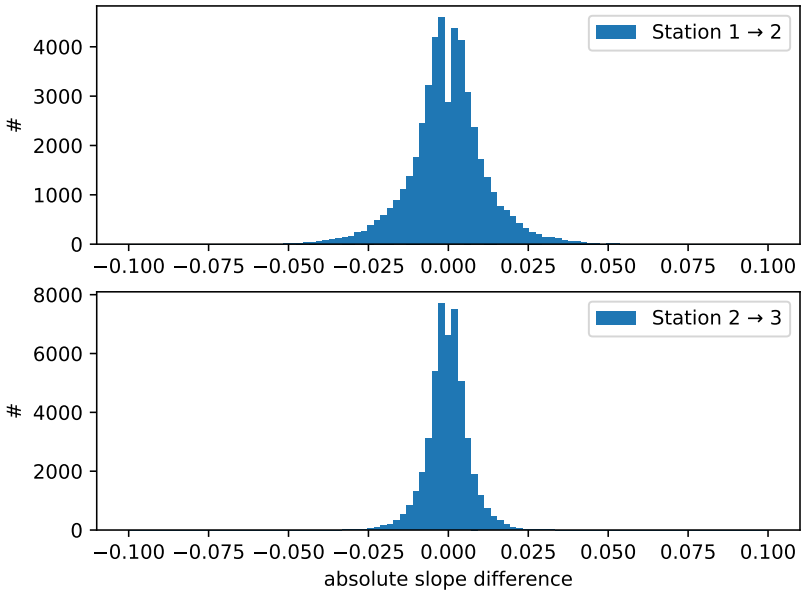


Figure 2: Slop difference for micro-tracks in x direction between different stations

References

[1] Letter of Intent for the LHCb Upgrade. Technical Report CERN-LHCC-2011-001. LHCC-I-018, CERN, Geneva, Mar 2011.

[2] A. Badalov. GPGPU opportunities at the LHCb trigger. Technical Report LHCb-PUB-2014-034. CERN-LHCb-PUB-2014-034, CERN, Geneva, May 2014.

[3] LHCb Collaboration. LHCb Tracker Upgrade Technical Design Report. Technical Report CERN-LHCC-2014-001. LHCB-TDR-015, Feb 2014.

[4] Christian Joram. LHCb Scintillating Fibre Tracker Engineering Design Review Report: Fibres, Mats and Modules. Technical Report LHCb-PUB-2015-008. CERN-LHCb-PUB-2015-008, CERN, Geneva, Mar 2015.

[5] R. Quagliani, Y. Amhis, F. Polci, and P. Billoir. Description of the hybrid seeding for the SciFi. Technical Report LHCb-INT-2015-025. CERN-LHCb-INT-2015-025, CERN, Geneva, Jun 2015.