Workshop on Optimization, Machine Learning, and Data Science

Book of Abstracts

April 10, 2018

Perspectives on Integer Programming in Sparse Optimization

Jeffrey Linderoth

University of Wisconsin-Madison

Algorithms to solve mixed integer linear programs have made incredible progress in the past 20 years. Key to these advances has been a mathematical analysis of the structure of the set of feasible solutions. We argue that a similar analysis is required in the case of mixed integer quadratic programs, like those that arise in sparse optimization in machine learning. One such analysis leads to the so-called perspective relaxation, which significantly improves solution performance on separable instances. Extensions of the perspective reformulation can lead to algorithms that are equivalent to some of the most popular, modern, sparsity-inducing non-convex regularizations in variable selection.

Based on joint work with Hongbo Dong (Washington State Univ.), Oktay Gunluk (IBM), and Kun Chen (Univ. Connecticut)

A Trust-Region Method for Nonconvex Finite-Sum Minimization

Robert Mohr
Institute of Operations Research
Karlsruhe Institute of Technology

February 11, 2018

Extended Abstract

We consider the minimization problem

$$\min_{x \in \mathbb{R}^d} F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x),$$

where $n, d \in \mathbb{N}$ and $f_i \in C^2(\mathbb{R}^d, \mathbb{R})$ for all $i = 1, ..., n$. Problems of this structure arise frequently in supervised machine learning in form of the empirical risk minimization problem, a well known example being the training of multilayer perceptrons. In these applications, n is the number of datapoints and typically very large.

Successful algorithms from unconstrained nonlinear optimization, such as BFGS and trust-region methods, are not well suited for the minimization of F, since they require the computation of the gradient and (approximate) hessian of the objective function in every iteration. For large n, these computations are expensive and inhibit fast progress in the early stages of the optimization process.

For this reason, stochastic gradient decent is currently the most widely used method in the "big data regime" and a large number of variants have been developed over the years, see for example the survey papers Bottou et al. (2017) and Curtis & Scheinberg (2017). One common drawback of most of these methods is that certain parameters (typically the step size) have to be determined by the user through experimentation. Therefore, it is highly desirable to devise methods that require no (or less) experimentation by the user. We argue that randomized variants of trust-region methods, which so far have not gotten a lot of attention in the machine learning community, could be an important step in this direction.

The method that we propose is called **A**daptive **D**ynamic **S**ample Size **T**rust-Region method, or ADST for short. In the beginning, the method operates only on averages of small samples of the functions f_i. During the optimization process the size of the samples is adaptively increased (or decreased) depending on the progress made on the objective function F. We prove that after a finite number iterations the sample size reaches n and the method becomes a full-batch trust-region method. As a result, our method provides a new and efficient way to apply trust-region methods to finite-sum minimization problems arising from machine learning applications with a large number of datapoints. Numerical experiments on logistic regression and multilayer perceptron training problems support our claim that our algorithm has significant advantages compared to current state-of-the-art methods.

The Mismatch Principle:

What Can the Lasso Learn About Non-Linear Observations?

Martin Genzel (TU Berlin)

In many real-world problems, one is given a collection of (random) *sample pairs* $(\boldsymbol{a}_1, y_1), \dots, (\boldsymbol{a}_m, y_m) \in \mathbb{R}^n \times \mathbb{R}$ where $\boldsymbol{a}_i \in \mathbb{R}^n$ is *data* (inputs) and $y_i \in \mathbb{R}$ *observations* (outputs). A typical problem issue is then the following: *What can we learn about the relationship between the input and the output variables?* Although one often does not impose very specific restrictions on the model, it is useful to think of some (unknown) parameters that determine the underlying observation rule. A prototypical example are *single-index models* for which the observations take the form

$$y_i = f(\langle \boldsymbol{a}_i, \boldsymbol{x}_0 \rangle), \quad i = 1, \dots, m,$$

where $f \colon \mathbb{R} \to \mathbb{R}$ can be unknown, non-linear, and noisy. The goal is then to recover the unknown (structured) parameter vector $\boldsymbol{x}_0 \in \mathbb{R}^n$. Another important scenario is *variable selection*: There exists a set of active variables $S = \{j_1, \dots, j_s\} \subset [n]$ such that

$$y_i = F(a_{i,j_1}, \dots, a_{i,j_s}), \quad i = 1, \dots, m,$$

for some unknown, non-linear function $F \colon \mathbb{R}^s \to \mathbb{R}$. Here, one is rather interested in finding S.

In this talk, we study how the *generalized Lasso* performs on such types of non-linear models:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \tfrac{1}{2m} \sum_{i=1}^{m} (y_i - \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle)^2 \quad \text{subject to} \quad \boldsymbol{x} \in K, \tag{P_K}$$

with $K \subset \mathbb{R}^n$ being a convex *constraint set* that encourages a certain structure of the solution (e.g., sparsity). It should be emphasized that (P_K) is a standard estimator that is widely-used in practice and does not require any specific knowledge of the underlying observation rule. The recent works of [1–4] show that the Lasso—although originally designed for linear regression—is surprisingly robust against non-linear distortions and can in fact handle much more complicated situations. A simplified and informal recovery guarantee may read as follows:

Theorem 1 (informal, cf. [3, Thm. 6.4]) *Suppose that* $\{(\boldsymbol{a}_i, y_i)\}_{i=1}^m$ *are independent samples of a joint random pair* $(\boldsymbol{a}, y) \in \mathbb{R}^n \times \mathbb{R}$*, where* $\boldsymbol{a}$ *is an isotropic, mean-zero sub-Gaussian random vector in* $\mathbb{R}^n$*. Fix an arbitrary* target vector $\boldsymbol{x}^\natural \in K \subset \mathbb{R}^n$*. Then, with high probability, any minimizer* $\hat{\boldsymbol{x}}$ *of* (P_K) *satisfies the following error bound:*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}^\natural\|_2 \lesssim \Big(\frac{w(K)^2}{m}\Big)^{1/4} + \rho(\boldsymbol{x}^\natural), \tag{1}$$

where $w(K)$ *denotes the* Gaussian width *of* K *and* $\rho(\boldsymbol{x}^\natural) \coloneqq \big\|\mathbb{E}[(\langle \boldsymbol{a}, \boldsymbol{x}^\natural \rangle - y)\boldsymbol{a}]\big\|_2$ *is called the* mismatch covariance*.*

Remarkably, this statement holds true for *every* choice of $\boldsymbol{x}^\natural$ and there are no assumptions on the output variable y. But in order to turn (1) into a meaningful error bound, one needs to ensure that the offset term $\rho(\boldsymbol{x}^\natural)$ is sufficiently small. If the target vector $\boldsymbol{x}^\natural$ can be chosen in such a way, Theorem 1 states that the Lasso yields a good estimator of $\boldsymbol{x}^\natural$. According to our initial problem issue, we can therefore formulate a general "recipe" to prove theoretical guarantees for the Lasso:

> *Determine a target vector* $\boldsymbol{x}^\natural \in K$ *that captures the "parametric" structure of the observation rule and minimizes the mismatch covariance* $\rho(\boldsymbol{x}^\natural)$ *at the same time.*

This *mismatch principle* particularly indicates when one can expect reasonable outcomes of (P_K) and when not. A crucial role is obviously played by the mismatch covariance because it measures the compatibility between the linear fit of (P_K) and the true (parametric) model. In the above examples, we would have to specify a target vector in $\operatorname{span}\{\boldsymbol{x}_0\} \cap K$ for single-index models and in $\{\boldsymbol{x} \mid \operatorname{supp}(\boldsymbol{x}) \subseteq S\} \cap K$ for variable selection, respectively. In fact, it turns out that in both cases (with Gaussian data) there always exists an appropriate choice of $\boldsymbol{x}^\natural$ such that $\rho(\boldsymbol{x}^\natural) = 0$.

This is joint work with Peter Jung (TU Berlin) and Gitta Kutyniok (TU Berlin).

Computing a Bivaraiate Partial Information Decomposition Measure

Abdullah Makkeh[*], Dirk Oliver Theis[†], Raul Vicente[‡]

Institute of Computer Science
Univeristy of Tartu, 51014 Tartu, Estonia

January 2018

Partial information decomposition is the decomposition of mutual information into shared, unique, and synergistic information. In other words, the decomposition allows quantifying the information any of the source signals has about the target source in a complex system.

The first consistent information decomposition is due to William and Beer [7]. They introduced the so-called *William Beer axioms* which are natural properties of shared information. From these axioms, they proposed the partial information lattice framework for partial information decomposition. Then they proposed a measure for shared information $I_{\min}$ which suffered from serious flows. The $I_{\min}$ measure provoked a series of papers trying to improve the measure [1–4].

Bertschinger et al. [2] proposed a measure for computing partial information decomposition. The four information quantities are obtained by solving a Convex Program. The Convex optimization is ill-conditioned and hard to solve. We reformulated their Convex Program as a Cone Programming over the exponential cone [5, 6]. But often scientists want to compute the partial information decomposition subjected to some constraints, so we show how to obtain subgradients in practice [5].

[*]abdullah.makkeh@ut.ee
[†]dotheis@ut.ee
[‡]raul.vicente.zafra@ut.ee

On big data, optimization and learning

Andrea Lodi

École Polytechnique de Montréal

In this talk I review a couple of applications on Big Data that I personally like and I try to explain my point of view as a Mathematical Optimizer – especially concerned with discrete (integer) decisions – on the subject. I advocate a tight integration of Machine Learning and Mathematical Optimization (among others) to deal with the challenges of decision-making in Data Science. For such an integration I try to answer three questions: 1) what can optimization do for machine learning? 2) what can machine learning do for optimization? 3) which new applications can be solved by the combination of machine learning and optimization?

Deep Learning Assisted Heuristic Tree Search for the Pre-Marshalling Problem

André Hottung[1], **Shunji Tanaka**[2], **Kevin Tierney**[1]

[1]Decision Support & Operations Research Lab, University of Paderborn, Germany
ahottung@mail.upb.de,kevin.tierney@upb.de

[2]Institute for Liberal Arts and Sciences & Department of Electrical Engineering, Kyoto University, Japan
tanaka@kuee.kyoto-u.ac.jp

Extended Abstract

One of the key challenges for operations researchers solving real-world problems is designing and implementing high-quality heuristics to guide their search procedures. Machine learning techniques are increasingly playing a role in operations research approaches, especially in terms of guiding branching and pruning decisions (see [Lodi and Zarpellon 2017] and [Dilkina et al. 2017]).

We integrate deep neural networks into a heuristic tree search procedure and call our approach Deep Learning assisted heuristic Tree Search (DLTS). We apply it to a well-known problem from the container terminals literature, the container pre-marshalling problem (CPMP). Our approach consists of a *policy* network that makes branching decisions and a *value* network that predicts a lower bound for given a node in the search tree. The approach learns good networks using examples from solved CPMP instances.

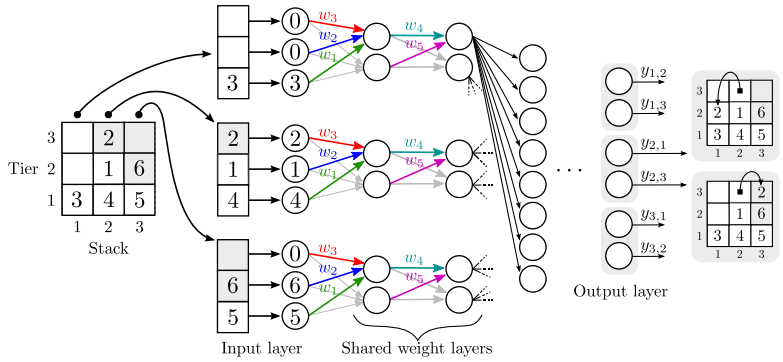


Figure 1. The policy network for the CPMP.

Figure 1 shows the basic idea of the policy network as applied to a CPMP problem with three tiers and three stacks. The goal of the CPMP is to find a minimal sequence of moves of containers such that each stack is sorted ascending from the bottom to the top. In this example, the policy network decides which of the containers 3, 2 or 6 should be moved next, and where it should be moved, as seen in the output layer on the right side.

Our approach results in the new state-of-the-art solver for the CPMP despite having very little heuristic information about the CPMP. We evaluate our approach on a large dataset of instances and on several different instance types. DLTS on the CPMP improves the quality of the solutions found over the state-of-the-art by over 4% in the same or less CPU time.

References

Dilkina, B., Khalil, E. B., and Nemhauser, G. L. (2017). Comments on: On learning and branching: a survey. *TOP*, 25(2):242–246.

Lodi, A. and Zarpellon, G. (2017). On learning and branching: a survey. *TOP*, 25(2):207–236.

Exact Mean Computation in Dynamic Time Warping Spaces*

Markus Brill[1], Till Fluschnik[1], Vincent Froese[1], Brijnesh Jain[2], Rolf Niedermeier[1], and David Schultz[2]

[1]Institut für Softwaretechnik und Theoretische Informatik, TU Berlin, Germany,
{brill, till.fluschnik, vincent.froese, rolf.niedermeier}@tu-berlin.de
[2]Distributed Artificial Intelligence Laboratory, TU Berlin, Germany,
{brijnesh-johannes.jain, david.schultz}@dai-labor.de

Time series such as acoustic signals, electrocardiograms, and internet traffic data are time-dependent observations that vary in length and temporal dynamics. Given a sample of time series, to filter out the corresponding variations, one major direction to time series averaging applies *dynamic time warping* (dtw). Time series averaging is often posed as an optimization problem [1–5] (here called DTW-Mean): Let $\mathcal{X} = \big(x^{(1)}, \ldots, x^{(k)}\big)$ be a sample of k time series $x^{(i)}$. A (Fréchet) *mean* in dtw-spaces is any time series z that minimizes the Fréchet function

$$F(z) = \frac{1}{k} \sum_{i=1}^{k} \Big(\mathrm{dtw}\Big(z, x^{(i)}\Big)\Big)^2 ,$$

where $\mathrm{dtw}(x, y)$ denotes the dtw-distance between time series x and y.

We discuss several problematic statements in the literature concerning the computational complexity of exact algorithms for DTW-Mean. We refute (supplying counterexamples) some false claims from the literature and clarify the known state of the art with respect to computing means in dtw-spaces. We develop a dynamic program as an exact algorithm for DTW-Mean. The time complexity of the proposed dynamic program is $O(n^{2k+1} \cdot k2^k)$, where k is the sample size and n is the maximum length of a sample time series. We apply the proposed exact dynamic program on small-scaled problems as a benchmark of how well state-of-the-art heuristics approximate a mean. Our empirical findings reveal that all tested heuristics suffer from relatively poor worst-case solution quality in terms of minimizing the Fréchet function, and the solution quality in general may vary quite a lot. Moreover, a further theoretical contribution is to show that in case of binary time series (both input and mean) there is an exact polynomial-time algorithm for mean computation in dtw-spaces.

*A conference version was accepted for publication at the SIAM International Conference on Data Mining (SDM18).

Continuous Optimization and Machine Learning

Stephen Wright

University of Wisconsin-Madison

Techniques for formulating and solving optimization problems have become central to modern machine learning and data analysis. We give an overview of how these techniques are applied in such areas as classification, structured model recovery, and deep learning. We address in particular the recent interest nonconvex optimization techniques, discussing both the applications and the algorithms that have been proposed and analyzed, including new theory for an approach from the 1980s: Newton/conjugate-gradient.

Algorithms Based on Unions of Nonexpansive Maps

Matthew K. Tam*

Given two closed sets A and B in $X = \mathbb{R}^d$, the *Douglas–Rachford* iterative scheme, starting at a point $x_0 \in \mathbb{R}^d$, is given by

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} \in T(x_n) \quad \text{where} \quad T := \frac{I + R_B R_A}{2},$$

and R_A denotes the *metric reflector* with respect to a set A. The scheme is well-known to converge when both A and B are convex. In 2014, Bauschke & Noll [1] proved the following.

Theorem 1 (Bauschke–Noll [1])**.** *Suppose that A and B are finite union of convex sets. Then the Douglas–Rachford algorithm converges locally around points in $A \cap B$.*

A case of particular interest arises when the set A is a *sparsity constraint* of the form

$$A = \{x :\in \mathbb{R}^d : \|x\|_0 \leq s\}$$

where $s < d$. In particular, such a set is the finite union of subspaces.

In this talk, we consider a framework for the analysis of iterative algorithms which can be described in terms of a structured set-valued operator which encompasses that of Bauschke & Noll [1] for the Douglas–Rachford scheme. More precisely, we consider fixed point schemes based on operators $T : X \rightrightarrows X$ of the following form.

Definition 2 (Union paracontracting)**.** *An operator $T : X \rightrightarrows X$ is said to be* union paracontracting *if there exist a finite index set, I, a collection of single-valued paracontracting operators, $\{T_i\}_{i \in I}$, and an active selector, $\phi : X \rightrightarrows I$, satisfying*

(P1) $\phi(x)$ is non-empty for every $x \in X$.

(P2) ϕ is outer semicontinuous (osc)*.*

such that T can be expressed in the form

$$T(x) := \{T_i(x) : i \in \phi(x)\} \quad \forall x \in X.$$

Note that, for set-valued maps, there are two notions of fixed points which both coincide for single-valued operators; the *fixed point set* denoted $\operatorname{Fix} T := \{x : x \in T(x)\}$, and the *strong fixed point set* denoted $\mathbf{Fix}\, T := \{x : T(x) = \{x\}\}$.

Our main result is the following theorem concerning local convergence around fixed points.

Theorem 3 (T. [2])**.** *Suppose $T : X \rightrightarrows X$ is union paracontracting with $x^* \in \mathbf{Fix}\, T$, and define*

$$r := \sup\left\{\delta > 0 : \phi(x) \subseteq \phi(x^*) \text{ for all } x \in \mathbb{B}(x^*; \delta)\right\}.$$

Then $r > 0$ and, for any $\epsilon \in (0, r)$, it holds that $\|y - x^*\| \leq \|x - x^*\|$ whenever $x \in \mathbb{B}(x^*; \epsilon)$ and $y \in T(x)$. Furthermore, if the initial point x_0 is contained in $\mathbb{B}(x^*; \epsilon)$ and $x_{n+1} \in T(x_n)$ for all $n \in \mathbb{N}$, then the sequence $(x_n)_{n \in \mathbb{N}}$ converges to a point $\overline{x} \in \operatorname{Fix} T \cap \mathbb{B}(x^*; \epsilon)$.

As a concrete application our theorem, we analyse the *forward-backward algorithm* applied to *sparsity constrained minimisation*

$$\min_{x \in X}\{f(x) : \|x\|_0 \leq s\},$$

where $f : X \to \mathbb{R}$ is convex and continuously differentiable with Lipschitz continuous gradient, ∇f, having Lipschitz constant $L > 0$.

*Institut für Numerische und Angewandte Mathematik, Universität Göttingen, 37083 Göttingen, Germany.
E-mail: m.tam@math.uni-goettingen.de

A sequential homotopy method for unconstrained optimization problems

Andreas Potschka

Universität Heidelberg

We consider the problem of finding a local minimum of a twice continuously differentiable function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. The main challenge for efficient numerical solution methods of this problem is the appropriate treatment of nonconvexity, degeneracy, and large scale. We attack these challenges through the successive partial solution of nonlinear homotopies, which aim to drive a proximal-type regularization of f to zero. If one of the homotopies can be solved entirely, we obtain a local minimum of f after a finite number of steps. Otherwise, the sequential homotopy approach generates a sequence of iterates whose accumulation points are stationary points of f. For the numerical solution of each homotopy subproblem, we employ a corrector-free path-following method based on backward step control with inexact solution of the linearized problems with a preconditioned conjugate gradient method. Numerical results on the unconstrained problems of the CUTEst test set indicate that our method performs competitively with a state-of-the art trust-region method in terms of computational speed, but tends to deliver better local optima in terms of the attained objective values. We close the talk with an outlook on extensions to constrained optimization problems and on challenges for the application to machine learning problems.

Solving the Time-Dependent TSP with Machine Learning Guidance

Imke Joormann

TU Braunschweig

In this talk, we consider the time-dependent traveling salesman problem (TDTSP), a generalization of the asymmetric traveling salesman problem (ATSP) to incorporate time-dependent cost functions. Since the traveling times on an arc can change with every minute, the IP formulation of the TDTSP is quite large and cannot be solved easily. We introduce multiple families of cutting planes for the TDTSP as well as different LP-based primal heuristics, a propagation method and a branching rule. We conduct computational experiments to evaluate the effectiveness of our approaches on several randomly generated instances.

The TDTSP has its origin in a real-world application, where, e.g., delivery routes in one city but for different days are planned. This results in lots of instances sharing (partly) the same structure. We explore how machine learning techniques can be used to exploit this structure and be incorporated in the Branch-and-Cut-and-Price solver.

DeepChem: Deep Learning Meets Nonlinear Optimization to Guide Chemical Development

Kathrin Hatz[1], Sadegh Mohammadi[1], and Linus Görlitz[1]

[1]Bayer AG, Crop Science, Research & Development, Monheim, Germany

Workshop on Optimization, Machine Learning, and Data Science
Braunschweig, April 12-13, 2018

Abstract

The guiding question in early crop protection research is how to identify chemicals which on the one-hand are new and economically synthesizable and on the other hand fulfill many requirements as diverse as biological activity in the respective target organism (e.g. in specific fungi or weed), safety for all remaining organisms (incl. humans) and safety for the environment (e.g. fast degradation in soil). In order to deal with the sheer size of the chemical space of all pharmaceutically or agronomically relevant molecules, which is estimated to be around 10^{33}, various in-silico approaches exist to virtually explore the chemical space in a structured way. Recent advances rely on deep learning to generate new and potentially interesting structures. However, these approaches are exclusively focused on representing structure information and are not capable of generating new molecules that fulfill a complete profile of different biological, chemical and safety properties. Our goal is to use deep learning methodologies to learn a representation of chemical structures that maps compounds with similar profiles (e.g. biological activity in the target organism) close in representation space. With that, we obtain guidance in the chemical space towards areas where new, potentially interesting compounds could be located. Mathematically, this task leads to a two-step problem: learning of the optimal embedding of a molecular structure combined with a prediction task mapping the embedding to the desired profile which could be, e.g., the biological activity combined with further requirements. Major challenges of this approach include

- the lack of data for the learning of the prediction task
- finding a proper problem formulation for the combination of the task (hierarchical optimization or multiobjective optimization - and how to balance the two tasks?)
- convergence problems, i.e. choosing an optimization approach that is tailored to the problem formulation, standard approaches often get stuck
- how to efficiently compute derivatives
- how to evaluate and deal with local minima
- how to optimally exploit known sparsity and structures during optimization within a GPU architecture.

First results for finding the optimal embedding for predicting other molecular properties like polarity and lipophilicity of a structure are promising and support the impression that using insights from nonlinear optimization within this deep learning task could significantly advance these approaches.

Distributed Non-Convex Optimization in Power Systems

A. Engelmann[1], Y. Jiang[2], B. Houska[2] and T. Faulwasser[1]

[1]Institute for Automation and Applied Informatics,
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

[2]School of Information Science and Technology,
ShanghaiTech University, Shanghai, China

A core problem in the operation of large-scale energy systems is the reliable computation of near-optimal energy-efficient operation points. These operating points are usually computed solving large-scale NLPs denoted as AC Optimal Power Flow (OPF) problems. We investigate the application of distributed, non-convex optimization algorithms to AC-OPF problems. Many distributed algorithms such as the Alternating Direction of Multipliers Method (ADMM) are tailored to convex optimization problems. Since the AC-OPF problem is inherently and strongly non-convex, there is no guarantee for these algorithms to converge. This limits their applicability for operation of critical infrastructures like power systems.

We demonstrate that the recently proposed Augmented Lagrangian Alternating Direction Inexact Newton (ALADIN), which provides convergence guarantees for non-convex problems, is well-suited for solving AC-OPF problems [1, 2]. A detailed comparison to ADMM reveals that ALADIN converges faster and to a higher level of accuracy compared to ADMM. To this end, we draw upon the IEEE 118 bus and IEEE 300 bus test cases.

References

[1] Alexander Engelmann, Tillmann Mühlpfordt, Yuning Jiang, Boris Houska and Timm Faulwasser. "Distributed AC Optimal Power Flow using ALADIN." IFAC-PapersOnLine **50.1** (2017): 5536-5541.

[2] Boris Houska, Janick Frasch and Moritz Diehl. "An Augmented Lagrangian Based Algorithm for Distributed Nonconvex Optimization." SIAM Journal on Optimization **26.2** (2016): 1101-1127.

Incorporating Prior Knowledge through Relevance Regularization

Christian Etmann, Peter Maass
Center for Industrial Mathematics, University of Bremen

Incorporating prior knowledge is instrumental for finding a good classification model. Inspired by how prior knowledge is incorporated in inverse problems, we introduce a new family of regularization methods called 'relevance regularization', which allows us to incorporate priors for sparsity and image data.

In many inverse problems, given some known function $f : \mathcal{X} \to \mathcal{Y}$ and a (noisy) measurement $y \in \mathcal{Y}$, the goal is to find a suitable $x^* \in \mathcal{X}$ such that $f(x^*) \approx y$. Here, x^* serves as an estimation of the underlying cause of y. The function f typically models e.g. a physical process or a measurement procedure. For example in X-ray computed tomography, a measured sinogram y is obtained from some (not directly measurable) tomography image x with the radon transform f (plus noise). Problems of this type are usually posed as a minimization problem

$$x^* \in \operatorname*{argmin}_{x \in \mathcal{X}} \mathcal{L}(f(x), y) + \mathcal{R}(x) \tag{1}$$

over the space of causes $\mathcal{X}$ (Mueller & Siltanen, 2012). The regularization term $\mathcal{R}(x)$ is often used to incorporate prior information about x:

- If the solution x^* is expected to be sparse, the choice $\mathcal{R}(x) := \lambda\|x\|_1$ promotes sparse solutions.
- If x is an image (often modelled as a function $x \in C^1(\Omega)$ on the rectangular domain Ω), then the *total variation* $TV_p(x) := \int_\Omega \|\nabla x(z)\|_p dz$ is used as a natural image prior $\mathcal{R}(x) = \lambda TV_p(x)$ in many image processing applications such as denoising (Rudin et al., 1992), superresolution (Babacan et al., 2008), inpainting (Getreuer, 2012) and (blind) deconvolution Chan & Wong (1998); Bioucas-Dias et al. (2006).

In classification, a similar approach for the incorporation of prior knowledge is only applied in linear models, e.g. in ℓ_1-regularized logistic regression, where we minimize

$$\frac{1}{n}\left[\sum_{i=1}^{n} \mathcal{L}(s(Wx^{(i)}), y^{(i)})\right] + \lambda\|W\|_1, \tag{2}$$

w.r.t. W for a labelled training dataset $\left\{(x^{(i)}, y^{(i)})\right\}$, where s denotes the softmax function. For a sufficiently large value of λ, the weight matrix W will be sparse, such that the prediction $s(Wx)$ only depends on a few entries of x.
On the other hand, in the more general case of non-linear parametric models (such as convolutional neural networks), this approach generally does not work as intended: If one simply penalizes the 1-norm of the model parameters during training, the parameters will indeed be sparse. This however does in general not imply that the prediction only depends of a few entries of x, which is the actual motivation.

The reason why this works for linear models but may fail for non-linear models is that the parameters W allow for a clear *interpretation* of the *relevance* of the entries of x. In a neural network, the relationship between its input and its output via its parameters is not evident at all, so that a penalty on these parameters is not fruitful. So how can these principles be transferred to non-linear models?
For this, we introduce a *relevance function*

$$\rho : \mathcal{X} \times \mathcal{Y} \times \mathcal{P} \to \mathcal{Z}, \tag{3}$$

which allows for the assessment of the discriminative relevance of individual parts of the input for the classifier. Examples for relevance functions include saliency maps (Simonyan et al., 2013), guided backpropagation (Springenberg et al., 2014) and VisualBackProp (Bojarski et al., 2016). With a suitable regularization function

$$\mathcal{R} : \mathcal{Z} \to \mathbb{R} \tag{4}$$

defined, we can influence the behavior of our model in a desired manner. We call this paradigm *relevance regularization*. We demonstrate these principles both on the classification of mass spectra through sparse relevances as well as on the problem of image classification on the ImageNet dataset using a TV prior.

References

Babacan, S Derin, Molina, Rafael, and Katsaggelos, Aggelos K. Total variation super resolution using a variational approach. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pp. 641–644. IEEE, 2008.

Extended Abstract

"Workshop on Optimization, Machine Learning, and Data Science"
April 12 - 13, 2018

Sandra Keiper (TU Berlin)

There are many applications, where signals fulfill a secondary structure constraint besides sparsity. That is, the nonzero entries of the signal x_0 stem from a finite or discrete alphabet. Those signals appear, for example, in error correcting codes[1] as well as massive Multiple-Input Multiple-Output (MIMO) channel[2] and wideband spectrum sensing.[3] There also exist several examples of applications, where the transmitted data originate from a general finite set $\mathcal{A} \subset \mathbb{R}$ such as in source decoding[4] or radar.[5]

In this talk we will focus on signals with entries from a bounded lattice and show that compressed sensing recovery guarantees for those signals can be improved significantly in some cases. More precisely, we will focus on the following two structural assumptions.

We first consider the general case $\mathcal{A} = \{-L_1, \dots, L_2\}$, $L_1, L_2 \in \mathbb{N}$, and then $\mathcal{A} = \{0, \dots, L\}$, $L \in \mathbb{N}$. Surprisingly, it will turn out that those alphabets exhibit quite different phenomena due to the positioning of the zero within the set. A second key observation is the fact that mainly the boundary elements play a role in the sense of $-L_1$ and L_2 in the case of bipolar finite-valued signals. Note, that the cases $\mathcal{A} = \{0, 1\}$ and $\mathcal{A} = \{-1, 0, 1\}$ are particularly included. We will see that the proposed recovery algorithm will exploit the structure of those signals exceptionally well.

For the reconstruction of finite-valued sparse signal we consider basis purusit with box constraints, i.e.,

$$\min \|x\|_1 \quad \text{subject to} \quad Ax = b \quad \text{and} \quad x \in [-L_1, L_2]^N. \tag{$P_\mathcal{F}$}$$

New null space properties for the recovery of finite-valued k-sparse signals using ($P_\mathcal{F}$), which allow equivalent conditions for unique recoverability of such signals can be introduced.[6] Using those properties one can analytically compute the phase transitions of all versions (adapted to the specific alphabet considered) of ($P_\mathcal{F}$) in the case of a *Gaussian matrix* A, i.e.,

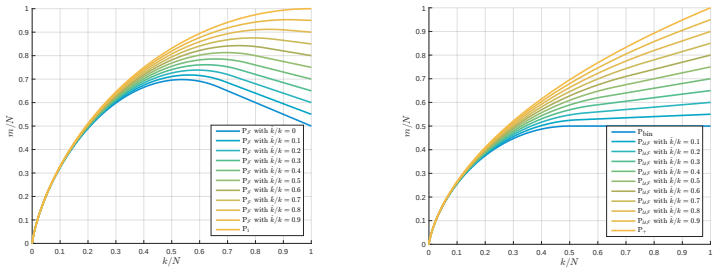
$$A = m^{-1/2} \, [a_{i,j}]_{i,j=1}^{m,N}, \quad \text{with i.i.d.} \quad a_{i,j} \sim \mathcal{N}(0, 1). \tag{1}$$

The content of the main theorem is illustrated in Figure 1.

Up until now, most measurement matrices, e.g. (1) that have been considered were centered, i.e., the expected value of each entry was assumed to be 0. A simple numerical experiment reveals that, when recovering sparse binary signals, i.e., $\mathcal{A} = \{0, 1\}$, using ($P_\mathcal{F}$), this might not be optimal. In Figure 2 we illustrated the results of the numerical experiment to recover a binary signal for different sparsity levels and number of measurements from 0/1-Bernoulli measurements using ($P_\mathcal{F}$). Two observations can be made:

For both the Gaussian and Bernoulli distribution, the numerical experiments indicate that using $m > N/2$ measurements secures recovery with high probability, independent of the sparsity level. Furthermore, in the Bernoulli case, and not in the Gaussian case, the numerical experiments suggest that the recovery of a sparse binary signal is equally probable to the recovery of an *saturated* binary signal, i.e., an signal which has only a few entries equal to zero. Using a probabilistic model, we provide conditions under which the recovery of both sparse and saturated binary signals is very likely. In fact, we also show that under the same condition, the solution of the boxed-constrained basis pursuit program can be found using boxed-constrained least-squares.

The talk is an overview of the results in [Keiper, Kutyniok, Lee, Pfander][6] and in [Flinth, Keiper].[7]



(a) (b)

Figure 1. Phase transition of the convex program ($P_\mathcal{F}$) for bipolar (a) and unipolar (b) signals according to the ratio of $\bar{k} = k - k_{\text{bnd}}$ to k, where is the size of the support and k_{bnd} the number of entries having largest amplitude. Successful recovery is related to the area above the curves.

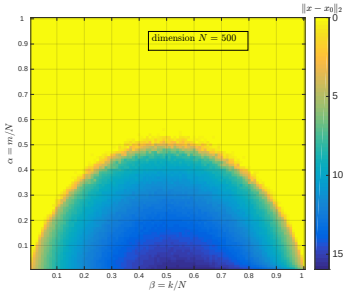


Figure 2. Reconstruction from 0/1-Bernoulli measurements via ($P_\mathcal{F}$). Ground truth $x_0 \in \mathbb{R}^{500}$. We repeated the experiment for each sparsity level and number of measurements 25 times.

Learning Algebraic Varieties from Samples

Bernd Sturmfels

MPI Leipzig

This lecture discusses the role of algebraic geometry in data science. We report on recent work with Paul Breiding, Sara Kalisnik and Madeline Weinstein. The goal is to determine a real algebraic variety from a fixed finite subset of points. Existing methods are studied and new methods are developed. Our focus lies on topological and algebraic features, such as dimension and defining polynomials. All algorithms are tested on a range of datasets and made available in a Julia package.

A primal-dual homotopy algorithm for sparse recovery with infinity norm constraints

Christoph Brauer

We consider the convex optimization problem

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \quad \text{s.t. } \|Ax - b\|_\infty \leq \delta \tag{P_δ}$$

with $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $\delta \geq 0$. By Fenchel-Rockafellar duality [Rockafellar, 1972], it holds that x^* is an optimal solution of (P_δ) if and only if there exists a $y^* \in \mathbb{R}^m$ such that

$$-A^\top y^* \in \partial \|x^*\|_1 \quad \text{and} \quad Ax^* - b \in \delta \partial \|y^*\|_1. \tag{1}$$

If (x^*, y^*) is such an optimal pair, then y^* is an optimal solution to the dual problem of (P_δ). Therefore, we refer to y^* as a dual solution.

In Brauer et al. [2018], we propose to solve a sequence of problems $(\mathrm{P}_{\delta^k})_{k=1,\dots,K}$ with $\delta^0 > \delta^1 > \cdots > \delta^K = \delta$. The underlying motivation is that the transition from an optimal pair (x^k, y^k) to a subsequent optimal pair (x^{k+1}, y^{k+1}) can be much less complex than solving (P_δ) directly. The basic idea behind our method is the following: We start with $(x^0, y^0) = (0, 0)$ which is an optimal pair for (P_{δ^0}) with $\delta^0 = \|b\|_\infty$. Given an optimal pair (x^k, y^k) for (P_{δ^k}), we first fix x^k and δ^k in (1) and determine a $y^{k+1} \neq y^k$ such that the optimality conditions are still satisfied at (x^k, y^{k+1}). Next, we fix y^{k+1} in (1) and seek a $x^{k+1} \neq x^k$ and a $t^{k+1} > 0$ such that the conditions are satisfied at (x^{k+1}, y^{k+1}) and with $\delta^{k+1} = \delta^k - t^{k+1}$.

Although the optimality conditions (1) are non-linear due to the occuring subdifferential, it turns out that they can be reformulated in a linear fashion if either x^* or y^* are fixed. Therefore, we propose to determine y^{k+1} and (x^{k+1}, t^{k+1}) by solving two linear programs, where the objective functions are chosen such that finite termination can be guaranteed. Furthermore, we propose a dedicated active-set strategy that provably works efficiently in this setting [Brauer et al., 2018].

In Brauer [2018], we show that the algorithm terminates after at most $(3^{m+n} + 1)/2$ iterations and consider an example where the algorithm needs exactly $(3^n + 1)/2$ iterations. Moreover, using the example of cross-validation for sparse linear discriminant analysis [Cai and Liu, 2011], we demonstrate that the availability of the entire homotopy path of (P_δ) can be particulary useful in the context of classification tasks. Further applications discussed in Brauer et al. [2018] include sparse dequantization [Brauer et al., 2016], sparse precision matrix estimation [Cai et al., 2011], Chebyshev estimation [Appa and Smith, 1973, Stiefel, 1959] and the Dantzig selector problem [Candès and Tao, 2007].

Attainable Regions of Bio-Chemical Reactions

Nidhi Kaihnsa

MPI Leipzig

In this talk we look at the mass-action kinetics of bio-chemical reactions. We then give a mathematical definition of attainable region. Attainable region is feasible set of an optimisation problem. We characterise this region for linear systems and present a conjecture for weakly reversible systems based on computational experiments. We also discuss a construction due to Cynthia Vinzant which is adapted to describe faces in the convex hull of trajectories.