3rd EDITION, JDSE 2018
PARIS-SACLAY

JUNIOR CONFERENCE ON DATA SCIENCE AND ENGINEERING

13th-14th September 2018

https://jdse-paris.github.io/jDSE2018/

KEYNOTES

Patrice Simard
Microsoft Research AI Lab, Redmond, United States of America

Talk: **Machine Learning -- What's next?**

Abstract:

For many Machine Learning (ML) problems, labeled data is readily available. When this is the case, algorithms and training time are the performance bottleneck. This is the ML researcher's paradise! Vision and Speech are good examples of such problems because they have a stable distribution and additional human labels can be collected each year. Problems that extract their labels from history, such as click prediction, data analytics, and forecasting are also blessed with large numbers of labels. Unfortunately, there are only a few problems for which we can rely on such an endless supply of free labels. They receive a disproportionally large amount of attention from the media.

We are interested in tackling the much larger class of ML problems where labeled data is sparse. For example, consider a dialog system for a specific app to recognize specific commands such as "lights on first floor off", "increase spacing between 2nd and 3rd paragraph", "make doctor appointment after Hawaii vacation". Anyone who has attempted building such a system has soon discovered that generalizing to new instances from a small custom set of labeled instances is far more difficult than they originally thought. Each domain has its own generalization challenges, data exploration and discovery, custom features, and decomposition structure. Creating labeled data to communicate custom knowledge is inefficient. It also leads to embarrassing errors resulting from over-training on small sets. ML algorithms and processing power are not a bottleneck when labeled data is scarce. The bottleneck is the teacher and the teaching language.

To address this problem, we change our focus from the learning algorithm to teachers. We define "Machine Teaching" as improving the human productivity given a learning algorithm. If ML is the science and engineering of extracting knowledge from data, Machine Teaching is the science and engineering of extracting knowledge from teachers. A similar shift of focus has happened in computer science. While computing is revolutionizing our lives, systems sciences (e.g., programming languages, operating systems, networking) have shifted their foci to human productivity. We expect a similar trend will shift science from Machine Learning to Machine Teaching.

The aim of this talk is to convince the audience that we are asking the right questions. We provide some answers and some spectacular results. The most exciting part, however, is the research opportunities that come with the emergence of a new field.

Bio:

Patrice Simard is a Distinguished Engineer in the Microsoft Research AI Lab in Redmond. He is passionate about finding new ways to combine engineering and science in the field of machine learning. Simard's research is currently focused on human teachers. His goal is to extend the teaching language, science, and engineering, beyond the traditional (input, label) pairs. Simard completed his PhD thesis in Computer Science at the University of Rochester in 1991. He then spent 8 years at AT&T Bell Laboratories working on neural networks. He joined Microsoft Research in 1998. In 2002, he started MSR's Document Processing and Understanding research group. In 2006, he left MSR to become the Chief Scientist and General Manager of Microsoft's Live Labs Research. In 2009, he became the Chief Scientist of Microsoft's AdCenter (the organization that monetizes Bing search). In 2012, he returned to Microsoft Research to work on his passion, Machine Learning research. Specifically, he founded the Computer-Human Interactive Learning (CHIL) group to study Machine Teaching and to make machine learning accessible to everyone.

Juliana Freire

Professor of Computer Science and Engineering and Data Science Executive Director, NYU Moore-Sloan Data Science Environment, United States of America

Talk: **Democratizing Urban Data Exploration**

Abstract:

The large volumes of urban data, along with vastly increased computing power, open up new opportunities to better understand cities. Encouraging success stories show that data can be leveraged to make operations more efficient, inform policies and planning, and improve the quality of life for residents. However, analyzing urban data often requires a staggering amount of work, from identifying relevant data sets, cleaning and integrating them, to performing exploratory analyses and creating predictive models that must take into account spatio-temporal processes. Our long-term goal is to enable domain experts to crack the code of cities by freely exploring the vast amounts of urban data. In this talk, we will present methods and systems that combine data management, analytics, and visualization to increase the level of interactivity, scalability, and usability for urban data exploration.

Bio:

Juliana Freire is a Professor of Computer Science and Engineering and Data Science at New York University. She holds an appointment at the Courant Institute for Mathematical Science, is a faculty member at the NYU Center for Urban Science and at the NYU Center of Data Science. She is the executive director of the NYU Moore-Sloan Data Science Environment, chair of the ACM SIGMOD and a council member of the Computing Community Consortium (CCC). Her recent research has focused on big-data analysis and visualization, large-scale information integration, web crawling and domain discovery, provenance management, and computational reproducibility. Prof. Freire is an active member of the database and Web research communities, with over 170 technical papers, several open-source systems, and 12 U.S. patents. She is an ACM Fellow and a recipient of an NSF CAREER, two IBM Faculty awards, and a Google Faculty Research award. She has chaired or co-chaired workshops and conferences, and participated as a program committee member in over 70 events. Her research grants are from the National Science Foundation, DARPA, Department of Energy, National Institutes of Health, Sloan Foundation, Gordon and Betty Moore Foundation, W. M. Keck Foundation, Google, Amazon, AT&T, the University of Utah, New York University, Microsoft Research, Yahoo! and IBM.

Christine Balagué

Professor at Institut Mines Telecom Business School, Titulaire de la Chaire Réseaux Sociaux, France

Talk: **Enjeux éthiques et responsabilité des technologies**

Abstract:

Les technologies d'intelligence artificielle et les usages croissants des systèmes algorithmiques impactent la vie quotidienne des individus et nos sociétés.

En 2018, la révolution digitale s'est retrouvée au cœur de nombreux débats sociétaux, le clivage devenant plus marqué entre des représentations de la technologie très positives d'une part et d'autres plus fermement négatives.

Ces débats sont liés aux enjeux éthiques qu'engendrent le développement massif des technologies et leurs usages dans nos sociétés. Les modèles dominants sont portés par les Etats-Unis et la Chine et portent des valeurs profondément différentes de celles qui ont crée l'Europe. Nous discuterons dans cet exposé les différents enjeux éthiques des technologies, depuis la recherche jusqu'aux applications, ainsi que des pistes futures permettant de développer un modèle plus responsable des technologies.

Bio:

Christine Balagué est Professeur et Titulaire de la Chaire réseaux sociaux et objets connectés à l'Institut Mines-Télécom Business School, et a été Vice-présidente du Conseil National du Numérique de 2013 à 2015. Ses recherches portent sur la modélisation du comportement des individus connectés, en particulier sur les réseaux sociaux et avec des objets connectés. Elle est également membre de la CERNA (Comité d'Ethique de la Recherche sur le Numérique d'Allistène) et de l'Institut de Convergences DATAIA sur les sciences de données et l'intelligence artificielle. En tant que VP du Conseil National du Numérique, elle a participé à différents travaux remis au gouvernement français sur les grandes questions du numérique (Neutralité du Net, Neutralité des plateformes, E-inclusion, E-éducation, E-santé, concertation nationale). Elle est également l'auteur de nombreux ouvrages sur le développement de l'Internet en France et sur les réseaux sociaux. Habilitée à Diriger des Recherches, Christine Balagué est docteur en Sciences de Gestion, diplômée de l'ESSEC et d'un Master d'économétrie à l'ENSAE.

MACHINE LEARNING

TALK SESSION

Thermodynamics of Restricted Boltzmann Machines

[Talk submission]

A. Decelle[1], G. Fissore[1,2], and C. Furtlehner[2]

[1]LRI, AO team, Bât 660 Université Paris Sud, Orsay Cedex 91405
[2]Inria Saclay - Tau team, Bât 660 Université Paris Sud, Orsay Cedex 91405

Abstract. The Restricted Boltzmann Machine (RBM), an important tool used in machine learning in particular for unsupervized learning tasks, is investigated from the perspective of its spectral properties. Starting from empirical observations, we propose a generic statistical ensemble for the weight matrix of the RBM and characterize its mean evolution. This let us show how in the linear regime, in which the RBM is found to operate at the beginning of the training, the statistical properties of the data drive the selection of the unstable modes of the weight matrix. A set of equations characterizing the non-linear regime is then derived, unveiling in some way how the selected modes interact in later stages of the learning procedure and defining a deterministic learning curve for the RBM. Finally, the analysis of a realistic RBM ensemble let us show how the model is found to operate in a compositional phase.

Keywords: Unsupervised learning, Generative models, Restricted Boltzmann Machine, Statistical physics, singular value decomposition

1 Motivation

In the last years progresses in machine learning have led to spectacular applications in many fields such as computer vision, image classification and speech recognition, giving results that were believed to be decades away [1]. While the practical applications are of tangible impact, however, our theoretical understanding of the models in use is poor. From this point of view, unsupervised learning presents specific challenges different from those of supervised learning, the former being the problem of automatically extracting structure from data while the latter refers to the ability to learn a rule that maps data to their appropriate labels. What we are interested in is the matter of automatically constructing generative models from a dataset, a problem which arises in the context of unsupervised learning.

While generative models in use can be arbitrarily complex, not even the most elementary models such as RBMs, a simple neural network with only one hidden layer, are well understood. Historically, the theoretical foundations of neural networks have been grounded in statistical physics [2][3] and in this work we show the effectiveness of this approach applied to the RBM model.

2 Empirical results

After a brief overview on the RBM model, we present a recently proposed training algorithm based on statistical physics [4]. A comparison to classical Monte Carlo based algorithms is then given, showing the substantial equivalence of the methods. This preliminary analysis paves the ground for a careful examination of the dynamics of learning, which are found to be independent on the specific training procedure used. The RBM is then studied in the linear regime, where mean-field theory from statistical physics helps in identifying the Singular Value Decomposition (SVD) of the RBM parameters as the SVD of the training data. On this basis, the analysis of the dynamical evolution of the SVD of the RBM parameters produces a detailed picture of how the structure of the training data is embedded into the model. This increases our understanding of the learning process and gives some clear insights to understand when the learning has come to an end, improving on current criteria based on flatness of a likelihood function and subjective quality of generated samples.

Finally, the SVD analysis let us differentiate the trained RBM parameters in a set of random parameters that represent noise and can be discarded and a set of structured non-random parameters. This is a first advance in an attempt to determine the proper statistical ensemble of the RBM model, in order to improve current theoretical treatments which are based on the approximation that parameters are random and independent [5].

3 Statistical Physics analysis

Starting from the empirical observations and exploiting the tools of statistical physics (mean-field theory and replica method), we propose a generic statistical ensemble for the weight matrix of the RBM and characterize its mean evolution. This let us show how in the linear regime, in which the RBM is found to operate at the beginning of the training, the statistical properties of the data drive the selection of the unstable modes of the weight matrix. A set of equations characterizing the non-linear regime is then derived, unveiling in some way how the selected modes interact in later stages of the learning procedure and defining a deterministic learning curve for the RBM.

Analyzing the thermodynamical properties of the realistic statistical ensemble of RBM that we have proposed, moreover, the model is found to operate in a ferromagnetic phase which may or may not be of compositional type, depending mainly on the distribution's kurtosis of the singular vectors components of W. Experiments on both artificial and real data show how the RBM operates in the ferromagnetic compositional phase.

4 Conclusion

The present work let us gain a much deeper understanding of how a RBM works. The point is to gain insights into the relationship between model and data. This allows us to give some elements of understanding on which properties of the data drive the learning and how they are represented in the model. Eventually this will lead us to identify and cure some flaws of present learning methods.

References

1. D. Silver et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.
2. W. Krauth and M. Mezard. Machine learning algorithms with optimal stability in neural networks. *Journal of Physics A: Mathematical and General*, 20:L745L752, 1987.
3. J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, 1982.
4. G. Marylou, E.W. Tramel, and F. Krzakala. Training restricted Boltzmann machines via the Thouless-Anderson-Palmer free energy. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 640–648, 2015.
5. R. Monasson and J. Tubiana. Emergence of compositional representations in restricted Boltzmann machines. *Phys. Rev. Let.*, 118:138301, 2017.

Wasserstein regularization for sparse multi-task regression
[Talk submission]

Hicham Janati[1,3], Marco Cuturi[2,3] and Alexandre Gramfort[1]

[1]Inria, [2]CREST, [3]Université Paris-Saclay

Abstract. Two important elements have driven recent innovation in the field of regression: sparsity-inducing regularization, to cope with high-dimensional problems; multi-task learning through joint parameter estimation, to augment the number of training samples. Both approaches complement each other in the sense that a joint estimation results in more samples, which are needed to estimate sparse models accurately, whereas sparsity promotes models that act on subsets of related variables. This idea has driven the proposal of block regularizers such as ℓ_1/ℓ_q norms, which however effective, require that active regressors strictly overlap. In this paper, we propose a more flexible convex regularizer based on unbalanced optimal transport (OT) theory. That regularizer promotes parameters that are close, according to the OT geometry, which takes into account a prior geometric knowledge on the regressor variables. We derive an efficient algorithm based on a regularized formulation of optimal transport, which iterates through applications of Sinkhorn's algorithm along with coordinate descent iterations. The performance of our model is demonstrated on synthetic simulations.

Keywords: Multi-task learning, Regression, Optimal transport

1 Motivation/Introduction

Consider multiple regression models in high dimensional settings (commonly referred to as $p \gg n$ regimes) which are known to be related. A natural assumption that combines both sparsity and joint estimation is to consider that each vector of regression coefficients is sparse, and that a common set of active features is shared across all tasks. This intuition has led to several seminal proposals of Lasso-type models, called multi-task Lasso (MTL) or multi-task feature learning (MTFL) [Argyriou et al., 2007, Obozinski and Taskar, 2006]. Both approaches are based on convex ℓ_1/ℓ_2 group-Lasso norms that promote block sparse solutions. However, in many applications the assumption of shared active features between all tasks can be too restrictive. For instance, in the context of functional brain imaging, where features are de facto brain regions, ℓ_1/ℓ_q models suggest that the exact same brain locations are active for each human subject in the study. This assumption is clearly not realistic [Gramfort et al., 2015].

In this work, we propose to handle non-overlapping supports in standard multi-task models using an *optimal transport distance* between the different parameters of

our regression models. We leverage the inherent ability of OT theory to compute a meaningful distance between probability measures with non-overlapping supports.

2 Methods

Multitask regression. Consider T regression datasets $(X^t, Y^t) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$, where n is the sample size of each set, and p is the dimension of the common space in which all observations lie. Our aim is to estimate, in a high-dimensional regime $n \ll p$, T linear regression models: $Y^t = X^t\theta^t + \epsilon^t, \quad t \in [\![T]\!]$ where $\theta^1, \dots, \theta^T \in \mathbb{R}^p$ are regression coefficients to be estimated from the samples X_t, the Y_t are the associated responses, and $\epsilon^1, \dots, \epsilon^T \in \mathbb{R}^n$ are i.i.d $\sim \mathcal{N}(0, \sigma^2)$.

Unbalanced Wasserstein distance Here we model task similarity using optimal transport. To do so, we will assume that we have a priori knowledge on the p features that form our vectors of observations in $\mathbb{R}^p$. Such a knowledge will be encoded as a $p \times p$ matrix $M \in \mathbb{R}^{p\times p}$ which describes some form of substitution cost. Consider now two probability vectors a, b in $\mathbb{R}^p_{++}$. Following [Cuturi, 2013], we use an entropy regularized definition of the Wasserstein distance. We denote by $\varepsilon > 0$ the entropy regularization strength. Moreover, to cope with inputs with different masses ($a^\top \mathbb{1}_p \neq b^\top \mathbb{1}_p$), Chizat et al. [2017] and Frogner et al. [2015] proposed independently to use a Kullback-Leibler divergence from the matrix to target marginals a and b:

$$W(a,b) \stackrel{\text{def}}{=} \min_{P\in\mathbb{R}_+^{p\times p}} \varepsilon \mathrm{KL}(P|K) + \gamma \mathrm{KL}(P\mathbb{1}|a) + \gamma \mathrm{KL}(P^\top \mathbb{1}|b) \ , \tag{1}$$

where $K = \exp(-M/\varepsilon)$. Large values of $\gamma > 0$ tend to strongly penalize unbalanced transports, and as a result penalize discrepancies between the marginals of P and a, b.

Problem statement. To induce supports proximity between coefficients, we introduce a latent variable $\bar{\theta}$ and propose to estimate the θ^t by minimizing the following regularized regression:

$$\min_{\substack{\theta^1,\dots,\theta^T \\ \bar{\theta}\in\mathbb{R}^p}} \sum_{t=1}^{T} \left[\frac{1}{2n} \|X^t\theta^t - Y^t\|^2 + \frac{\mu}{T} W(\theta^t, \bar{\theta}) + \frac{\lambda}{T} \|\theta^t\|_1 \right] \ , \tag{2}$$

where $\lambda > 0$ and $\mu > 0$ are positive regularization parameters.

When $\mu = 0$, solving (2) boils down to the estimation of T independent Lasso models, one for each task. When the θ^t are fixed, the minimization w.r.t. $\bar{\theta}$ consists in estimating the barycenter of the θ^t according to the distance W. By increasing μ, one forces all the coefficients to be closer according to W.

Optimization strategy By combining equations (1) and (2), we get a cost function that is jointly convex in $(\theta^t)_t$ and $\bar{\theta}$ (since the Kullback-Leibler is jointly convex). Strong convexity is given by the entropy terms $\mathrm{KL}(P^t, K)$. Thus, we solve it by alternating the minimization with respect to $(P^1, \dots, P^T, \bar{\theta})$ using Generalized Sinkhorn algorithm and each θ^t using proximal coordinate descent.

3 Results

We compare the performance of our algorithm against: a Lasso ran independently on each task and a *Dirty model* ([Jalali et al., 2010]) *Dirty model* solves the problem recalled in (3). Consider a decomposition of the regression coefficients into a *common* (across tasks: columns have the same support) and a *specific* part: $\theta = \theta_c + \theta_s \in \mathbb{R}^{p \times T}$. Then each part is penalized differently, this allows for a partial overlap between supports. When $\theta_s = 0$ (resp. $\theta_c = 0$) one falls back to MTFL (resp. Lasso). Each model is ran on its appropriate grid of hyperparameters. We report the best (AUC) of the Precision-Recall curve knowing the true coefficients.

$$\min_{\substack{\theta^c, \theta^s \\ \in \mathbb{R}^{p \times T}}} \sum_{t=1}^{T} \frac{1}{2n} \|X^t \theta_c^t + X^t \theta_s^t - Y^t\|^2 + \mu \|\theta_c\|_{2,1} + \lambda \|\theta_s\|_1 \ , \tag{3}$$

The boxplots of Figure 1 show the distribution of the best AUC scores. Coefficients are generated as sparse images of 576 features. The linear operator applies a gaussian blurr and a down-sampling operation by a factor of 9 (64 samples). The signal noise ratio is set to 2 and the number of tasks to 3. When the overlap is 0%, MTW reaches an AUC of 0.9 for 29% of runs whereas Lasso and Dirty get only 9% both. Only in the case of a perfect overlap, Dirty is better than alternative methods.

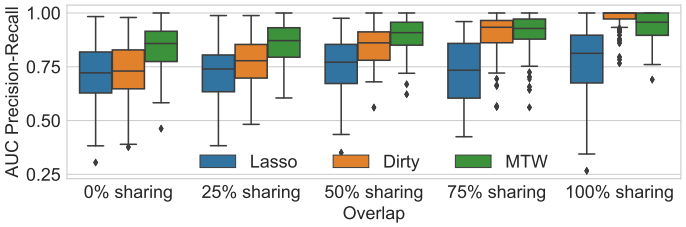


Fig. 1. Boxplot of AUC scores computed on the estimated coefficients versus ground truth with Dirty models (in particular MTFL), independent Lasso estimators, and Multi-task Wasserstein (MTW). Plots obtained with 60 independent runs on synthetic data. MTW outperforms other models when overlap is less than 50%.

4 Discussion/Conclusion

The seminal work of Caruana [1993] has motivated a series of contributions leveraging the presence of related learning tasks to improve statistical performance. Our work is one of them in the context of sparse high dimensonial regression. Using Optimal Transport to model proximity between coefficients, we proposed a convex formulation of MTL that does not require any overlap between the supports, contrarily to previous literature. We show how the MTW model can be solved efficiently relying on fast coordinate descent iterations and Sinkhorn's algorithm. Our simulations demonstrate that even when supports overlap partially, MTW outperforms Dirty models that are mixtures of ℓ_1 and block-sparse ℓ_1/ℓ_2 norms. Extensions of the model to support signed coefficients will be pursued in future work.

Bibliography

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*. 2007.

G. Obozinski and B. Taskar. Multi-task feature selection. In *ICML. Workshop of structural Knowledge Transfer for Machine Learning*, 2006.

A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *IPMI 2015*, July 2015.

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *NIPS*, 2013.

L. Chizat, G. Peyré, B. Schmitzer, and F-X. Vialard. Scaling Algorithms for Unbalanced Transport Problems. *arXiv:1607.05816 [math.OC]*, 2017.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *NIPS*, 2015.

A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A Dirty Model for Multi-task Learning. *NIPS*, 2010.

R. Caruana. Multitask Learning: A Knowledge-Based Source of Inductive Bias. *Proceedings of the Tenth International Conference on Machine Learning*, 1993.

Automated machine learning with Monte Carlo Tree Search

[Talk submission]

Herilalaina Rakotoarison and Michèle Sebag

TAU, CNRS - INRIA - LRI, University of Paris-Saclay

Abstract. The sensitivity of machine learning (ML) algorithms w.r.t. their hyper-parameters and the difficulty of finding the ML algorithm and hyper-parameter setting best suited to a given dataset has led to the rapidly developing field of automated machine learning (AutoML), at the crossroad of meta-learning and structured optimization. Several international AutoML challenges have been organized since 2015, motivating the development of the Bayesian optimization-based approach Auto-Sklearn and the Bandit-based approach Hyperband . In this paper, a new approach, called *Monte Carlo Tree Search for Algorithm Configuration* (MOSAIC), is presented, fully exploiting the tree structure of the algorithm portfolio and hyperparameter search space. Experiments (on 133 datasets of the OpenML repository) show that MOSAIC performances match that of Auto-Sklearn.

Keywords: Model selection, Hyper-parameter optimization, Monte Carlo Tree Search, AutoML

1 Motivation

The progress of the machine learning (ML) field are witnessed as an explosion of applications in all fields, from computer vision to recommendation systems. However, the diversity of ML algorithms and their sensitivity w.r.t. their hyper-parameters make it a difficult task to find the approach best suited to the application at hand. This difficulty makes all the more serious the announced shortage of machine learning experts in the next decade. The problem of automatically finding the best setting for an ML problem, referred to as AutoML, has attracted interest since the late 1980s, with several AutoML international challenges organized in the last decade. These challenges have primed the design and deployment of numerous automated machine learning platforms (AutoMLP in the following).

At the current state of the art in AutoMLPs are `Auto-Sklearn` [1] and `Hyperband`[2]. `Auto-Sklearn` relies on the Bayesian-based approach. Auto-Sklearn involves two extra components: meta-learning components to warm-start the Bayesian optimization procedure, and model ensemble strategy to build a more robust classifier. Auto-Sklearn involves 15 classifiers, 14 feature preprocessing methods, and 4 data preprocessing methods. `Hyperband` tackles hyperparameter optimization as a resource allocation task by exploring a large number of randomly sampled configurations, subject to computational constraints (cut-off time).

2 Main contribution

Monte-Carlo Tree Search [3] extends the celebrated multi-armed bandit algorithm [4] to tree-structured search spaces. Each round of Monte Carlo Tree Search consists of four steps:

- **Selection**: In each node of the tree, the child node is selected w.r.t. their value.
- **Expansion**: The algorithm adds one or more nodes to the tree.
- **Playout**: When reaching the limits of the visited tree, a roll-out strategy is used to select moves until reaching a terminal node and computing the associated reward.
- **Backpropagation**: The reward value is propagated back, i.e. it is used to update the value associated to all nodes along the visited path up to the root node.

Our main contribution is to tackle AutoML as a one-player game, adapting Monte-Carlo Tree Search (MCTS) to the exploration of the complex and structured decision space Λ made of all pre-processing, feature selection, model selection and hyper-parameter optimization, settings. Formally, the proposed approach called *Monte-Carlo Tree Search for AlgorIthm Configuration* (MOSAIC), tackles the following optimization problem:

$$\text{Find } \lambda^* = \operatorname*{argmin}_{\lambda \in \Lambda} L(\lambda, D_{train}, D_{valid}), \tag{1}$$

where Λ is the set of hyper-parameter settings, L is a loss function to assess the learning performance on the validation set D_{valid} of the model learned on the training set D_{train}. In MOSAIC, the order of the choice follows the domain knowledge: choice of pre-processing method and its parameters then the machine learning algorithm and its parameters. The MOSAIC parallels the MCTS strategy:

- The tree-path from the root node to an internal node represents a partial solution;
- In each non-terminal node, the choice of a child node is conducted using the standard UCB [4] rule in the finite case; When reaching the limits of the visited tree, a roll-out random strategy is applied until reaching a terminal node and computing the associated reward.
- The reward associated to a full tree-path is computed.
- After the evaluation of a terminal node, the reward is backpropagated to update the value in each node of the visited tree path; this value will support the choice among the child nodes in the next tree-walk.

In its current version, the search space of MOSAIC is defined from the *scikit-learn* machine learning environment [5], with 13 data preprocessing methods and 17 classifiers.

3 Results

This section presents an empirical evaluation of MOSAIC compared to the vanilla version of Auto-Sklearn which is obtained by disabling the meta-learning and ensembling component of Auto-Sklearn. We use the same settings for all experiments: memory limited to 3GB and time budget of one hour on one CPU. We evaluates 10 times both Auto-Sklearn and MOSAIC on 102 (binary and multi-class) datasets. Then compute a ranking of the two approaches according to their average test performance (Balanced

accuracy) for each dataset and finally report the average rank over all datasets of MOSAIC and Auto-Sklearn. Figure 1 reports the average rank of the two methods showing that MOSAIC slightly outperforms Vanilla Auto-Sklearn as the search goes on.

Fig. 1. Average rank of MOSAIC and Vanilla Auto-Sklearn across 102 datasets.

4 Discussion and Perspectives

The main contribution of the paper is the MOSAIC AutoML platform, adapting and extending the Monte-Carlo Tree Search setting to tackle the structured optimization problem of algorithm selection and configuration. The merits of the approach are demonstrated as it matches the performance of the mature Auto-Sklearn, which dominates the state of the art in the last 3 years. Two perspectives for further research will be: (1) improving the sampling of continuous hyper-parameter values, by taking inspiration from AlphaGo and (2) taking advantage in the search of the meta-features describing the current dataset, and to capitalize the models learned from the past datasets.

References

1. Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28*, pages 2962–2970. 2015.
2. Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. 2016.
3. Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.
4. Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.
5. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Infinite-Task Learning with Vector-Valued RKHSs
[Talk submission]

Alex Lambert[1], Romain Brault[2], Zoltan Szabo[3], Maxime Sangnier[4], Florence d'Alché-Buc[1]

1: Télécom ParisTech - 2: Centrale-Supélec - 3: École Polytechnique - 4: Sorbonne Université

Abstract Machine learning has witnessed the success of solving tasks depending on a hyperparameter. While multi-task learning is celebrated for its capacity to solve jointly a finite number of tasks, learning a continuum of tasks for various loss functions is still a challenge. A promising approach, called *Parametric Task Learning*, has paved the way in the case of piecewise-linear loss functions. We propose a generic approach, called *Infinite Task Learning*, to solve jointly a continuum of tasks via Vector-Valued Reproducing Kernel Hilbert Spaces. We provide generalization guarantees to the suggested scheme and illustrate its efficiency in cost-sensitive classification, quantile regression and density level set estimation.

Keywords: operator-valued kernels, vv-rkhs, quantiles, functional learning

1 Motivation/Introduction

Several fundamental problems in machine learning and statistics can be phrased as the minimization of a loss function described by a hyperparameter. The hyperparameter might capture numerous aspects of the problem: (i) the tolerance w. r. t. outliers as the ϵ-insensitivity in SVR, (ii) importance of smoothness or sparsity such as the weight of the l_2-norm in Tikhonov regularization, l_1-norm in LASSO, (iii) Density Level-Set Estimation (DLSE), see for example one-class support vector machines One-Class Support Vector Machine (Schölkopf et al., 2000), (iv) confidence as examplified by Quantile Regression (QR, Koenker et al., 1978), or (v) importance of different decisions as implemented by Cost-Sensitive Classification (CSC, Zadrozny et al., 2001).

For some of these problems such as QR, CSC or DLSE, one is usually interested in solving the parametrized task for several hyperparameter values. When dealing with a finite number of those hyperparameters, multi-task learning (Evgeniou et al., 2004) is then a relevant setting, enabling to take benefit from the relationship between close parameterized tasks while keeping local properties of the algorithms: ν-property in DLSE (Glazer et al., 2013) or quantile property in QR (Takeuchi, Le, et al., 2006).

Eventually, it can be advantageous to allow the hyperparameter to change, possibly among infinitely many values in order to provide a prediction tool able to deal with any value of the hyperparameter. In their seminal work, (Takeuchi, Hongo, et al., 2013) extend the multi-task learning setting by considering an infinite number of parametrized tasks in a framework called Parametric Task Learning. They prove that under a piecewise-linearity assumption on the loss function, one recovers the task-wise solution for the whole spectrum of hyperparameters, at the cost of having a piecewise-linear model.

While being able to find the task-wise solution is a desired property, the strong assumption on the loss function and the restriction to a piecewise-linear model in the hyperparameter might be a hindrance. In this paper, we define a new family of tasks, called Infinite Task Learning, in which

the piecewise linearity assumption on the loss is relaxed and whose goal is to learn a function with values in the space of continuous functions over the hyperparameter space. We propose to solve ITL in the context of vv-RKHS, shown to be adapted to multi-task learning (Micchelli et al., 2005). Due to space limitation, only the Quantile Regression problem is presented here.

2 The Infinite-Task learning framework

A *supervised parametrized task* is defined as follows. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random variable with joint distribution $\mathbf{P}_{X,Y}$; $\mathbf{P}_{X,Y}$ is assumed to be fixed but unknown. Instead we have access to n independent identically distributed observations called training samples: $S := ((x_i, y_i))_{i=1}^n \sim \mathbf{P}_{X,Y}^{\otimes n}$. Let Θ be the domain of hyperparameters, and $v_\theta \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function associated to $\theta \in \Theta$. Let $\mathcal{H} \subset \mathcal{F}(\mathcal{X}; \mathcal{Y})$ denote our hypothesis class; the goal is to find a minimizer of the expected risk

$$R^\theta(h) := \mathbf{E}_{X,Y}[v_\theta(Y, h(X))], \tag{1}$$

QR*:* Assume $\mathcal{Y} \subseteq \mathbb{R}$ and $\theta \in [0, 1]$. For a given hyperparameter θ, Quantile Regression aims at predicting the θ-quantile of the real-valued output conditional distribution $\mathbf{P}_{Y|X}$. The task can be tackled (Koenker et al., 1978) using the pinball loss defined in Eq. (2).

$$v_\theta(y, h(x)) = |\theta - \mathbb{1}_{\mathbb{R}_-}(y - h(x))||y - h(x)| \tag{2}$$

The ITL framework aims at solving jointly a continuum of parametrized tasks. To that end, the following optimization problem is considered

$$\min_{h \in \mathcal{H}} R(h) := \mathbf{E}_{X,Y}\left[\int_\Theta v_\theta(Y, h(X)(\theta))d\theta\right]. \tag{3}$$

Note that h is now a function-valued function, since at each point x we want a solution $h(x)$ to be able to predict a value at each hyperparameter. To modelize this, $\mathcal{H} \subset \mathcal{F}(\mathcal{X}; \mathcal{F}(\Theta; \mathcal{Y}))$ is chosen to be the vv-RKHS associated with the operator valued kernel $K(x, z) = k_\mathcal{X}(x, z)I_{\mathcal{H}_{k_\Theta}}$ where both $k_\mathcal{X}$ and k_Θ are classical scalar kernels, one on the input space, the other on the hyperparameter space. Since solving this problem without knowing $\mathbf{P}_{X,Y}$ is impossible, we rely on some empirical risk minimization strategy, along with a Quasi-Monte Carlo approximation with anchors $(\theta_j)_{j=1}^m$ to compute the integral. We also add a penalization based on the RKHS norm in $\mathcal{H}$, so that the problem to solve becomes

$$\operatorname*{arg\,min}_{h \in \mathcal{H}_K} \widetilde{R}_S(h) + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}_K}, \quad \lambda > 0. \tag{4}$$

where $\widetilde{R}_S(h) := \frac{1}{nm}\sum_{i,j=1}^{n,m} v_{\theta_j}(y_i, h(x_i)(\theta_j))$.

3 Guarantees for the ITL scheme

Thanks to the choice of $\mathcal{H}$, Eq. (4) becomes amenable to optimization thanks to the following finite expansion.

Proposition 1 (Representer). *Assume that for $\forall \theta \in \Theta, v_\theta$ is a proper lower semicontinuous convex function with respect to its second argument. Then Eq. (4) has a unique solution h^*, and $\exists$ $(\alpha_{ij})_{i,j=1}^{n,m} \in \mathbb{R}^{n \times m}$ such that $\forall (x, \theta) \in \mathcal{X} \times \Theta$*

$$h^*(x)(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_{ij} k_\mathcal{X}(x, x_i)k_\Theta(\theta, \theta_j).$$

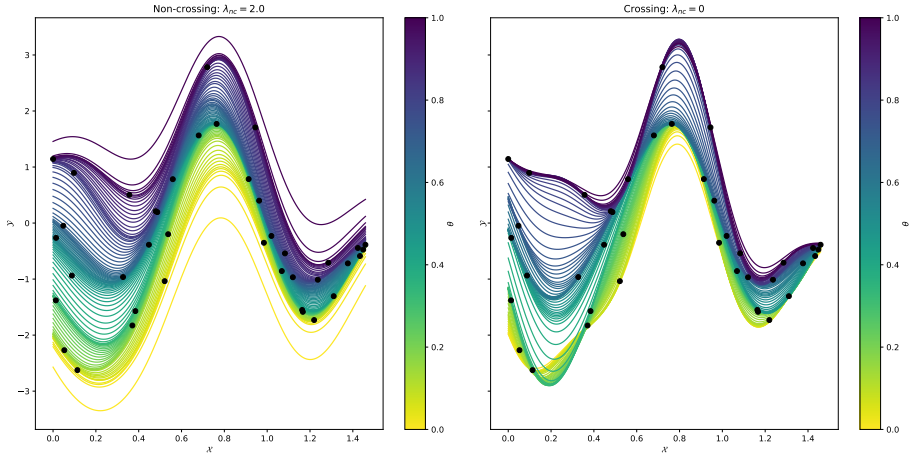


Figure 1: Impact of crossing penalty on toy data. Left plot: strong non-crossing penalty ($\lambda_{nc} = 2$). Right plot: no non-crossing penalty ($\lambda_{nc} = 0$). The plots show 100 quantiles of the continuum learned, linearly spaced between 0 (blue) and 1 (red).

In Prop. 2, we derive generalization error to the resulting estimate by stability argument (Bousquet et al., 2002), extending the work of Audiffren et al. (2013) to Infinite-Task Learning. We are especially interested in the effect of the two approximations, the one related to the size of the training sample and the other captured by m, the number of locations taken in the integral approximation. The key insight of Prop. 2 is that despite the two approximations (n, m), it is possible to get excess risk guarantees, highlighting the role of m and n.

Proposition 2 (Generalization). *Let $h^* \in \mathcal{H}$ be the solution of Eq. (4) for the* QR *or* CSC *problem with Quasi Monte Carlo approximation. Under mild conditions on the kernels k_X, k_Θ and $\mathbf{P}_{X,Y}$, one has*

$$R(h^*) \leqslant \widetilde{R}_S(h^*) + \mathcal{O}_{\mathbf{P}_{X,Y}}\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{\log(m)}{m}\right).$$

Concerning the implementation of ITL, the optimization is performed on the $(\alpha_{ij})_{i,j=1}^{n,m}$ vector of size nm with L-BFGS on a smoothed version of the pinball loss. Moreover, having a continuous model in the hyperparameter allows us to design new penalty to enforce the non-crossing phenomenon between quantiles, namely

$$\widetilde{\Omega}_{nc}(h) = \frac{\lambda_{nc}}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left| -\frac{\partial h}{\partial \theta}(x_i)(\theta_j) \right|_+ \tag{5}$$

The Fig. 1 illustrates the efficiency of this new constraint made possible by the continuum scheme.

4 Discussion/Conclusion

Infinite Task Learning with vv-RKHS is a novel nonparametric framework aiming at jointly solving parametrized tasks for a continuum of hyperparameters. This approach allows to recover several existing multi-task approaches and extends Parametric-Task Learning to nonparametric models and a larger class of loss functions.

References

Audiffren, Julien and Hachem Kadri (2013). "Stability of Multi-Task Kernel Regression Algorithms." In: *Asian Conference on Machine Learning (ACML)*. Vol. 29. PMLR, pp. 1–16.

Bousquet, Olivier and André Elisseeff (2002). "Stability and generalization." In: *Journal of Machine Learning Research* 2, pp. 499–526.

Evgeniou, Theodoros and Massimiliano Pontil (2004). "Regularized multi–task learning." In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 109–117.

Glazer, Assaf, Michael Lindenbaum, and Shaul Markovitch (2013). "q-OCSVM: A q-quantile estimator for high-dimensional distributions." In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 503–511.

Koenker, Roger and Gilbert Bassett Jr (1978). "Regression quantiles." In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Micchelli, Charles A. and Massimiliano Pontil (2005). "On Learning Vector-Valued Functions." In: *Neural Computation* 17, pp. 177–204.

Schölkopf, Bernhard et al. (2000). "New support vector algorithms." In: *Neural computation* 12.5, pp. 1207–1245.

Takeuchi, Ichiro, Tatsuya Hongo, et al. (2013). "Parametric task learning." In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1358–1366.

Takeuchi, Ichiro, Quoc V Le, et al. (2006). "Nonparametric quantile estimation." In: *Journal of Machine Learning Research* 7, pp. 1231–1264.

Zadrozny, Bianca and Charles Elkan (2001). "Learning and making decisions when costs and probabilities are both unknown." In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 204–213.

Predictive Data Mining for Multi-Variate Time Series in a Distributed Environment

[Talk submission]

Authors: Jingwei Zuo, Karine Zeitouni, Yehia Taher, and Raef Mousheimish

David Laboratory, University of Versailles UVSQ

Abstract. With the emergence of IoT concept (Internet of Things) in recent years, the multivariate data mining from IoT sensors is becoming a novel research hot spot. In particular, early classification of time series aims at predicting events as early as possible, which may help mitigating risks or anticipate actions. This work is based on a previous proposal, which combines stream data processing and early classification, to generate complex event processing rules on multivariate time series mining. Precisely, we propose novel optimizations, and leverage a distributed computing framework, namely Spark, to scale up the algorithm for Big Data context.

Keywords: Time Series Data Mining, Shapelets, Early Classification, Multivariate Time Series, Distributed Algorithm

1 Motivation/Introduction

Early classification, an emerging subject in data mining, refers to predicting events occurrences as early as possible [3]. Being applied to numerous time-dependent contexts such as IoT (Internet of Things) data flows, apart from the accuracy considered by classic classifiers, early classifier takes simultaneously the earliness into account.

Our research interest is in predictive analytic and specifically over classified time series mining. Based on a relatively new concept for time series data mining, which is called **shapelet** [7], we focus on the data flows from IoT sensors which are always multivariate, sequential, and massive. Shapelet-based mining algorithm can be potentially applied in processing IoT data flows. However, we need to highlight the fact that the current algorithm [4] on data flows mining is still relatively lethargic, specifically when the data is expanded to multi-dimensional and large scale.

In recent years, some research initiatives [1, 4] have scaled up the data dimensions in time series analytic. A typical approach of multivariate time series mining discussed in [4] presents a composed algorithm of USE **(Univariate Shapelet Extraction)** and SEE **(SEquence Extraction)**, which allows to extract class features of time series data, and generates the rules to predict the events proactively. However, the current approach based on central processing (under Python) is not suitable for large data scale as required by many real-world applications.

Therefore, in this article, we present a novel distributed approach in multivariate time series mining. Our work makes contributions to: (i) reducing the complexity of USE & SEE; (ii) scaling up the improved algorithm to apply in the context of big data.

2 Optimization of USE & SEE

To achieve early classification, our work takes advantage of a new primitive for data mining that emerged recently, which is called **Time Series Shapelets**[7]. Basically, feature extraction is the precondition of an effective and efficient classifier. Shapelet is such kind of features that are special subsequences and particularly discriminating in time series. The optimal algorithm to extract the univariate Shapelet proposed in [4] has a complexity of $\mathcal{O}(n^2m^3 \log m)$ where **m** is the length of time series and **n** is the number of time series in the dataset.

The time complexity of **USE** is impacted by the computation of the similarity of time series, such as Euclidean distance[7], Dynamic Time Warping (DTW)[2], MASS[4]. Therefore, we intend to improve the performance by choosing the most efficient method based on Nearest Neighbor Algorithms due to its easy-design feature. By deciding a distance threshold between a shapelet and the sub-sections in time series of different classes, we can then compute shapelets' **Information Gain**, which serves as the criterion of shapelets' selection. For univariate time series, USE is capable of extracting features for early classification. However, when time series scales up to multivariate data, the challenge is then boiled down to find the potential interactions/relations between different variables. **SEE** serves to extract the shapelets' sequential relations in diverse dimensions. As shown in **Figure 1**, with the input of a set of shapelets generated by USE, SEE combines these shapelets and filters the most characteristic sequential relations as output by using Information Gain or Term Frequency-Inverse Document Frequency **(TF-IDF)**. Finally, the output, namely Time-Annotated Sequences (TAS), constitutes the features extracted from multivariate time series for a early classifier, according to the accuracy and earliness criteria set by the user.

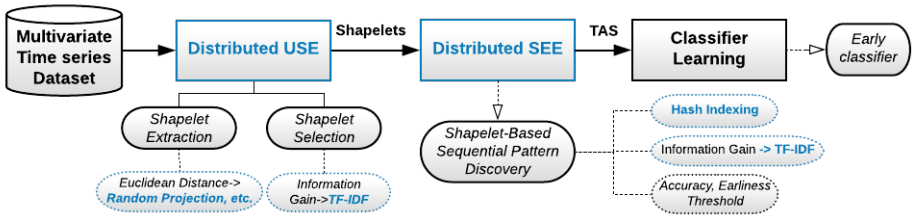


Fig. 1. Global process of Early classification for Multivariate Time Series

3 Preliminary results and future work

As there is an explicit difference between the data scale of time series mining and classic data mining, to train the classifier, the amount of data needed for time series mining are several orders of magnitude higher than that of classic data mining. Therefore, for time series mining, a distributed environment to execute the program is necessary to make the feature extraction process more efficient. To this end, as illustrated in **Figure 1**, we have proposed some modifications of the current algorithm(shown with blue color): (i) Using Hash Table to index shapelets and sequences during the generation process of class features, which allows to visibly reduce the occupied memory space on Spark cluster and save the communication costs between clusters nodes; (ii) During the sequential pattern discovery process, the pattern selection is then based on TF-IDF, rather than Information Gain, which caused a nm^2 higher time complexity and a waste of clusters resource.

Currently, by implementing only (i) on Spark, we have improved the SEE algorithm: our tests on the dataset **Wafer**, show a gain in performance for the distributed algorithm of **210% (in local mode, 8G RAM, 1 worker)**, and **560% (in cluster mode, 213.8G RAM, 6 workers)** compared to the original SEE.

To reduce the complexity of USE, which is $\mathcal{O}(n^2m^3 \log m)$ according to [4], enormous potential methods emerged over recent years could be inspired, such as Symbolic Aggregate approXimation (SAX), Random Matrix Projection[6][5], or even a classic concept in text mining TF-IDF, with a better performance than Information Gain, which allows to reduce the time complexity of Shapelet's Extraction to $\mathcal{O}(n^2m^3)$. Besides, Shapelet-based algorithm is not the only method to do the Time-Series Prediction, Markov logic networks, or Recurrent Neural Networks (RNN)[1] could also be potentially taken into consideration.

4 Discussion/Conclusion

In this work, we proposed a distributed approach to address the problem of massive multivariate data flows mining. Compared to classic classifiers, the early classification of time series data takes not only the accuracy, but also the earliness of prediction into account, which allows to predict the event as early as possible.

USE & SEE, two separable engines, extract and select shapelet features, as well as the potential interactions/relations between shapelets of different dimensions, and learn Time-Annotated Sequences (TAS) which serve as the extracted features for early classification. We optimized the SEE by TF-IDF and parallelized it in a distributed environment. In addition, we followed our theoretical hypothesis with practical experiments, tested over a real-life data set. The satisfactory results proved the efficiency of the distributed approach, and testified its competitiveness. Different optimizations are in our planning list. Specifically, to reduce the complexity of USE algorithm and to adjust it in a distributed environment.

References

1. Zhengping C, S Purushotham, and et al. Recurrent neural networks for multivariate time series with missing values. *Nature Scientific Reports*, 8(1):6085, 2018.
2. Eamonn J. Keogh and Michael J. Pazzani. Derivative dynamic time warping. In *In First SIAM International Conference on Data Mining*, 2001.
3. Yu-Feng Lin, Hsuan-Hsu Chen, Vincent S. Tseng, and Jian Pei. Reliable early classification on multivariate time series with numerical and categorical attributes. In *Advances in Knowledge Discovery and Data Mining*, pages 199–211, Cham, 2015.
4. Raef Mousheimish, Yehia Taher, and Karine Zeitouni. Automatic learning of predictive cep rules: Bridging the gap between data mining and complex event processing. DEBS '17, pages 158–169, New York, NY, USA, 2017. ACM.
5. Thanawin Rakthanmanon and Eamonn Keogh. *Fast Shapelets: A Scalable Algorithm for Discovering Time Series Shapelets*, pages 668–676.
6. Djamel-Edine Yagoubi, Reza Akbarinia, Florent Masseglia, and Dennis Shasha. Radiussketch: Massively distributed indexing of time series. In *DSAA 2017: IEEE International Conference on Data Science and Advanced Analytics*, pages 1–10, 2017.
7. Lexiang Ye and Eamonn Keogh. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22(1):149–182, Jan 2011.

STATISTICAL THEORY FOR DATA SCIENCE

TALK SESSION

Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions

Boris Muzellec and Marco Cuturi

CREST, ENSAE

Abstract. Embedding complex objects as vectors in low dimensional spaces is a longstanding problem in machine learning. We propose in this work an extension of that approach, which consists in embedding objects as elliptical probability distributions, namely distributions whose densities have elliptical level sets. We endow these measures with the 2-Wasserstein metric, with two important benefits: *(i)* For such measures, the squared 2-Wasserstein metric has a closed form, equal to the sum of the squared Euclidean distance between means and the squared Bures metric between covariance matrices. The latter is a Riemannian metric between positive semi-definite matrices, which turns out to be Euclidean on a suitable factor representation of such matrices, which is valid on the entire geodesic between these matrices. *(ii)* The 2-Wasserstein distance boils down to the usual Euclidean metric when comparing Diracs, and therefore provides the natural framework to extend point embeddings. We show that for these reasons Wasserstein elliptical embeddings are more intuitive and yield tools that are better behaved numerically than the alternative choice of Gaussian embeddings with the Kullback-Leibler divergence. In particular, and unlike previous work based on the KL geometry, we learn elliptical distributions that are not necessarily diagonal. We demonstrate the advantages of elliptical embeddings by using them for visualization, to compute embeddings of words, and to reflect entailment or hypernymy.

Keywords: optimal transport, embeddings, dimensionality reduction

1 Motivation

One of the holy grails of machine learning is to compute meaningful low-dimensional embeddings for high-dimensional complex objects. That ability is crucial to tackle advanced tasks, such as inference on texts using word embeddings, image understanding, or concise representations for nodes in a huge graph. When those embeddings live in a 2 or 3 dimensional space, they can also be used for data visualization.

Early references in this field focused on creating *isometric* embeddings in target low dimensional Euclidean spaces $\mathcal{Y} = \mathbb{R}^d$, building upon strong mathematical foundations [Johnson and Lindenstrauss, 1984]. Given input points $x_1, \dots, x_n$, the goal was to compute embeddings $\mathbf{y}_1, \dots, \mathbf{y}_n$ in $\mathbb{R}^d$ whose pairwise distances $\|\mathbf{y}_i - \mathbf{y}_j\|_2$ would not depart from the original distances $d_{\mathcal{X}}(x_i, x_j)$. Starting with metric multidimensional scaling (mMDS), several approaches have refined this intuition [Hinton and Roweis, 2003, Maaten and Hinton, 2008]. More general criteria, such as reconstruction error, co-occurence, or relational knowledge, be it in metric learning [Weinberger and Saul, 2009] or for word embeddings [Mikolov et al., 2013] can be used to obtain vector embeddings, whose distance or more generally dot-products $\langle \mathbf{y}_i, \mathbf{y}_j \rangle$ must comply with some desiderata.

Probabilistic Embeddings Our work belongs to a recent trend, pioneered by Vilnis and McCallum, who proposed to embed data points as *probability measures* in $\mathbb{R}^d$ [2015]. Usual point embeddings can be regarded as a very particular and degenerate case of probability measures, in which the mass is infinitely concentrated on a single point (a Dirac). Probability measures that are more spread-out, or event multimodal, provide therefore additional flexibility. To exploit this, Vilnis and McCallum proposed to embed words as *Gaussians* endowed with the Kullback-Leibler (KL) divergence or the usual ℓ_2 metric [Smola et al., 2007]. Athiwaratkun and Wilson have extended the latter approach to mixtures of Gaussians [2017]. The Kullback-Leibler and ℓ_2 geometries on measures have, however, an important drawback: these geometries do not coincide with the usual Euclidean metric between point embeddings when the variances of these Gaussians collapse and become Diracs. Indeed, the KL divergence between two Gaussians diverges to ∞ when their variances become small, whereas the ℓ_2 distance saturates and becomes 1 no matter where the Gaussians are located. Numerical issues arising from these degeneracies are avoided in these works by often times switching back to a simple Euclidean metric at test time..

2 Contributions

We propose in this work a comprehensive framework for probabilistic embeddings, in which point embeddings are seamlessly handled as a particular case. We consider arbitrary families of elliptical distributions, not necessarily Gaussians, and focus in particular on uniform elliptical distributions, that are more intuitive to handle because of their compact support. The cornerstone of our approach lies in the use of the 2-Wasserstein distance, which can handle degenerate measures and admits a closed form, both for the metric and its gradient [Gelbrich, 1990] in its original Riemannian formulation and more amenable Euclidean parameterization. We provide numerical tools to carry out the computation of embeddings in different scenarii, both to optimize with respect to the metric as is done in multidimensional scaling, or with respect to a dot-product, as shown in our applications to word embeddings for entailment, similarity and hypernymy tasks.

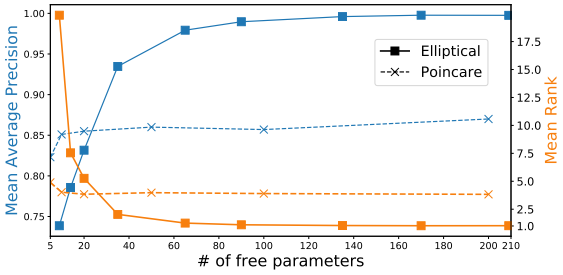


Fig. 1: Reconstruction performance of our elliptical embeddings against Poincare embeddings (values reported from [Nickel and Kiela, 2017]) on the hypernymy evaluation task, evaluated by mean retrieved rank (lower=better) and MAP (higher=better).

Preprint & Implementation This work is to appear at NIPS 2018. A preprint is avaliable on arxiv [Muzellec and Cuturi, 2018]. Python code for reproducing the experiments in this preprint is available in the following repository: `https://github.com/BorisMuzellec/EllipticalEmbeddings`.

Bibliography

Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. *arXiv preprint arXiv:1704.08424*, 2017.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the wasserstein space of elliptical distributions, 2018. To appear at NIPS 2018.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6341–6350. Curran Associates, Inc., 2017.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. *A Hilbert Space Embedding for Distributions*, pages 13–31. 2007.

Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. *Proceddings of the International Conference on Learning Representations (ICLR)*, 2015. arXiv preprint arXiv:1412.6623.

K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

Finite Sample Bounds for Superquantile Linear Prediction

Talk submission

Evgenii Chzhen, Joseph Salmon and Zaid Harchaoui

Université Paris-Est, Télécom ParisTech, University of Washington

Abstract. We present a non-asymptotic theoretical analysis of superquantile linear prediction, based on a method originally proposed by Rockafellar, Uryasev and Zabarankin. Super-quantile regression allows one to learn linear predictors with strong guarantees on the test error when the testing distribution may differ from the training distribution. Indeed classical statistical machine learning methods trained using empirical risk minimization work under the assumption that the testing distribution and the training distributions are identical. Should this assumption fail to be satisfied at test time, classical linear predictors may behave unpredictably and perform arbitrarily badly.
The notion of α-superquantile allows one to model such catastrophic risks in a precise manner. Instead of minimizing the average of the loss, we then minimize the superquantile of the loss. The associated minimization problem enjoys an intuitive interpretation owing to its Fenchel dual representation. We establish non-asymptotic bounds for kernel-based methods trained by minimizing the new objective, demonstrating the stronger robustness of the approach compared to classical counterparts when the testing distribution departs from the training distribution. We present numerical illustrations using a first-order optimization algorithm in different settings showing the interest of the approach.

Keywords: statistical learning

1 Motivation/Introduction

The empirical risk minimization and maximum likelihood principles lie at the core of many of statistical machine learning methods for prediction. In order to predict on unseen data, one usually minimizes the empirical loss on the training data, with a regularization penalty (or a constraint alternative). Machine learning methods are being deployed in several safety-critical areas, from healthcare to transportation, where the unseen data may depart from the training data in an unexpected way, and where the automated predictions are being used to feed decision-making processes: statistical machine learning methods showing robustness to distributional shifts and enjoy provable *theoretical safety guarantees* are needed.

The key to build safer machine learning methods is to put in question the empirical or expected loss minimized during training. By construction, should extreme

losses be suffered, the empirical or expected loss would hardly be affected. Instead, if one considers the empirical or expected quantiles of the loss, then extreme losses can be accounted for. The robust optimization and risk analytics communities have been developing over the years an array of concepts that one can encompass under the term of "risk measures" to model extreme losses. In particular, *superquantile*, also known in robust optimization and risk analytics as the Conditional Value at Risk (CVaR), provides an attractive scalar representation of a random variable for risk-averse decision making. Contrarily to classical quantiles, superquantiles are more stable with respect to perturbations of the underlying random variable. We analyze here alternatives to linear prediction methods, including kernel-based methods, that enjoy provable theoretical safety guarantees. The proposed methods minimize either a given empirical super-quantile of the loss or the integrated empirical super-quantiles of the loss. The theoretical guarantees provided characterize the rate of convergence with respect to the sample size of the maximum loss that may be suffered should the testing distribution departs from the training distribution within a radius of a φ-divergence between the two distributions. Compared to classical generalization bounds for classical methods such as kernel ridge regression or kernel logistic regression, the proposed bounds highlight the *distributional robustness* of the methods compared to their classical counterparts, that is a lower maximum loss if the testing data departs significantly from the training data.

2 Related work

Superquantile linear regression was proposed and explored in the series of papers Rockafellar and Uryasev [2000, 2002], Rockafellar et al. [2008]. The approach has been mostly explored for (regular) un-regularized linear regression. The notion of superquantile (SQ), also known as the Conditional Value at Risk in robust optimization and risk analytics, has been extensively studied academically, and perhaps not widely used enough, in quantitative finance. The notion also underlies chance-constrained robust optimization, such as chance-constrained linear optimization, conic optimization, etc. [Ben-Tal et al., 2009]. The superquantile is a particular instance of a coherent risk measure [Artzner et al., 1999]. Coherent risk measures can be encompassed in the Optimized Certainty Equivalent framework introduced in [Ben-Tal and Teboulle, 2007], which we use to derive our results.

Superquantiles have also been used in the same spirit as ours for reinforcement learning Chow et al. [2015]. We mention the work of [Shafieezadeh-Abadeh et al., 2015] where a related approach, yet with ambiguity set defined with Wasserstein distances, has been studied for the specific case of logistic regression. The authors also have established finite-sample bounds although with quite different techniques tailored to deal with Wasserstein distances. Therefore the two works complement each other nicely. We also mention the work of Duchi et al. [2016] who used similar tools than ours yet with a different perspective, namely providing confidence intervals for solutions of empirical risk minimization problems.

Our superquantile approach on learning bares some similarities with robust statistics Huber and Ronchetti [2009], but is radically different in essence. In robust statistics, one is mostly interested in avoiding large losses induced by outliers by removing or ignoring them. In contrast, our framework is rather inspired by robust optimization Ben-Tal et al. [2009], where one is instead attempting to achieve the best performance should the data at test time be different from the training data. Early proposals

of robust statistical machine learning methods can be found in Ben-Tal et al. [2009]. These methods were mostly developed on a case-by-base basis for each method and for each type of distributional shift, stated directly in terms of the mean or the variance of the distribution. We also mention that superquantile regression is different from quantile regression even though the names may sound similar. In quantile regression, one learns from data a linear predictor to predict a particular quantile of the response. In superquantile regression, one learns instead a linear predictor to predict well on testing data that may be drawn from a different distribution than the training data.

We present here a general approach where one minimizes the superquantile of the loss instead of the expectation of the loss usually. We establish the non-asymptotic bounds characterizing the robustness of the proposed superquantile linear predictors including kernel-based predictors. The bounds apply in particular to the original superquantile unregularized linear regression Rockafellar and Uryasev [2000]. The bounds also apply to a broader family of linear predictors, namely kernel ridge regression. We also show that the proposed linear predictors can be easily implemented using the method of simple dual averages, leveraging a dual representation of superquantiles. We present numerical illustrations that highlight the interest of the approach compared to the classical ridge regression.

Bibliography

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9(3):203–228, 1999.

A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Math. Finance*, 17(3):449–476, 2007.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2009.

Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making: A cvar optimization approach. In *NIPS*, pages 1522–1530. 2015.

J. Duchi, P. Glynn, and H. Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *ArXiv e-prints*, 2016.

P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley series in probability and statistics. Wiley, second edition, 2009.

R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *J. Bank. Finance*, 26(7):1443–1471, 2002.

R. T. Rockafellar, S. Uryasev, and M. Zabarankin. Risk tuning with generalized linear regression. *Math. Oper. Res.*, 33(3):712–729, 2008.

S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *NIPS*, pages 1576–1584. 2015.

DEEP LEARNING

TALK SESSION

Optimizing deep video representation to match brain activity

Hugo Richard[1], Ana Luisa Pinho[1], Bertrand Thirion[1], Guillaume Charpiat[2]
{hugo.richard, ana-luisa.grilo-pinho, bertrand.thirion, guillaume.charpiat}@inria.fr

[1]PARIETAL Team, INRIA, 1 Rue Honor d'Estienne d'Orves, 91120 Palaiseau, France
[2]TAU team, INRIA, LRI, Paris-Sud University, France

Abstract. The comparison of observed brain activity with the statistics generated by artificial intelligence systems is useful to probe brain functional organization under ecological conditions. Here we study fMRI activity in ten subjects watching color natural movies and compute deep representations of these movies with an architecture that relies on optical flow and image content. The association of activity in visual areas with the different layers of the deep architecture displays complexity-related contrasts across visual areas and reveals a striking foveal/peripheral dichotomy.
See full paper: `https://ccneuro.org/2018/proceedings/1071.pdf`

Keywords: deep learning, video encoding, brain mapping

1 Introduction

The understanding of brain functional architecture has long been driven by subtractive reasoning approaches, in which the activation patterns associated with different experimental conditions presented in event-related or block designs are contrasted in order to yield condition-specific maps (Poline & Brett, 2012). A more ecological way of stimulating subjects consists in presenting complex continuous stimuli that are much more similar to every-day cognitive experiences.

The analysis of the ensuing complex stimulation streams proceeds by extracting relevant features from the stimuli and correlating the occurrence of these features with brain activity recorded simultaneously with the presentation of the stimuli.

From the different layers of the deep neural networks, we build video representations that allow us to segregate (1) occipital and lateral areas of the visual cortex (reproducing the results of (Güçlü & van Gerven, 2015)) and (2) foveal and peripheric areas of the visual cortex. We also introduce an efficient spatial compression scheme for deep video features that allows us to speed up the training of our predictive algorithm. We show that our compression scheme outperforms PCA by a large margin.

2 Methods

We use a deep neural network trained for action recognition to build deep representations of more than four hours of color natural movies. We use Temporal Segment

Network (TSN) (Wang et al., 2016), a deep network pretrained on the largest action recognition dataset available (Kay et al., 2017).

The TSN network takes raw frames and optical flow fields as inputs and creates low-level to high-level abstractions of the videos using two dedicated streams. Each activity in both streams can be considered as specific features or representations of the video.

If we were to extract all network activities of the movies we would need to store more than 6 millions floats per frame in the dataset. Such a representation would be highly redundant. In order to keep the volume of data reasonable, in each stream we only focus on four convolutional layers L_1, L_2, L_3, L_4 ranked by complexity. We further compress the data using a spatial down-sampling procedure and use temporal smoothing so that we get one representation every two seconds of video, which allows us to match the acquisition rate of fMRI scanners.

10 subjects were scanned while watching the movies. In order to link extracted deep video features to the internal representation of videos in each subject we use a simple linear model to fit their brain activity in each voxel.

The use of a very simple model allows us to posit that the performance of the predictive model from a particular video representation is mostly linked to the biological suitability of the video representation.

Figure 1 gives an overview of the pipeline used to extract and process deep video features to estimate the brain activity of subjects.

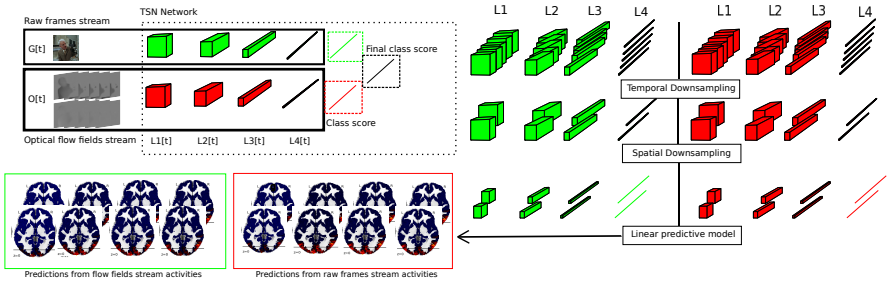


Fig. 1. Feature extraction and regression scheme: at each time frame we compute and extract the activities of four layers $L_1, \cdots, L_4$ of the temporal segment network on a single frame and on a stack of 5 consecutive optical flow fields. The extracted activities are spatially and temporally down-sampled and then used to predict brain activity of subjects exposed to the video stimuli.

3 Results

The extracted deep network features lead to different prediction performance depending on the down-sampling procedure, the stream used and the localization of target voxels.

We show that preserving the channel structure of the network during spatial compression procedure is key for developing an efficient compression scheme.

We compare three spatial compression schemes for network activities: (1) Standard principal component analysis (PCA) with 2000 components; the transformation

is learned on training sessions before it is applied to all sessions. (2) Average pooling inside channels (APIC) which computes local means of activities located in the same channel. The APIC approach strongly outperforms PCA. When using APIC we predict correctly up to 850 times more voxels than when using PCA.

Depending on the considered region of the brain, the best fitting representation varies. We show that the compressed activities of different layers show contrasts between low-level (retinotopic) versus high-level (object-responsive) areas, but also between foveal and peripheral areas.

The difference between the prediction score from high level layer activity and low level layer activity of both streams ($L_4^{flow} - L_2^{flow}$ and $L_4^{rgb} - L_2^{rgb}$) yields a clear contrast between occipital (low-level) and lateral (high-level) areas. This highlights a gradient of complexity in neural representation along the ventral stream which was also found in (Güçlü & van Gerven, 2015).

The difference between predictions score from low-level layers activity of flow fields stream and high level layers activity of raw frames stream ($L_1^{flow} - L_4^{rgb}$) yields a contrast that does not match boundaries between visual areas; instead, it does coincide with the retinotopic map displaying preferred eccentricity.

4 Discussion

Reproducing the results of (Güçlü & van Gerven, 2015) we have shown that lateral areas are best predicted by the last layers of both streams whereas occipital areas are best predicted by first layers of both streams. We have also shown that foveal areas are best predicted by last layers of the raw frames stream and that peripheric areas are best predicted by the first layers of the flow fields stream. We have introduced a compression procedure for video representation that does not alter too much the channel structure of the network, yielding tremendous gains in performance compared to PCA.

In conclusion, our study provides key insights that areas have a role linked to their retinotopic representation when performing action recognition. Future studies should focus on finessing this result by using a network tuned for other tasks.

References

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... others (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Poline, J.-B., & Brett, M. (2012, Aug). The general linear model and fmri: does love last forever? *Neuroimage*, *62*(2), 871–880.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20–36).

Generative Neural Networks for Global Optimization with Gradients

Talk Submission

Louis Faury[1⋆] and Olivier Fercoq[2]

[1] Criteo AI Lab, France
[2] LTCI, Télécom ParisTech, Université Paris-Saclay, France

Abstract. The aim of global optimization is to find the global optimum of arbitrary classes of functions, possibly highly multimodal ones. We focus on the subproblem of global optimization for differentiable functions and we propose an Evolutionary Search-inspired solution where we model point search distributions via Generative Neural Networks. This approach enables us to model diverse and complex search distributions based on which we can efficiently explore complicated objective landscapes. In our experiments we show the practical superiority of our algorithm versus classical Evolutionary Search and gradient-based solutions on a benchmark set of multimodal functions, and demonstrate how it can be used to accelerate Bayesian Optimization.

Keywords: Generative neural networks, non-convex optimization, evolutionary search.

1 Introduction

Problem formulation and related work In global optimization the task is to find the global optimum of an objective function f over a compact set $\mathcal{X}$. For general classes of functions, this cannot be performed greedily and requires the exploration of the associated landscape. Evolutionary Strategy (ES) is a state-of-the-art framework that has recently seen a growing interest in the machine learning community [6], tackling the global optimization problem when one only has access to a zeroth order oracle of the objective f. One popular ES algorithm is the Natural Evolution Strategies (NES) [9], where a search distribution p_θ is used to generate points x where f will be subsequently evaluated. From these function evaluations, NES produces a search gradient on the parameters towards lower expected objective. Formally, it minimizes:

$$J(\theta) = \mathbb{E}_{x\sim p_\theta}\left[f(x)\right] \tag{1}$$

by gradient descent, encouraging p_θ to steer its probability mass in area of smallest objective value, and relying on internal noise to explore the objective landscape. The search gradient can be estimated from samples using a Monte-Carlo estimate of the *score function* estimator [7]:

$$\frac{\partial}{\partial\theta}J(\theta) = \mathbb{E}_{x\sim p_\theta}\left[f(x)\frac{\partial}{\partial\theta}\log p_\theta(x)\right] \tag{2}$$

⋆ PhD candidate at LTCI.

NES improves over this plain gradient by using the natural gradient [1], and reports state-of-the-art performances on a large continuous optimization benchmark. However, the gradient estimator (2) requires a closed-form for p_θ, which is therefore almost always chosen to be Gaussian - θ describing the mean and variance parameters. This simplification taken for the sake of tractability can however significantly slow down the optimization process, as the search ellipsoids produced by the Gaussian distribution can provide a poor fit to the areas of small values of the objective landscape. Using more diverse and general search distributions is therefore an interesting way to improve ES. One of the most promising class of models for fitting search distributions is the class of generative neural networks [3] that have been shown to be able to model non-trivial multimodal distributions [2].

Contribution We propose to replace Gaussian search distributions by generative neural networks, which allows us to improve both convergence speed and quality of found minima by having more general and complex search distributions. To the best of our knowledge, this is the first time that neural generative models are proposed for optimization purposes. To further improve convergence speed, our method leverages gradient information, which is often available in machine learning related global optimization tasks, like hyper-parameters optimization [5], sample-variance penalization [8] or merit functions optimization in Bayesian Optimization.

2 Method

NES chooses the search distribution p_θ to be Gaussian, although any parametric distribution could be used to search the objective landscape. A general way to construct one is to apply a parametric transformation to an initial random variable u, which probability distribution we note $\mathcal{P}$. In our case, the parameter vector θ of this transformation has to be adapted or learned so that p_θ can be used to explicitly optimize f. Neural networks are able to generate complex transformations and their weights and biases can be learned quickly thanks to gradient back-propagation, and therefore they constitute good candidates for generating p_θ from $\mathcal{P}$.
We note G_θ the neural network parametrized by θ (the weights and biases of the network), mapping the noise $u \sim \mathcal{P}$ into points $x \in \mathcal{X}$. As our goal is to generate queries x with low-value of the objective, (1) is a natural cost function for training G_θ. However, note that we don't have access to a closed form of p_θ and therefore ideas similar to NES cannot be applied. Still, (1) can be rewritten as:

$$J(\theta) = \mathbb{E}_{u\sim\mathcal{P}}\left[f(x(\theta, u))\right] \tag{3}$$

where $x(\theta, u)$ is the output of G_θ with input u. This allows us to compute an estimate of J's gradient with respect to θ, known as the *pathwise derivative* estimator [7]:

$$\nabla_\theta J(\theta) \simeq \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial\theta} f(x(\theta, u_i)) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial x}{\partial\theta}(\theta, u_i)^T \nabla_x f(x(\theta, u_i)) \tag{4}$$

where $\{u_1, \ldots, u_N\}$ is a collection of samples from $\mathcal{P}$. This stochastic estimate of the gradient is then fed to a stochastic gradient descent algorithm. Note that this estimator requires access to f's derivatives, and that the Jacobian term $\frac{\partial x}{\partial\theta}$ of the neural network can be easily computed via back-propagation.

3 Results

In our full paper [4], we compare our method with two state-of-the-art evolutionary algorithms - NES and Covariance Matrix Adaptation Evolution Strategies (CMA-ES), and with a repeated greedy derivative-based algorithm used in the Bayesian Optimization (BO) community. We obtain from competitive to best results on a benchmark testbed of multimodal functions, and demonstrate empirically the good scalability of our algorithm. We also show how our method can be incorporated within the BO procedure to provide acceleration and scalability.

4 Conclusion

We propose to use neural generative models to optimize multimodal black-box functions for which gradients are available. We show the merits of our approach on a benchmark set of multimodal functions by comparing with state-of-the-art zeroth order methods and a repeated gradient-based greedy method. We also demonstratre how to use this idea in the Bayesian Optimization framework by efficiently optimizing acquisition functions. Other applications of this new method are numerous. In future work, we wish to combine recent methods [5] with our algorithms to optimize hyper-parameters of machine learning models. Another promising application of our method is the efficient global optimization of deep neural networks, which we also plan to tackle in future work.

References

1. Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 1998.
2. Yanshuai Cao, Gavin Weiguang Ding, Kry Yik-Chau Lui, and Ruitong Huang. Improving GAN training via Binarized Representation Entropy (BRE) Regularization. *arXiv preprint arXiv:1805.03644*, 2018.
3. Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via Maximum Mean Discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.
4. Louis Faury, Flavian Vasile, Clément Calauzènes, and Oliver Fercoq. Neural generative models for global optimization with gradients. *arXiv preprint arXiv:1805.08594*, 2018.
5. Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, 2015.
6. Tim Salimans, Jonathan Ho, Xi Chen, and Ilya Sutskever. Evolution Strategies as a scalable alternative to Reinforcement Learning. *arXiv preprint arXiv:1703.03864*, 2017.
7. John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, 2015.
8. Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, 2015.
9. Daan Wierstra, Tom Schaul, Jan Peters, and Juergen Schmidhuber. Natural Evolution Strategies. In *Evolutionary Computation*. IEEE, 2008.

Neuroevolution with CMA-ES for the tuning of a PID controller of nonholonomic car-like mobile robot

Mohamed Outahar and Eric Lucet

CEA, LIST, Interactive Robotics Laboratory, Gif-sur-Yvette, F-91191, France

Abstract. In the field of mobile robotics, finding an optimal control policy is a challenging task. PID controllers have been widely used in the industry. However, tuning a PID controller is not easy, especially to take into consideration the fluctuation in the precision of the perception. We propose a **neuroevolution** algorithm to find **the optimal parameters** of the controller in real time. The controller is tuned by a neural network which is trained by with the **covariance matrix adaption evolution strategy (CMA-ES)** . The neural network takes into account both the error and the uncertainty of the measurement the tuning of the parameters. The level of uncertainty in the measurement is given by the **the covariance matrix of the Kalman filter**.

Keywords: Neuroevolution, Machine learning, Neural network, Gradient-free optimization, Robotics, mobile robot, Control theory, PID controller

1 Introduction

In the last few years, neuroevolution [1] has gained interest in the research community. It has been shown to outperform reinforcement learning algorithms in certain situations where the search space is non-convex and noisy or the gradient is not available [2]. Neuroevolution describes a method to optimize neural networks with evolutionary algorithms. The algorithm is extremely parallelizable and scalable [3]. This document aims to demonstrate a method to automatically tune a PID controller of a car-like mobile robot using neuroevolution. The CMA-ES optimization algorithm was chosen to optimize the neural network. CMA-ES being also noted CMA with ES standing for Evolution Strategy which is a family of algorithms that is loosely based on biological evolution (hence the name). Multiple Evolutionary algorithms exist and they are all based on the same basic steps of population generation, evaluation, selection and reproduction. The CMA-ES has outperformed many algorithms in black box optimization problems [4]. That is why this algorithm has been chosen to tune PID controllers in many cases with promising results [5] [6].

As seen in figure 1, the system is composed of a robot controlled by a PID controller. The state of the robot is observed by an extended Kalman filter (EKF). The EKF provides the state $\hat{x}$ and the corresponding covariance matrix

P. The core concept of the document is to use the covariance matrix with the corresponding error as inputs to a neural network which outputs the parameters of the controller in real time.Both of the CMA-ES blocks are used to define and optimize the neural network in order to adapt the behavior of the robot to the level of uncertainty in the measurements.

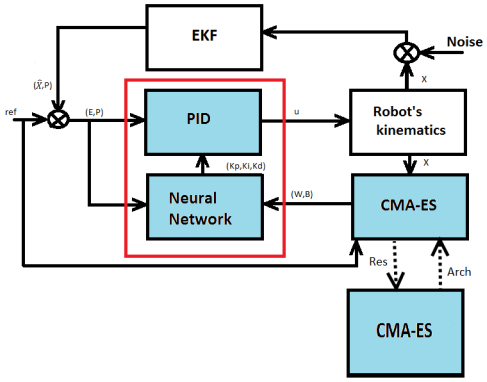


Fig. 1. Control bloc diagram.

2 PID tuning using a neural network

The goal here is to find the optimal parameters K_P, K_I and K_D to control the robot, by taking into consideration the error and the covariance matrix of the EKF. A neural network is used because if offers both adaptability and efficiency.

2.1 PID controller

PID controllers are wildly used in the industry. This is due to their reliability and simplicity. PID has shown to have good preferences in multiple cases [7]. The general formula for the PID controller is as follows:

$$C(t) = K_P e(t) + K_I \int_0^t \mathrm{e}(\tau)\,\mathrm{d}\tau + K_D \frac{d}{dt} e(t) \tag{1}$$

with e(t)=actual(t)-target(t)
K_P, K_I and K_D are the proportional, integral and derivative gains respectively. Even thou the controller is easy to implement, the tuning of its parameters is not a simple task and is a large area of research [8]. The three actions of proportional, integral and derivative have different and some concurrent effects. For example,

the proportional term decreases the rise time while the derivative term increases it, while both are essential for the stability of complex systems. This is the reason of the difficulty of finding optimal gains.

2.2 Neural network

Neural networks are highly connected systems that are used to model complex, non-linear functions.In a simple representation of a neural network, the outputs of the each layer are multiplied by the weights and summed together with the biases and passed through the activation functions. Activation functions are what makes the system capable of modeling non-linear behavior, it can be represented graphically by neurons.

A big part of the progress done in this area is due to the backpropagation algorithm. This algorithm allows the neural network to learn patterns and desired behaviors. However The backpropagation algorithm uses the gradient to optimize the neural network. In this case, the gradient is not available, therefore the backpropagation algorithm can not be used. 1

3 Neuroevolution

A neural network is used to find the optimal parameters to control the robot efficiently, even with the presence of uncertainty. In traditional neural networks, the backpropagation algorithm is used to update the weights and biases. Here, an evolutionary algorithm is used instead. The choice was made because of the need of exploration in our problem and because neuroevolution is a gradient free method, which reduces execution time by orders of magnitude [2].

3.1 CMA-ES

The CMA-ES is an evolutionary algorithm [9]. it had been used because it outperformed most black box optimization algorithms. The algorithm starts off by generating a population of candidates. Those candidates are evaluated and put in order of fitness. From the top preforming candidates, a percentage is selected to regenerate the new population. The new population is again reevaluated and the cycle continues until a termination condition is met. The termination condition is based on number of generations or the resemblance between parents and offspring.

3.2 Objective function

The CMA-ES algorithm takes an objective function as an input, and has the neural network parameters as outputs. the objective function is critical to the performance of the optimization. In our case it will be set to take into consideration the absolute error between the non noisy signals and the reference. In

other words, the CMA-ES will tweak the neural network parameters in order to minimize the influence of the noise on the system. This is done to force the neural network to learn to control the system based on the level of noise (EKF's covariance matrix).

4 Results and perspectives

Multiple implementations varying in complexity, were realized for this work. At first a fix PID controller was optimized with the CMA-ES to adapt to fluctuations in the precision of the perception. After this initial phase, a neural network was used to tune a PID controller on line. The neural network was optimized by CMA-ES. The architecture of the neural network was chosen by the user. One of the latest Implementations describes the complete system, where both CMA-ES blocks work to have an optimal system.

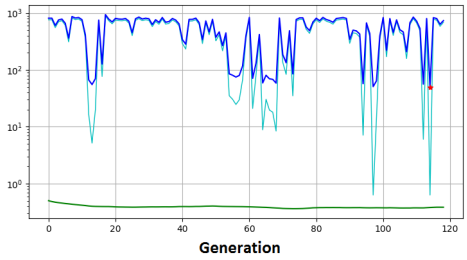


Fig. 2. Evolution of the objective function across generations. The size of the step between generations is displayed in green, the change in the objective function in cyan and the minimum objective function of each generation in blue. The red asterisk is the overall minimum objective function found by CMA-ES [10].

In figure 2 we see the evolution of the objective function throughout the generations. We notice that CMA-ES finds local optima but successfully overcomes them until the fix number of function evaluations has been achieved. As mentioned before the backpropagation algorithm can not be used. Therefore we can not compare the training phase of the two algorithms to evaluate the performance of the developed system. However we can evaluate the performance of the developed system by comparing it to other controllers which are used in these cases. These types of tests have been carried and showed promising results.

References

1. K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evolutionary Computation*, vol. 10, no. 2, pp. 99–127, 2002.

2. T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution Strategies as a Scalable Alternative to Reinforcement Learning," *ArXiv e-prints*, 2017.
3. X. Zhang, J. Clune, and K. O. Stanley, "On the relationship between the openai evolution strategy and stochastic gradient descent," *CoRR*, vol. abs/1712.06564, 2017.
4. I. Loshchilov, "Cma-es with restarts for solving cec 2013 benchmark problems," 06 2013.
5. M. s. Saad, H. Jamaluddin, and I. Mat Darus, "Pid controller tuning using evolutionary algorithms," vol. 7, pp. 139–149, 01 2012.
6. K. Marova, "Using CMA-ES for tuning coupled PID controllers within models of combustion engines," *CoRR*, vol. abs/1609.06741, 2016.
7. Y. Wakasa, S. Kanagawa, K. Tanaka, and Y. Nishimura, "PID Controller Tuning Based on the Covariance Matrix Adaptation Evolution Strategy," *IEEJ Transactions on Electronics, Information and Systems*, vol. 130, pp. 737–742, 2010.
8. B. Doicin, M. Popescu, and C. Patrascioiu, "Pid controller optimal tuning," *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, June 2016.
9. N. Hansen, "The cma evolution strategy: A tutorial," 2010.
10. N. Hansen, "Cma-es source code."

OPTIMIZATION

TALK SESSION

Stochastic algorithms for ICA

Talk submission

Pierre Ablin⋆, Alexandre Gramfort⋆, Jean-François Cardoso‡ and Francis Bach◦

⋆: INRIA, Université Paris-Saclay
‡: CNRS - Institut d'Astrophysique de Paris
◦: INRIA - Département d'informatique de l'ENS

Abstract. Independent component analysis (ICA) is a widely spread data exploration technique, where observed signals are assumed to be linear mixtures of independent components. Infomax, one of the first and most used algorithms for inference of the latent parameters, maximizes a log-likelihood function which is non-convex and decomposes as a sum over signal samples. We introduce a new majorization-minimization framework for the optimization of the loss function. We derive an online algorithm for the streaming setting, and an incremental algorithm for the finite-sum setting, which outperform the state-of-the-art on large scale problems.

Keywords: Independent component analysis, stochastic optimization, latent variable estimation

1 Motivation

In its classical and most widely spread form, ICA makes the assumption that a random vector $\boldsymbol{x} \in \mathbb{R}^p$ is a *linear mixture* of independent sources. It means that there exists a *source* vector $\boldsymbol{s} \in \mathbb{R}^p$ of statistically independent features and a *mixing matrix* $A \in \mathbb{R}^{p \times p}$, such that $\boldsymbol{x} = A\boldsymbol{s}$. The aim of ICA is to recover A from some realizations of $\boldsymbol{x}$.

One of first and most employed ICA algorithm is Infomax [1]. It assumes that each feature of $\boldsymbol{s}$ follows a given *super-Gaussian* distribution d. The likelihood of $\boldsymbol{x}$ given A writes:

$$p(\boldsymbol{x}|A) = \frac{1}{|\det(A)|} \prod_{i=1}^{p} d([A^{-1}\boldsymbol{x}]_i). \tag{1}$$

It is more convenient to work with the *unmixing matrix* $W := A^{-1}$ and the negative log-likelihood, yielding a cost function $\ell(\boldsymbol{x}, W) := -\log(p(\boldsymbol{x}|W^{-1}))$:

$$\ell(\boldsymbol{x}, W) = -\log|\det(W)| - \sum_{i=1}^{p} \log(d([W\boldsymbol{x}]_i)) \ . \tag{2}$$

The underlying *expected risk* is then:

$$\mathcal{L}(W) := \mathbb{E}_{\boldsymbol{x}}[\ell(\boldsymbol{x}, W)] = -\log|\det(W)| - \sum_{i=1}^{p} \mathbb{E}[\log(d([W\boldsymbol{x}]_i))] \ . \tag{3}$$

Given a set of n i.i.d. samples of $\boldsymbol{x}$, $X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{p \times n}$, the *empirical risk* is:

$$\mathcal{L}_n(W) := \frac{1}{n} \sum_{j=1}^{n} \ell(\boldsymbol{x}_j, W) = -\log|\det(W)| - \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n} \log(d([WX]_{ij}))] \ . \tag{4}$$

This article focuses on the inference of W in two cases: the *finite-sum* setting, where W is found by searching for a minimizer of $\mathcal{L}_n$, and the *online* setting where a stream of samples arriving one by one is considered. Infomax solves the finite-sum problem [2] by using a stochastic gradient method. Unfortunately, $\mathcal{L}_n$ is not convex, hence it is hard to find a good step-size policy which fits any kind of data. More recently, several full-batch second-order algorithms have been derived for the exact minimization of $\mathcal{L}_n$ [3]. Full-batch methods are robust and can show quadratic convergence speed, but an iteration can take a very long time when the number of samples n is large.

We introduce a set of surrogate functions for ℓ, allowing for a majorization-minimization method. We derive incremental and online algorithms for this problem. Through experiments, we observe that the proposed methods performs better than the state-of-the-art, while enjoying the robust property of guaranteed decrease.

2 Surrogate functions

The density d is assumed symmetric and *super-Gaussian* in the sense that $-\log(d(\sqrt{x}))$ is an increasing concave function over $(0, +\infty)$ [4]. Following [4], there exists a function f such that:

$$G(y) := -\log(d(y)) = \min_{u \geq 0} \frac{uy^2}{2} + f(u), \tag{5}$$

and the minimum is reached for an unique value denoted as $u^*(y)$. We introduce a new cost function $\tilde{\ell}(\boldsymbol{x}, W, \boldsymbol{u})$ where $\boldsymbol{u} \in \mathbb{R}_+^p$, which writes:

$$\tilde{\ell}(\boldsymbol{x}, W, \boldsymbol{u}) := -\log|\det(W)| + \frac{1}{2}\sum_{i=1}^{p} u_i \, [W\boldsymbol{x}]_i^2 + \sum_{i=1}^{p} f(u_i) \ , \tag{6}$$

and the associated empirical risk, for $U = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n] \in \mathbb{R}_+^{p \times n}$:

$$\tilde{\mathcal{L}}_n(W, U) := \frac{1}{n}\sum_{j=1}^{n} \tilde{\ell}(\boldsymbol{x}_j, W, \boldsymbol{u}_j) = -\log|\det(W)| + \frac{1}{2n}\sum_{i=1}^{p}\sum_{j=1}^{n} U_{ij} \, [WX]_{ij}^2 + \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n} f(U_{ij}). \tag{7}$$

Following Eq. (5), we have:

Lemma 1 (Majorization). *Let $W \in \mathbb{R}^{p \times p}$. For any $U \in \mathbb{R}_+^{p \times n}$, $\mathcal{L}_n(W) \leq \tilde{\mathcal{L}}_n(W, U)$, with equality if and only if $U = u^*(WX)$, that is, $\forall i, j, \ U_{ij} = u^*([WX]_{ij})$. Further, W minimizes $\mathcal{L}_n$ if and only if $(W, u^*(WX))$ minimizes $\tilde{\mathcal{L}}_n$.*

A natural algorithm relies on alternatively minimizing with respect to W and U. The rest of the paper focuses on the minimization of $\tilde{\mathcal{L}}_n$ rather than $\mathcal{L}_n$.

3 Stochastic minimization of the loss function

Using an EM strategy, $\tilde{\mathcal{L}}_n(W, U)$ is minimized by alternating descent moves in U and in W.

3.1 M-step: Descent in W

Expanding $[WX]_{ij}^2$, the middle term in the new loss function (6) is quadratic in the rows of W:

$$\tilde{\mathcal{L}}_n(W, U) = -\log|\det(W)| + \frac{1}{2}\sum_{i=1}^{p} W_{i:} A^i W_{i:}^\top + \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n} f(U_{ij}) \ , \tag{8}$$

where $W_{i:}$ denotes the i-th row of W, and the A^i's are $p \times p$ matrices given by:

$$A_{kl}^i := \tfrac{1}{n}\textstyle\sum_{j=1}^{n} U_{ij} X_{kj} X_{lj} \ . \tag{9}$$

Therefore, when U is fixed, with respect to W, $\tilde{\mathcal{L}}_n$ is the sum of the log det function and a quadratic term. It can be *partially* minimized in closed-form, with a *multiplicative update* of one of its rows. Let $i \in [1, p]$, and $\boldsymbol{m} \in \mathbb{R}^p$. Define $M \in \mathbb{R}^{p \times p}$ such that $M = I_p$ except its i-th row which is equal to $\boldsymbol{m}$. With respect to $\boldsymbol{m}$, $\tilde{\mathcal{L}}_n(MW, U)$ is of the form $-\log(|m_i|) + \frac{1}{2}\boldsymbol{m}K\boldsymbol{m}^\top$ where we define $K = WA^iW^\top \in \mathbb{R}^{p \times p}$. This expression can be minimized exactly by canceling the gradient, yielding:

$$\boldsymbol{m} = ((K^{-1})_{ii})^{-1/2}(K^{-1})_{i:} \ . \tag{10}$$

3.2 E-step : Descent in U

When only one sample $X_{:j} = \boldsymbol{x}_j \in \mathbb{R}^p$ is available, the operation $U_{:j} \leftarrow u^*(W\boldsymbol{x}_j)$ minimizes $\tilde{\mathcal{L}}_n(W, U)$ with respect to the j-th row of U. As seen previously, we only need to compute the A^i's to perform a descent in W, hence one needs a way to accumulate those matrices.

Incremental algorithm. To do so in an incremental way, a memory $U^{\text{mem}} \in \mathbb{R}^{p \times n}$ stores the values of U. When a sample $\boldsymbol{x}_j$ is seen by the algorithm, we compute $U^{\text{new}}_{:j} = u^*(W\boldsymbol{x}_j)$, and update the A^i's as:

$$A^i \leftarrow A^i + \frac{1}{T}(U^{\text{new}}_{ij} - U^{\text{mem}}_{ij})\boldsymbol{x}_j\boldsymbol{x}_j^\top \ . \tag{11}$$

The memory is then updated by $U^{\text{mem}}_{:j} \leftarrow U^{\text{new}}_{:j}$ enforcing $A^i = \frac{1}{n}\sum_{j=1}^n U^{\text{mem}}_{ij}\boldsymbol{x}_j\boldsymbol{x}_j^\top$ at each iteration.

Online algorithm. When each sample is only seen once, there is no memory, and a natural update rule following [5] is:

$$A^i \leftarrow (1 - \rho(n))A^i + \rho(n)U_{ij}\boldsymbol{x}_j\boldsymbol{x}_j^\top \ , \tag{12}$$

where n is the number of samples seen, and $\rho(n) \in [0, 1]$ is a well chosen factor. Setting $\rho(n) = \frac{1}{n}$ yields the unbiased formula $A^i(n) = \frac{1}{n}\sum_{j=1}^n U_{ij}\boldsymbol{x}_j\boldsymbol{x}_j^\top$. A more aggressive policy $\rho(n) = \frac{1}{n^\alpha}$ for $\alpha \in [\frac{1}{2}, 1)$ empirically leads to faster estimation of the latent parameters.

4 Discussion

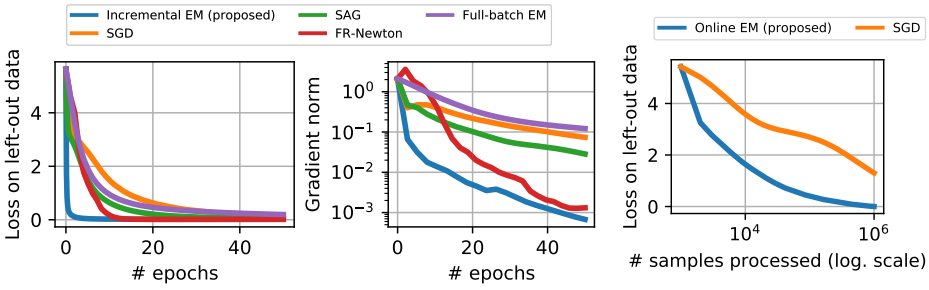


Fig. 1. Behavior of different algorithms on real data. Left and middle: finite sum problem. Right: online problem.

Fig. 1 compares several algorithms and shows some convergence measures on a real EEG dataset of size $p = 30$, $n = 10^6$. The proposed algorithms have the same cost per iteration as SGD, and clearly outperform the state-of-the-art. Furthermore, to the best of our knowledge, they are the first ICA algorithms that provably decrease the loss function at each step.

Bibliography

[1] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[2] J-F Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal processing letters*, 4(4):112–114, 1997.

[3] M. Zibulevsky. Blind source separation with relative newton method. In *Proc. ICA*, volume 2003, pages 897–902, 2003.

[4] J. Palmer, K. Kreutz-Delgado, B. D. Rao, and D. P. Wipf. Variational EM algorithms for non-gaussian latent variable models. In *Proc. NIPS*, pages 1059–1066, 2006.

[5] O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

Celer: dual extrapolation for the Lasso
[Talk submission]

Mathurin Massias, Alexandre Gramfort, and Joseph Salmon

INRIA/Télécom ParisTech, Université Paris-Saclay

Abstract. To accelerate solvers for ℓ_1-regularized problems, state-of-the-art approaches consist in reducing the size of the optimization problem at hand. In the context of regression, this can be achieved either by discarding irrelevant features (screening) or by prioritizing features likely to be included in the support of the solution (working set). Duality comes into play at several steps in these techniques. We propose an extrapolation technique in the dual that leads to the construction of an improved dual point. This enables a tight control of the stopping criterion, as well as better screening performance of Gap Safe rules. Finally, we propose a working set strategy based on our new dual point, which improves state-of-the-art time performance on the Lasso.

Keywords: Sparsity, Lasso, Duality, Acceleration

1 Motivation/Introduction

Following the seminal work on the Lasso/Basis Pursuit [Tibshirani, 1996, Chen and Donoho, 1995], convex sparsity-inducing regularizations have had a major impact on machine learning [Bach et al., 2012]. In machine learning applications, the default method to optimize such problems is coordinate descent Fu [1998], Friedman et al. [2010]. Since by design only a few features are included in the optimal solution (the *support*), state-of-the-art techniques rely on limiting the size of the (sub-)problem to consider. Various approaches can be distinguished: *screening* techniques [Wang et al., 2013, Fercoq et al., 2015], following the seminal work of El Ghaoui et al. [2012], strong rules [Tibshirani et al., 2012] or correlation screening [Xiang and Ramadge, 2012]. The current state-of-the-art safe rules are so-called Gap Safe rules [Ndiaye et al., 2017] and they rely on the estimation of the duality gap, which itself requires to know a good dual optimal point. Alternatively, *working sets*(WS) techniques [Fan et al., 2008, Johnson and Guestrin, 2015] select a subset of important features according to a particular criterion, and approximately solve the subproblem restricted to these features. A new subset is then defined, and the procedure is repeated. For WS, duality also comes into play, both in the stopping criterion of the subproblem solver and in the WS definition.

2 Methods

The Lasso problem and its dual problem are:

$$\hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{\tfrac{1}{2} \|y - X\beta\|^2 + \lambda \|\beta\|_1}_{\mathcal{P}^{(\lambda)}(\beta)} , \quad \text{and} \quad \hat{\theta}^{(\lambda)} = \underset{\theta \in \Delta_X}{\arg\max} \underbrace{\tfrac{1}{2} \|y\|^2 - \tfrac{\lambda^2}{2} \left\|\theta - \tfrac{y}{\lambda}\right\|^2}_{\mathcal{D}^{(\lambda)}(\theta)} ,$$

where $\lambda > 0$ controls the trade-off between data-fitting and regularization, and $\Delta_X = \{\theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leq 1\}$ is the (rescaled) dual feasible set. The duality gap is defined by $\mathcal{G}^{(\lambda)}(\beta, \theta) := \mathcal{P}^{(\lambda)}(\beta) - \mathcal{D}^{(\lambda)}(\theta)$, for any pair $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$.

Proposition 1. *Strong duality holds for the Lasso, and* $\hat{\theta}^{(\lambda)} = \frac{1}{\lambda}(y - X\hat{\beta}^{(\lambda)})$.

Because of Proposition 1, a canonical choice of dual point during coordinate descent iterations (*i.e.,* , corresponding to a sequence of iterates β^t converging to $\hat{\beta}$) is *residuals rescaling*. It consists in choosing a dual feasible point proportional to the residual $r^t := y - X\beta^t$, see for instance Mairal [2010]: $\theta^t_{\text{res}} := r^t / \max(\lambda, \|X^\top r^t\|_\infty)$.

Building on the work on nonlinear regularized acceleration by Scieur et al. [2016], we propose a better dual point. Instead of relying only on the last residual r^t, we extrapolate previous residuals r^t, r^{t-1}, r^{t-2}, etc.

Definition 1 (Extrapolated dual). *For a fixed number of iterations* $K = 5$*, let*

$$r^t_{\text{accel}} = \begin{cases} r^t, & \text{if } t \leq K, \\ \sum_{k=1}^{K} c_k r^{t+1-k}, & \text{if } t > K, \end{cases} \tag{1}$$

where $c = (c_1, \ldots, c_K)^\top \in \mathbb{R}^K$ *is defined as* $c = z/z^\top \mathbf{1}_K$*, and* z *solves the linear system* $(U^t)^\top U^t z = \mathbf{1}_K$ *with* $U^t = [r^{t+1-K} - r^{t-K}, \ldots, r^t - r^{t-1}] \in \mathbb{R}^{n \times K}$*. Then,*

$$\theta^t_{\text{accel}} := r^t_{\text{accel}} / \max(\lambda, \|X^\top r^t_{\text{accel}}\|_\infty). \tag{2}$$

For iterations $\geq K$*, the* K *last values of the residuals are used to extrapolate the limit of the sequence* (r^t)*, and this extrapolation is rescaled to provide a feasible dual point.*

3 Results

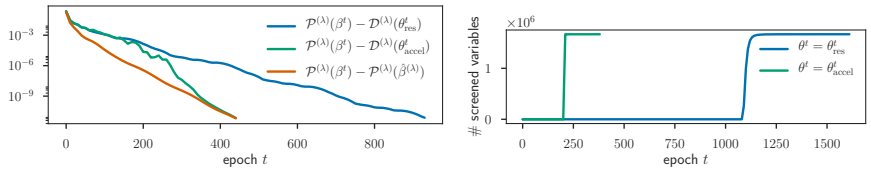


Fig. 1. Left: Duality gaps evaluated with θ_{res} and θ_{accel}, along with the true suboptimality gap. Performance is measured for coordinate descent on the *leukemia* dataset, for $\lambda = \lambda_{\max}/20$. Our gap quickly gets close to the true suboptimality, while the canonical approach constantly overestimates it. Right: Number of variables discarded by the Gap Safe rule as a function of CD epochs, depending on the dual point used, for $\lambda = \lambda_{\max}/5$ (Finance dataset). Our better dual point helps to screen variables earlier.

Better dual point Figure 1 (Left) shows that θ_{accel} gives a duality closer to the true suboptimality gap than θ_{res}, meaning that the proposed construction is indeed better.

Improved screening The gap safe screening rule is: $|x_j^\top \theta| < 1 - \|x_j\| \sqrt{\frac{2}{\lambda^2}\mathcal{G}^{(\lambda)}(\beta, \theta)} \Rightarrow \hat{\beta}_j^{(\lambda)} = 0$. Its performance depends strongly on how well θ approximates $\hat{\theta}^{(\lambda)}$. If the duality gap is large, the left-hand side is small, resulting in fewer discarded features. Figure 1 (Right) shows that θ_{accel} helps discarding more features than θ_{res}, accelerating CD solvers and achieving safe feature identification earlier.

Working sets Working set (WS) approaches involve two nested iteration loops: in the outer one, a set of features $\mathcal{W}_t \subset [p]$ is defined. In the inner one, an iterative algorithm is launched to solve the problem restricted to $X_{\mathcal{W}_t}$ (*i.e.,* considering only the features in $\mathcal{W}_t$). We propose a WS construction based on an aggressive use of Gap Safe rules.

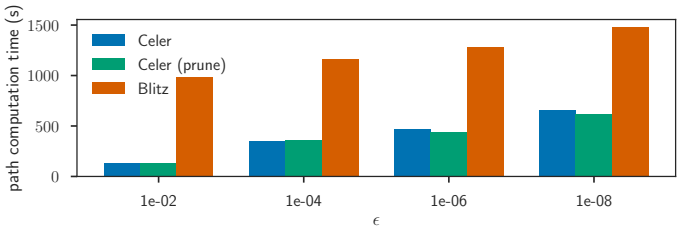


Fig. 2. Times to solve the Lasso path to precision ϵ for a grid of 100 λ's, from $\lambda_{\max}$ to $\lambda_{\max}/100$. Celer outperforms the state-of-the-art package Blitz. Safe and prune versions behave similarly.

As it appears in Gap Safe screening, the critical quantity measuring the importance of feature j is $d_j(\theta) := \frac{1-|x_j^\top \theta|}{\|x_j\|}$, because $d_j(\theta) > \sqrt{\frac{2}{\lambda^2}\mathcal{G}^{(\lambda)}(\beta, \theta)} \Rightarrow \hat{\beta}_j^{(\lambda)} = 0$. Rather than discarding feature j from the problem if $d_j(\theta)$ is too large, twe create a WS with the coordinates achieving the lowest $d_j(\theta)$'s values. A possibility would be to set $r \in]0, 1[$ and creating a WS with features such that $d_j(\theta) < r\sqrt{2\mathcal{G}^{(\lambda)}(\beta, \theta)/\lambda^2}$. However, a pitfall for this strategy is that the WS size is not explicitly under control: an inaccurate choice of r leads to extremely large WS, and which limits the benefits. Instead, to achieve a good control on the working set growth, we reorder the $d_j(\theta)$'s in a non-decreasing way: $d_{j_p}(\theta) \geq \cdots \geq d_{j_1}(\theta)$. Then, for a given working set size p_t, we choose $\mathcal{W}_t = \{j_1, \ldots, j_{p_t}\}$. When subproblems are solved with the same precision ϵ as considered for stopping the outer-loop and if the WS $\mathcal{W}_t$ grows geometrically (*e.g.,* $p_{t+1} = 2p_t$) and monotonically (*i.e.,* $\mathcal{W}_t \subset \mathcal{W}_{t+1}$), then convergence is guaranteed. Indeed, this growth strategy guarantees that as long as the problem has not been solved up to precision ϵ, more features are added, eventually starting the inner solver on the full problem until it reaches an ϵ-solution. We also introduce an unsafe version, *prune*. The initial WS size is set to $p_1 = 100$. This avoids two common WS issues: working sets growing one feature at a time, or too quickly. We have coined this strategy Celer (Constraint Elimination for the Lasso with Extrapolated Residuals). Figure 2 shows that it outperforms state-of-the-art Blitz [Johnson and Guestrin, 2015].

4 Discussion/Conclusion

We want to generalize Celer to sparse logistic regression and ℓ_1-regularized GLMs.

Bibliography

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

S. S. Chen and D. L. Donoho. Atomic decomposition by basis pursuit. In *SPIE*, 1995.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

W. J. Fu. Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.*, 7(3):397–416, 1998.

J. Friedman, T. J. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1, 2010.

J. Wang, J. Zhou, P. Wonka, and J. Ye. Lasso screening rules via dual polytope projection. In *NIPS*, pages 1070–1078, 2013.

O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the lasso. In *ICML*, pages 333–342, 2015.

L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *J. Pacific Optim.*, 8(4):667–698, 2012.

R. Tibshirani, J. Bien, J. Friedman, T. J. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74(2):245–266, 2012.

Z. J. Xiang and P. J. Ramadge. Fast lasso screening tests based on correlations. In *ICASSP*, pages 2137–2140, 2012.

E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res.*, 18(128):1–33, 2017.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

T. B. Johnson and C. Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. In *ICML*, pages 1171–1179, 2015.

J. Mairal. *Sparse coding for machine learning, image processing and computer vision*. PhD thesis, École normale supérieure de Cachan, 2010.

D. Scieur, A. d'Aspremont, and F. Bach. Regularized nonlinear acceleration. In *NIPS*, pages 712–720, 2016.

A Stochastic Fixed Point Method for Empirical Risk Minimization

[Talk submission]

Rui YUAN* Robert M. GOWER** Olivier FERCOQ***

Télécom ParisTech

Abstract. We study the problem of minimizing the average of a large number of smooth convex functions with a convex regularizer. We propose and analyze a stochastic fixed point method which at each iteration samples only a mini-batch of data points and updates a handful of parameters associated to features. In particular, our stochastic fixed point method includes two well known method dfSDCA and Quartz as special cases, thus our work unifies these two methods under one single framework and proposes a more general setting which guarantees theoretically the convergence of the optimal points in the ridge regression problem. The work is concluded with numerical experiments in some more general machine learning problems.

Keywords: stochastic fixed point relaxation method, Quartz, dfSDCA

1 Introduction

This paper is concerned with solve the following problem

$$\min_{x\in\mathbb{R}^d} P(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \phi_i(\langle a_i, x\rangle) + \lambda g(x). \tag{1}$$

In the context of machine learning, $\{a_i\}_{1\leq i\leq n}$ are independent samples of an observation and x is a predictor. The function ϕ_i is the loss incurred by the predictor on sample a_i and is often referred to as the *fidelity term*, since it tells us how *faithful* the predictor is in relation to the sample a_i. The constant $\lambda > 0$ is a regularization parameter and the function g is a regularizer, chosen as to discourage complex solutions. In machine learning, (1) is called the regularized empirical risk minimization problem. It gets this name since it is often interpreted as an empirical estimation of an integral. The setup (1) is also used to describe other applications outside of machine learning, including many best-fit or regression problems. Here we are interested in case when the number of samples n is very big, i.e. millions, billions, and much larger than the dimension d of our predictor. That is, we are in a big data setting.

* rui.yuan@polytechnique.edu
** gowerrobert@gmail.com
*** olivier.fercoq@telecom-paristech.fr

We assume that the functions g and ϕ_i for all $i = 1 \dots n$ are convex to guarantee that the local minimum is the global minimum of the problem (1). We also assume that g and ϕ_i are smooth to guarantee the existence of their derivatives. Thus, the minimum (global or local) is the point x^* such that $\nabla P(x^*) = 0$. The classic algorithm to handle these kinds of problem is the gradient descent method. From any initial point x^0 at time 0, the algorithm makes progress towards the solution at each iteration by taking a step in the direction of the steepest descent. As the functions ϕ_i for all $i = 1 \dots n$ are smooth, the convergence of the algorithm is guaranteed.

We notice that at each iteration, we need to calculate the full gradient $\nabla P(x^t) = \frac{1}{n}\sum_{i=1}^{n} \phi_i'(\langle a_i, x^t \rangle) a_i + \lambda \nabla g(x^t)$, which means we need to go through all the samples once at each single iteration. Not only the gradient descent method, all classic optimization algorithms, such as Newton's method or coordinate descent (go through one single coordinate of each sample), have the same phenomenon. When the number of data samples is huge, this leads to iterations which are too expensive. This is why SGD (*stochastic gradient descent*) methods have become popular: at each iteration, the SGD methods need only a single data sample a_i to make progress towards the solution.

However, the SGD methods have a significant drawback, the iterates are now themselves stochastic and furthermore, the stochastic gradients have a high variance and do not converge to zero when approximating the solution. Thus, a stepsize regime converging to zero needs to be used in conjunction with SGD type methods. This stepsize regime needs to be calibrated to each problem application, which is costly in both time and the patience of the user. This issue has led to the development of stochastic variance reduced methods that converge to the solution without needing to tune a stepsize regime. Here we follow the development of a particular type of stochastic variance reduced method, by developing a family of stochastic fixed point type methods.

To describe this class of fixed point method, first note that the solution x^* to (1) satisfies

$$\nabla P(x^*) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\phi_i'(\langle a_i, x^* \rangle)}_{\stackrel{\text{def}}{=} -\alpha_i^*} a_i + \lambda \nabla g(x^*) = 0,$$

which in turn is equivalent to solve the following equations for variables (α^*, x^*):

$$x^* = \nabla g^{-1}\left(\frac{1}{\lambda n} A\alpha^*\right) \tag{2}$$

$$\alpha_i^* = -\phi_i'(a_i^\top x^*), \quad \text{for } i = 1, \dots, n. \tag{3}$$

where $A = [a_1, \dots, a_n] \in \mathbb{R}^{d \times n}$ and ∇g^{-1} is the operator such that $\nabla g^{-1} \circ \nabla g = id$, which exists due to the convexity of g. As (2) and (3) are sufficient conditions for optimality, we now refer to them as the *optimality conditions*.

Since our optimality conditions are now a fixed point equation, we can apply fixed point methods. We adapt a standard relaxed fixed point method to an equivalent stochastic reformulation of our optimality conditions. As in SGD, we select randomly one sample at each iteration to update variables. The state-of-the-art optimization algorithms applying this type of method are Quartz [2] and dfSDCA [3] that we will resume briefly in Section 2. However, they have different visions of parameters setting. Our objective is to analyze the possibility of getting a more general choice of parameters for this kind of methods and develop variant methods from the fixed point method to improve the convergence rate.

2 Objectives

A natural fixed point strategy for finding a solution to the fixed point equations (2) and (3) is, from a given x^0, α_i^0 for $i = 1, \ldots, n$, to iterate alternatively apply the two equations (2) and (3) at each step, i.e. we do a fixed-point iteration. However, this method tends to be divergent if the iterated operator is not a contraction which is usually the case. So we can not apply the Banach Fixed Point Theorem. Besides, the method is costly since it requires a full sweep through the data at each iteration. To guarantee the convergence, and simultaneously keep the cost of each iteration low, we will use relaxation parameters $\theta, \gamma_i \in (0, 1]$ for $i = 1, \ldots, n$, and we will select one sample $i \in \{1, \ldots, n\}$ from a distribution $\mathcal{D}$. The resulting *Stochastic fixed point relaxation method* is given by

$$\alpha_i^{t+1} = (1 - \gamma_i)\alpha_i^t - \gamma_i \phi_i'(\langle a_i, x^t \rangle), \quad \text{for selected } i, \quad \text{where } i \sim \mathcal{D} \tag{4}$$

$$x^{t+1} = (1 - \theta)x^t + \theta \nabla g^{-1}\left(\frac{1}{\lambda n} A\alpha^t\right). \tag{5}$$

If we replace α_i^t and α_i^{t+1} by α_i^*, x^t and x^{t+1} by x^* in the above update, we obtain then the same fixed points solutions of equations (2) and (3). This is because that equations in (4) and (5) are the convex combination of the identity and the solution expression in equations (2) and (3). So we still apply a fixed-point iteration, but for equations (4) and (5). Intuitively, if we choose good relaxation parameters θ and γ, the iterated operator can be contractive. So the algorithm will converge.

We notice that if θ and γ are closed to zero, the iterated operator can be contractive. However, θ and γ are considered as the step size of the update. So we need θ and γ to be as big as possible, i.e. closed to one, so that the algorithm will converge as fast as possible. Hence, the trade-off between contraction and convergence rate leads to the optimal relaxation parameters θ^* and γ^*.

The Quartz [2] and the dfSDCA [3] methods are both instantiations of the stochastic fixed point relaxation method. For the Quartz, we take $\gamma_i = \frac{\theta}{p_i} \in (0, 1]$ with p_i the probability of sample i being selected. The optimal setting is that, $p_i^* = \frac{L\|a_i\|^2 + \lambda n}{\sum_{i=1}^n \left(L\|a_i\|^2 + \lambda n\right)}$ and $\theta^* = \frac{\lambda n}{\sum_{i=1}^n \left(L\|a_i\|^2 + \lambda n\right)}$ where ϕ_i is $L-$smooth for $i = 1 \ldots n$. For the dfSDCA, we take $\theta = 1$ and $\gamma_i = \eta\lambda n \in (0, 1]$ for all i with i following the uniform sampling. The optimal setting takes $\eta = \frac{1}{L + \lambda n}$ with L the smooth constant for all the functions ϕ_i. However, the proof of the convergence of these methods rely on different proof techniques and different conditions. In addition, the choice of parameters is limited in the sense that the parameter γ_i needs to be proportional to $1/p_i$, same case in the dfSDCA in considering $p_i = 1/n$ for the uniform sampling. In this paper, we analyze this class of methods for a specific problem - the ridge regression problem and prove the convergence of the method with a setting of parameters which has more freedom than the one of Quartz and dfSDCA. The proof is inspired by the analyze of [1].

References

1. A. Alves Riberio and P. Richtárik. The Complexity of Primal-Dual Fixed Point Methods for Ridge Regression. *ArXiv e-prints*, January 2018.

2. Zheng Qu, Peter Richtárik, and Tong Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 865–873, Cambridge, MA, USA, 2015. MIT Press.
3. Shai Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. *CoRR*, abs/1602.01582, 2016.

Optimal mini-batch size for stochastic variance reduced methods

[Poster/Talk submission]

Nidham GAZAGNADOU, Robert M. GOWER

LTCI, Telecom ParisTech, France
first.last@telecom-paristech.fr

Abstract. This work presents an analysis leading to the optimal mini-batch size for SAGA, a recent stochastic variance reduced method used for training many machine learning models without tuning any stepsize parameter. This study is based on the recent *JacSketch* framework, which unifies several stochastic variance reduced methods for different type of sampling. Analyzing the total complexity of this SAGA optimization method leads us to an approximation of the optimal mini-batch size.

Keywords: Empirical risk minimization, Stochastic variance reduced methods, Stochastic average gradient descent, Optimal mini-batch

1 Context and motivation

The empirical risk minimization problem often arises when training classical machine learning models. Today, gigantic datasets (sometimes several terabytes) from Internet, images or text, are used to train machine learning algorithms, such as logistic regression for classification or conditional random fields. Thus, one cannot perform the required minimization by computing a full gradient descent (GD) because it would be too costly.

In order to address this issue and solve such problems efficiently, the optimization community has revived an old method from the 1950's, the stochastic gradient descent (SGD) method [1]. SGD has established itself as a reference method for minimizing the empirical risk thanks to its scalability. Yet, in order to converge one has to tune a problem dependent stepsize sequence which often leads to suboptimal results.

This last issue has been tackled by the recent development of stochastic variance reduced gradient methods, which do not require any stepsize tuning to ensure convergence, like SAG/SAGA [2,3]. Theses methods keep an estimate of the gradient and update it using only a randomly subsampled set of the data at each iteration. This size of this subsampled set is referred to as the *mini-batch size*. The idea of mini-batching lies in the fact that, at each iteration, the gradient computed on a subsampled dataset randomly picked might be a good trade-off between a stochastic gradient, computed at a single random point, and the full gradient, computed over all the dataset. This question could be addressed for SGD too, but in this work we only focus on stochastic variance reduced gradient methods.

Though ensuring linear convergence and relieving the user of this stepsize tuning, one still has to set the mini-batch parameter. The aim of this project is to determine an approximation of the mini-batch size for SAGA by minimizing the total complexity of the optimization algorithm. To that end, we extend the mini-batching and complexity studies of the recent JacSketch [4] methods, which include mini-batch SAGA.

2 Methods

Authors of [4] computed the iteration complexity for mini-batch SAGA sketches (if one uses the right stepsize). Since each step of mini-batch SAGA computes τ stochastic gradients, the total complexity is τ times the iteration complexity which we note $K_{\text{total}}(\tau)$. Theorem 3.6 & Theorem 4.19 of [4] imply that, in the case of mini-batch SAGA with τ-nice (uniform sampling without replacement of mini-batch of size τ), the total complexity boils down to

$$K_{\text{total}}(\tau) = \max\left\{\frac{4\tau(\mathcal{L}_1+\lambda)}{\mu}, n + \frac{n-\tau}{n-1}\frac{4(L_{\max}+\lambda)}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right) \quad , \tag{1}$$

where $\mathcal{L}_1$ denotes the expected smoothness constant, which measures how smooth the expected stochastic gradient is, μ is the strong convexity parameter and $L_{\max}$ the largest smoothness constant of the individual subsampled functions.

So, our goal is to find the optimal τ that minimizes this total complexity of the optimization method. Unfortunately, the computation of the expected smoothness constant $\mathcal{L}_1$ turns out to be intractable for large n because it requires browsing all the τ-combinations from n. By finding upper bounds of $\mathcal{L}_1$, we show that $K_{\text{total}}(\tau)$ is bounded by $\hat{K}_{\text{total}}(\tau)$, the maximum of two computable terms $g(\tau)$ and $h(\tau)$. These terms $g(\tau)$ and $h(\tau)$ can be expressed as functions of several smoothness constants and other parameters of the minimization problem, such as the n or d. In Figure 1 we present one of our total complexity bounds for a traditional SAGA method with τ-nice sampling.

Thus, a preliminary work before studying the minimization of the total complexity is to find an upper bound of $\mathcal{L}_1$. To get the sharpest bounds, several tools are used such as properties of the smoothness constants or matrix concentration inequalities [5], which give us more insight on how the sampling randomness affects the bound, and thus the complexity.

We already succeeded to find an explicit value of the mini-batch size for SAGA with τ-nice sampling. In this case, computing the optimal τ boils down to finding the intersection of two linear functions as shown in Figure 1.

3 Experiments

With experiments solving ridge regression, we aim to estimate how tight are our upper bounds of the expected smoothness constant $\mathcal{L}_1$ and how accurate is our estimate of the optimal mini-batch size $\hat{\tau}^*$. We run simulations on both artificially generated data (gaussian and diagonal features matrices $A = [a_1, \ldots, a_n]$) and real data from LIBSVM datasets (abalone, housing) from [7].

Depending on the structure of the feature matrix A, one can see emerging some regimes of τ for which one particular bound is sharper than the others. Those experiments also clearly show that a heuristic 'bound' we introduce almost overlaps with the true $\mathcal{L}_1$. Much in-depth experiments are in progress to determine whether our different estimates of the optimal mini-batch size are close to its empirical value. Because we upper bound $\mathcal{L}_1$, one can think that our method might underestimate the optimal value of τ. Indeed, looking jointly at (1) and Figure 1 clarifies this idea: having a loose upper bound of $\mathcal{L}_1$ is equivalent to shifting up the g curve and leads to a left shift of the optimum.

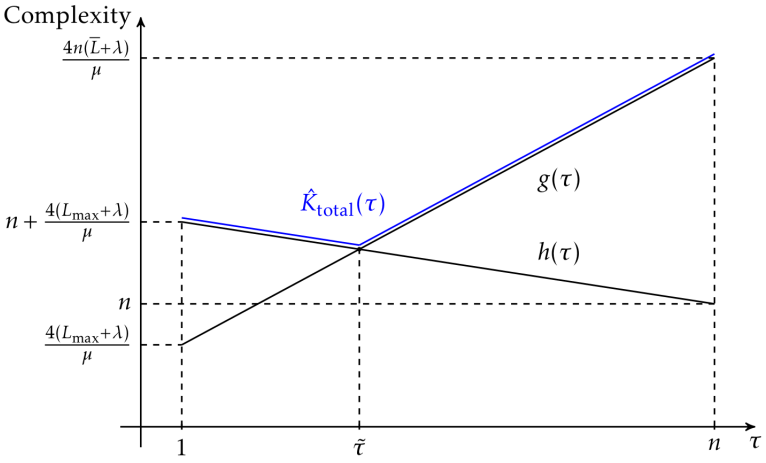


Fig. 1: Outline of the optimal mini-batch size $\tilde{\tau}$ for SAGA with identity weights.

4 Conclusion and future work

In this work, we proposed to minimize the total complexity of stochastic variance reduced methods, such as SAGA, in order to determine the optimal mini-batch size. We succeeded to get explicit estimates of the mini-batch size for SAGA with τ-nice sampling by minimizing an upper bound of the complexity.

Sharper results are also proved through matrix concentration inequalities and numerical experiments are in progress to see if our optimal mini-batch estimates do improve the convergence of stochastic variance reduced gradient methods. The analysis could also be extended for another well-known stochastic variance reduced method, the SVRG [6] method, but we leave it for later work.

References

[1] Robbins, H.; Monro, S. "A Stochastic Approximation Method". The Annals of Mathematical Statistics. 22 (3): 400, 1951

[2] M. Schmidt, N. Le Roux, F. Bach. "Minimizing Finite Sums with the Stochastic Average Gradient". Mathematical Programming, 162(1):83-112, 2016.

[3] A. Defazio, F. Bach, S. Lacoste-Julien. "SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives". Advances in Neural Information Processing Systems (NIPS), 2014.

[4] R. M. Gower, P. Richtárik, and F. Bach. "Stochastic Quasi-Gradient Methods: Variance Reduction via Jacobian Sketching". In: arxiv:1805.02632, 2018.

[5] J. A. Tropp. "An introduction to matrix concentration inequalities". In: arXiv:1501.01571, 2015.

[6] R. Johnson and T. Zhang. "Accelerating stochastic gradient descent using predictive variance reduction". Advances in Neural Information Processing Systems (NIPS), 315-323, 2013.

[7] C. C. Chang and C. J. Lin. "LIBSVM : A library for support vector machines". In:ACM Transactions on Intelligent Systems and Technology 2.3, 127, 2011.

TEXT PROCESSING

TALK SESSION

Automated extraction of food-drug interactions from scientific articles

Tsanta Randriatsitohaina

LIMSI, CNRS, Univ. Paris-Sud, Universit Paris-Saclay, F-91405 Orsay

Abstract. In this paper, we are interested in the extraction of food-drug interactions (FDI), a task which is similar to the extraction of relation between terms in specialized texts. We present a supervised classification method and the results of a first set of experiments. Despite the imbalance of classes, the results are encouraging. We have identified the most relevant classifiers according to the steps of our method. We have also observed the important impact of the semantic tags of terms used as features.

Keywords: Food-Drug Interaction, Semantic relation, Specialized corpora, Supervised learning

1 Introduction/Motivation

Although knowledge bases (KB) or terminologies exist in specialized domains, updating this information often requires to access unstructured data such as scientific literature. The problem occurs deeply when focusing on a new type of knowledge which has no recording in terminological resources yet. Thus, while drug interactions [1] or drug adverse effects [2] are listed in databases such as DrugBank [1] or Theriaque [2], other information such as food drug interactions is barely listed in KB, mainly sparsed in the scientific literature and recorded in textual form. However food-drug interactions correspond to various types of adverse drug effects and lead to harmful consequences on the patients health and well-being. We focus on extracting these interactions as a relation acquisition task given the references to a food and a drug.

In this article[3], our experiments rely on the extraction method in a single sentence as proposed by [6] using approach in [4] to form instances: all couples *food-drug*, *food_supplement-drug* or *food-side_effect* appearing in the same sentence are extracted to form positive instances if they are in relation and negative otherwise. Then we define a two-step method: detection of relevant relations and classification as proposed by [3]. As in [9], we have not explicitly introduced handcrafted features. Instead, we followed the logic of [7] to generalize the named entities replacing them with their semantic tag. Among the five classifiers we have experimented, four are mentioned in the state of the art: a linear SVM and decision tree [4], a Bayesian classifier [8], and a multilayer perceptron [9] and the last one is a logistic regression classifier. We compare the performance of these classification algorithms with default parameters provided by Scikit-Learn[4] When mining scientific article abstracts, we face several difficulties: (1) drug and food occurrences are very variable. Drug mentions are the international non proprietary name or the active drug substances. Foods may be reference to a particular nutrient, food component or family; (2) interactions are described in a rather precise way in the texts. It leads to a limited number of examples; (3) interactions are heterogeneously annotated in an unbalanced learning set. Our contribution are: (1) the selection of relevant sentences for food-drug interaction, (2) the classification of positive relations.

[1] `https://www.drugbank.ca/`

[2] `http://www.theriaque.org`

[3] This work was supported by the ANR through the grant ANR-16-CE23-0012 (MIAM project).

[4] `http://scikit-learn.org/stable/`

2 Experiments

2.1 Data

Our data consists of 2,341 positive instances and 25,231 negative instances extracted on 639 abstracts of scientific articles collected from the PubMed portal by the query: (FOOD DRUG INTERACTIONS"[MH] OR "FOOD DRUG INTERACTIONS*") AND ("adverse effects*") annotated with `Brat`. The corpus collection process and annotation scheme are detailed in [5]. Positive instances are categorized into 21 types of relation but we have grouped these relations into 4 groups: no relation, direct food-drug interaction (349 instances), drug adverse effect (1,242 instances), relation without precision (724 instances). Then we vectorized these sentences based on word counting: each sentence is represented by a vector corresponding to the number of occurrences of each word of the whole vocabulary in the sentence.

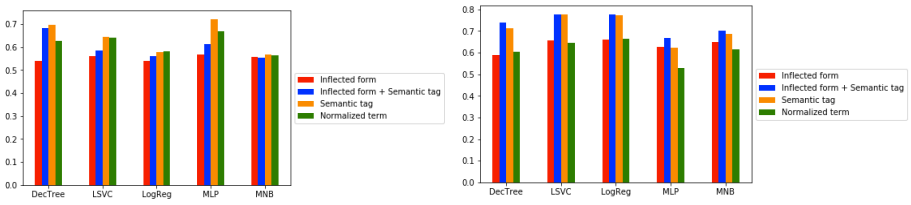
2.2 Evaluation and features

We evaluate our models using F1-score from 10-fold cross-validation. F1-score (F1) is the harmonic average of the precision (P) and recall (R) such that

$$F1 = 2.\frac{P * R}{P + R} \; P = \frac{correct\ positive\ results}{all\ positive\ results} \; R = \frac{correct\ positive\ results}{all\ returned\ results}$$

To train our models, we used four sets of features:

1. Inflected form of words (word form as it occurs in text)
   *ex: Bioavailability enhancement by **grapefruit juice** noted with **dihydropyridine calcium antagonists** does not occur with amlodipine.*
2. Inflected form of words and terms (i.e. the noun phrases conveying specialized concepts) followed by their semantic tag
   *ex: Bioavailability enhancement by **grapefruit juice /food/** noted with **dihydropyridine calcium antagonists /drug/** does not occur with amlodipine.*
3. Generalization of terms with their semantic tag without losing information about the entities
   *ex: Bioavailability enhancement by **food** noted with **drug** does not occur with amlodipine.*
4. Normalization of the arguments of the relations (replaced by `arg1` and `arg2`)
   *ex: Bioavailability enhancement by **arg1** noted with **arg2** does not occur with amlodipine.*



(a) Step 1: Binary classification (b) Step 2: Multi-class classification

Fig. 1: F1-score on 10-fold cross

3 Results

Step 1 : Binary classification. Figure 1a presents the results obtained to identify the sentences containing relevant relations. Depending on the descriptors used, F1-score varies between 0.54 and 0.71. Best results are obtained with decision trees and perceptron (MLP) using semantic tags.

Step 2 : Multi-class classification. Figure 1b presents the results obtained when recognizing group of relations. As in the previous step, the use of semantic tags of terms has a positive impact on the results. Among the models used, the Naive Bayes classifier is the one leading to the weakest results. We also obtain good results with Logistic Regression, Decision Tree and linear SVC.

4 Conclusion and future work

We propose a first step towards the extraction of food-drug interaction. These first experiments show that features including semantic tags lead to best results. As perspectives, we will pursue the next two steps of our method: (1) recognition of the various types of relations and (2) identification of related entities (food, medicine, disease, etc.). The preliminary results presented in this article need to be improved. We are considering the use of other classification methods such as convolutional deep neural networks using word embedding. We also want to study the impact of other descriptors (word lemmas, part-of-speech tags, syntactic relations, semantic tags of terms with different levels of granularity, etc.) or sampling methods to reduce the imbalance of the data.

References

1. Aagaard, L., Hansen, E.: Adverse drug reactions reported by consumers for nervous system medications in europe 2007 to 2011. BMC Pharmacology & Toxicology 14, 30 (June 2013)
2. Aronson, J., Ferner, R.: Clarification of terminology in drug safety. Drug Safety 28(10), 851–70 (2005)
3. Ben Abacha, A., Chowdhury, M.F.M., Karanasiou, A., Mrabet, Y., Lavelli, A., Zweigenbaum, P.: Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification. Journal of Biomedical Informatics 58, 122–132 (October 2015)
4. Cejuela, J.M., Vinchurkar, S., Goldberg, T., Prabhu Shankar, M.S., Baghudana, A., Bojchevski, A., Uhlig, C., Ofner, A., Raharja-Liu, P., Jensen, L.J., Rost, B.: LocText: relation extraction of protein localizations to assist database curation. BMC Bioinformatics 19(1), 15 (Jan 2018)
5. Hamon, T., Tabanou, V., Mougin, F., Grabar, N., Thiessard, F.: Pomelo: Medline corpus with manually annotated food-drug interactions. In: Proceedings of Biomedical NLP Workshop associated with RANLP 2017. pp. 73–80. Varna, Bulgaria (September 2017)
6. Lee, K., Kim, B., Choi, Y., Kim, S., Shin, W., Lee, S., Park, S., Kim, S., Tan, A.C., Kang, J.: Deep learning of mutation-gene-drug relations from the literature. BMC Bioinformatics 19(1), 21 (Jan 2018)
7. Liu, S., Tang, B., Chen, Q., Wang, X.: Drug-drug interaction extraction via convolutional neural networks. Computational and Mathematical Methods in Medicine 2016 (10 2016)
8. Ramani, A.K., Bunescu, R.C., Mooney, R.J., Marcotte, E.M.: Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. Genome Biol 6(5), R40–R40 (April 2005)
9. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. J. Mach. Learn. Res. 3, 1083–1106 (Mar 2003)

User modelling for the characterisation of eating behaviors

[Paper submission]

Sema Akkoyunlu

UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, France

Abstract. Dietary guidelines provided by public health agencies have not been very successful so far, as most people do not comply with nutritional recommendations. Several reasons can contribute to this phenomenon. First, dietary guidelines are generic: they do not take into account personal preferences or constraints. Second, usually, they bear on single categories and food items, advising to promote, limit, or substitute foods for others without considering that food items are rarely consumed alone but are eaten as part of structured meals. Our objective is to find clusters of consumers with regards to their consumption data. We propose a novel approach to describe users consumption data. Classical clustering methods can then be used in order to identify subgroups of consumers.

Keywords: user modelling, eating behavior, Doc2Vec

1 Motivation

Most chronic diseases such as diabetes, obesity and cardiovascular diseases are correlated to unhealthy eating habits [9]. Public health agencies created dietary guidelines targeted to the general population to help people to adopt healthier eating habits. However the compliance to the dietary guidelines are relatively low although the awareness about food based dietary guidelines is rather good [3]. Several causes contribute to this phenomenon: cultural and personal preferences, difficulty of implementing dietary changes, availability and price of food items [7]. Nutritionists stress the fact that it is essential to understand consumers' eating behavior in order to make practical food-based recommendations because making changes is challenging [1].

User modeling is defined as the creation of a representation of the user for the purpose of customization [10]. It is crucial to understand the application area in order to find measurable properties of a behavior. In nutrition, dietary behavior is modeled using two main types of methods: theoretical ones and empirical ones. We only focus on empirical ones as our goal is to learn dietary behaviors based on consumption data. The main approach consists in applying PCA on the matrix of food item occurrences. The literature in this domain is unclear about which method to choose as existing experiments are not easily reproducible. Moreover, it has been noted that in the scenario where users should use a nutrition recommender system, more complex information would be needed [8].

User profiling is often used for personalization as it is required to analyze users' behavior for adapting the recommendations [2]. Most of the time, user profiling consists in finding an embedding space appropriate for the application. In our application case, we want to create a representation of users based on their consumption data in order to discover eating behaviors.

We propose two methods to describe users' consumption data: a food item based approach and a meal based approach. These two approaches rely on two hypotheses about eating behavior. The item based approach relies on the hypothesis that an eating behavior is represented by the occurrence of consumption of food items during a given amount of time. However, this approach aggregates too much information and may not be sufficient for describing eating behaviors. The meal based approach describes users based on the meals they have consumed. Indeed, the way that people compose their meals may be an indicator of eating behavior.

2 Our approach

Dietary data is collected with consumption diaries in which users write down every item they have eaten. For instance, the user U_1 ate lunch $m_1 = \{\text{soup, rice, beans, tiramisu}\}$. Data are collected during one week.

The food item approach consists in representing a user by a vector of occurrence of food items and applying a matrix factorization algorithm such as PCA or NMF. It infers a latent space in which users are represented. Then any clustering algorithm is applied for discovering subgroups of users. This is the state of the art regarding user profiling for eating behavior.

However in the food item approach, the fact that food items are consumed with other food items is not taken into account. A meal based approach can provide extra information for clustering users as the way people structure their meals may be a better modelisation for closeness in eating behaviors. In this approach, users are described by meals (i.e sets of food items). This representation raises the question of the distance to be used for clustering. Clustering requires a distance between meals, i.e a distance defined between sets of items. However, there is no trivial way of computing such a distance as there is no distance defined between food items.

We overcome this limitation by considering each meal as a short text. NLP methods can then be applied for learning a document embedding for meals. It is then possible to compute distances between meals, i.e a mapping that converts documents into vectors. For this purpose, we use a popular algorithm for learning document embedding, Doc2Vec developed by Le and Mikolov in [5] as an extension of Word2Vec [6].

Doc2Vec is an algorithm that learns an embedding for documents. It is a neural network with a single hidden layer of which the task is a prediction task. Doc2vec is proposed with two flavors: DBOW (Distributed bag of words) and DMPV (Distributed Memory Paragraph Vector). DBOW is a simpler model in which the word order is ignored whereas DMPV is more complex as more parameters are learned. In our case, the word order in the declared meals is meaningless as participants were not asked to write down the food items in the order they ate them. Hence, we chose to use the DBOW version of the algorithm. Besides, it has been shown that DBOW outperforms DMPV[4] on similar tasks.

After applying Doc2Vec on meals, a user is represented as a set of points in the meal space. As a first approach, we choose to compute the mean of those points for the

description of users. Traditional clustering methods will be applied using the cosine distance.

3 Conclusion

We introduce a novel approach for describing users' food consumption data and will compare it with the method in the state of art. The results of clustering based on the meal approach can give a finer insight about the different consumption styles from a nutrition point of view but also for personalized recommendation purposes (e.g collaborative filtering).

References

1. Bier, D. M., Derelian, D., German, J. B., Katz, D. L., Pate, R. R., and Thompson, K. M. Improving compliance with dietary recommendations. *Nutrition Today 43*, 5 (sep 2008), 180–187.
2. Eskandanian, F., Mobasher, B., and Burke, R. A clustering approach for personalizing diversity in collaborative recommender systems. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (New York, NY, USA, 2017), UMAP '17, ACM, pp. 280–284.
3. Ivens, B. J., and Smith Edge, M. Translating the Dietary Guidelines to Promote Behavior Change: Perspectives from the Food and Nutrition Science Solutions Joint Task Force. *J Acad Nutr Diet 116*, 10 (Oct 2016), 1697–1702.
4. Lau, J. H., and Baldwin, T. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016* (2016), pp. 78–86.
5. Le, Q., and Mikolov, T. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* (2014), ICML'14, JMLR.org, pp. II–1188–II–1196.
6. Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *CoRR abs/1301.3781* (2013).
7. Webb, D., and Byrd-Bredbenner, C. Overcoming consumer inertia to dietary guidance. *Advances in Nutrition 6*, 4 (jul 2015), 391–396.
8. Wendel, S., Dellaert, B. G., Ronteltap, A., and van Trijp, H. C. Consumers' intention to use health recommendation systems to receive personalized nutrition advice. *BMC Health Services Research 13*, 1 (apr 2013).
9. World Health Organization. Diet, nutrition and the prevention of chronic diseases: report of a joint who/fao expert consultation. Tech. rep., 2003.
10. Yampolskiy, R. Behavioral modeling: an overview.

POSTER SESSION

All student papers accepted for an oral presentation are also invited to present their work in the poster session.

The list of submissions accepted for the poster session, in addition to those accepted for oral presentation, are as follows:

- [P1] Improving the generalization capacity of nonlinear regression and classification models with generative models. Nabil Benmerad. *DCBrain - Microsoft AI Factory*
- [P2] H_/ H_infinity Robust Observer for Actuator Fault Detection and Diagnosis of a Quadrotor. Eslam Abouselima, Said Mammar and Dalil Ichalal. *CEA*
- [P3] Identification of causal factors leading people to choose high protein food. Irène Demongeot, Antoine Cornuejols and Nicolas Darcel. *AgroParisTech & INRA*
- [P4] Logical approach to identify Boolean networks modelling cell differentiation. Stéphanie Chevalier, Andrei Zinovyev, Christine Froidevaux and Loïc Paulevé. *LRI*
- [P5] A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization. Robin Vogel, Aurélien Bellet and Stéphan Clémençon. *Télécom ParisTech*
- [P6] A Latent Model for Representation Learning on Networks. Abdulkadir Celikkanat and Fragkiskos Malliaros.*CentraleSupelec*
- [P7] Graph Matching and Transfer Learning: How to learn a new model from an existing one. Jiang You *AgroParisTech & INRA*
- [P8] Dimension Reduction Adapted to Paleogenomics. Séverine Liegeois, Olivier François and Flora Jay. *LRI*
- [P9] Memory Bandits: Toward The Switching Bandit Problem Best Resolution. Reda Alami, *Inria & Orange Labs*.
- [P10] Transcriptomics data explorations to decipher iron homeostasis in the pathogenesis yeast Candida glabrata. Thomas Denecker and Gaëlle Lelandais, *CEA*