New Interactions Between Statistics and Optimization

Jonathan Niles-Weed (New York University),
Lnac Chizat (École Polytechnique Fdrale de Lausanne),
Rachel Ward (University of Texas at Austin),
Francis Bach (École Normale Supérieure)

May 22nd–May 27th

1 Introduction to the field

Machine learning is the process by which a computer autonomously extracts knowledge from data. After years of continual progress, machine learning is now able to carry out difficult tasks such as image classification, language translation, or genomic data analysis. It typically works as follows: first a human designs a computer program with many free parameters, and then an algorithm adjusts these parameters so that the program achieves a good performance on training data. The design and analysis of such algorithms is done by researchers in optimization. But the actual goal is to solve the task for new data that does not belong to the training set. Understanding when and why good performance on training data also leads to good performance on unseen data is the task of researchers in statistics.

While optimization and statistics are both relevant to advance machine learning, they are distinct research communities, with different tools and languages. In the past, important breakthroughs have occurred when these two communities have interacted, an example being the introduction of the Stochastic Gradient Descent algorithm, used to train most machine learning programs in todays large scale applications. In this spirit the objective of the BIRS workshop New Interactions Between Statistics and Optimization was to facilitate new interactions and collaborations between the statistics and optimization communities, with the intention of sparking new ideas at the interface of the two fields.

2 Overview of the workshop

The workshop brought in twenty-two experts in statistics, optimization and related applied fields to present the latest developments and explore new directions in the field. There were fourteen talks and one open problem session during the workshop. The talks covered the following themes:

1. Implicit statistical regularization associated to common optimization algorithms
2. Representation costs associated with neural networks architectures
3. Statistical analysis of overfitting methods in high dimension
4. Statistics within optimization

These specialized topics were complemented by a talk by Lydia Liu who invited us to take a broader perspective and consider different statistical settings and models in order to take into account the interactions between algorithms and the environment in which they are deployed. She presented the various work on this topic completed as part of her PhD thesis.

3 Presentation Highlights

3.1 Implicit regularization

Statisticians have traditionally considered optimization as a black box: they study an estimator characterized as the unique minimizer of a well-posed optimization problem, or an approximation thereof. In this viewpoint, the optimization algorithm plays no statistical role and the task of optimization theory is confined to providing, for each problem, the most efficient and reliable algorithms.

However, in modern practice the choice of the algorithm strongly interferes with the statistical properties of the resulting estimator. A classical occurrence of this interference is via the practice of early stopping, i.e. stopping an iterative optimization algorithm before it comes close to a minimizer. Indeed, the *optimization path*, that is the sequence of intermediate estimators generated by an iterative optimization algorithm is specific to each algorithm, and therefore the “early stopped” estimator is algorithm dependent. This phenomenon has been studied for decades under the name *iterative regularization*.

More recently, practitioners have realized that, for supervised learning, simply minimizing (to global optimality) the empirical risk without any regularization sometimes led to state of the art performance in many machine learning tasks, but also that this performance highly depends on the algorithm. This phenomenon is possible because the empirical risk has, in overparameterized settings, an infinity of minimizers with very different statistical properties. The way in which each optimization algorithm selects a specific minimizer is called the *implicit bias* of the algorithm.

Several talks in the workshops have presented the latest advances in the understanding of algorithmic regularization.

Matus Telgarsky: Two implicit bias proof techniques The workshop opened with this topic of implicit bias. Matus started with a review and classification of the various implicit bias analyses that have been proposed for the gradient descent (GD) algorithm for linear models in classification. He in particular proposed a classification into *strong* and *weak* implicit bias guarantees: the former guarantees that the algorithm returns an estimator characterized as being optimal for some auxiliary optimization problem while the latter only attempts at directly providing statistical guarantees, bypassing “optimization” completely. The former is a stronger characterization but the latter is the one that matters from the statistical perspective and is more broadly available.

Loucas Pillaud-Vivien: The role of stochasticity in learning algorithms The talk of Loucas pertained to the difference between the implicit bias of the stochastic gradient descent (SGD) and GD method. While there is no difference in the basic case of linear regression, he presented a setting – the so-called diagonal two-layer linear neural networks – where SGD is provably biased towards sparser solutions than GD.

Nadav Cohen: Generalization in Deep Learning Through the Lens of Implicit Rank Minimization This talk also dealt with GD in a regression setting but for a more complex class of models, starting from matrix factorization, to tensor factorization and concluding with hierarchical tensor factorization. In all these settings, Nadav has shown that the GD algorithm exhibits an implicit bias towards low rank solutions (this is theoretically guaranteed in the early phases of training, and often empirically observed in later phases).

Arthur Jacot: Regimes of Training in Deep Neural Networks: a Loss Landscape Perspective In the same context of deep linear neural networks, Arthur presented an analysis with a different point of view: instead of focusing on the optimization dynamics, he described the loss landscape in which this algorithm navigates. He in particular has shown that depending on the scale of the initialization, the properties of this landscape wildly differ, going from a well-behaved strongly convex landscape for large scales to a landscape full of saddle points for small initializations.

3.2 Representation costs

For classification with the logistic loss, it has been shown in various contexts that GD selects a classifier for which the ratio of the margin over the ℓ_2-norm of the parameters is maximized. This observation naturally leads to the statistical (or approximation theoretical) question of characterizing which kind of functions can be represented by a certain parameterized model when the parameters span a ℓ_2-ball in parameter space. Equivalently, one can ask, given a target function to learn, what is the smallest norm of parameters with which it can be represented? The importance of this question is that it indicate which types of solutions are favored by particular architectures, and therefore clarifies the *inductive bias* of modern machine learning procedures. This line of inquiry has undergone recent progress, much of which was presented at the workshop.

Rebecca Willett: Linear layers in nonlinear interpolating networks Prior work on representation costs has mostly focused on the setting of ReLU (nonlinear) networks with a single hidden layer, or multi-layer networks, all of whose layers are linear. Both settings are quite far from the multi-layer, nonlinear networks used in practice. To begin to bridge this gap, Rebecca presented recent work on the representation costs for multi-layer networks whose first $L - 1$ layers are linear and whose last layer uses a nonlinear ReLU activation. The results reveal that the structure encouraged by this architecture is more complicated than that revealed in prior work on the one-layer or multi-layer linear case, and that this structure includes an interplay between "sparsity-inducing" properties, which encourage simple solutions, and "alignment-inducing" properties, which encourages collapse to low-dimensional subspaces.

Suriya Gunasekar: A convolution property and proof using polynomial representation Most neural networks used for image classification employ convolutional layers with multiple channels (e.g., for inputs which are three-channel RGB images). Even when the number of layers is small, the convolutional structure makes characterizing the representation costs implied by the network theoretically challenging. Suriya presented a key step of a recent proof connecting representation costs in such networks to certain semi-definite programs (SDP). The proof is based on polynomial representations of discrete convolution operators, and is used to show that an auxiliary SDP representation of the network possesses rank-one solutions.

Alberto Bietti: How can kernels help us understand deep architectures? Exactly answering the representation cost question is generally out of reach for more complicated models. A work-around in order to understand the effect of compositionality and of certain computational structure (convolution, pooling, patch extraction, etc) is to study a kernel method with the same structure. This approach was presented by Alberto who has shown how kernel methods which include these kinds of structure where able to better represent functions that are stable under certain groups of transformations.

3.3 Benign overfitting

In classical supervised learning theory, one fixes a learning problem and studies how the prediction error decreases as a function of the number n of observed samples. In this context, methods that interpolate or overfit often lead to a sub-optimal, or even sometimes increasing, error when the observations are noisy. Benign overfitting refers to apparently paradoxical observation that in various settings – which are typically high-dimensional – overfitting leads to predictions that are almost optimal. This phenomenon, which was first discovered empirically, has led to a fruitful line of works that studies the performance of overfitting predictors in high-dimensional regimes.

Ohad Shamir: The Implicit Bias of Benign Overfitting Though originally observed in the context of deep learning, benign overfitting is now recognized to be a more general feature of high-dimensional estimation tasks, even in simple models such as linear regression. The wealth of such examples has led to the informal understanding that overfitting may essentially *always* be benign in sufficiently high-dimensional settings.

Ohad presented results complicating this understanding, and showing that for linear regression and binary classification, overfitting is only benign in some very specific scenarios. In effect, this reveals a type of "implicit bias" for the benign overfitting phenomenon: overfitting implicitly biases certain types of solutions, and the overfitting is benign only if those solutions correspond to the optimal ones.

Theodor Misiakiewicz: Kernels in high-dimension: implicit regularization, benign overfitting and multiple descent The talk of Theodor consisted in a tutorial on the analysis of benign overfitting for high-dimensional kernel ridge regression. He first presented a general heuristic that one can replace the high-dimensional (potentially random) features by Gaussian random variables with matching first two moments, as long as the dimension d and the number of samples n scale polynomially $\log d \asymp \log n$. For this Gaussian model, one can derive the test error rather conveniently and observe various phenomena such as benign overfitting and non-monotonicity of the error. He finally presented various specific contexts where these results have been rigorously proved.

3.4 Statistics within optimization

In most of the talks presented above, although optimization practice inspired statistical theory and vice-versa, there was still a clear distinction between the two fields; with an auxiliary or implicit optimization problem at their interface. In this paragraph, we discuss the works presented during the workshop where this distinction is more blurry.

Nati Srebro: Early Stopping, Regularization, Interpolation and Uniform Convergence Nathi presented the general program through which one can combine optimization and statistics to better understand supervised machine learning. Until now, research works were typically following this line of thoughts: (i) one identifies a complexity measure R over the space of predictors such that, in certain contexts, the optimization algorithm (say, GD or SGD) converges to the minimizer of R over 0-training loss predictors, (ii) one then studies the generalization guarantees of this estimator via uniform convergence. He presented recent theoretical analyses that directly combine these two steps (optimization and statistics) and avoid resorting to uniform convergence. This leads to finer guarantees and is a "weak implicit bias" analysis, in terms of the classification proposed by M. Telgarsky in his talk.

Tomas Vakeviius: A tutorial on offset Rademacher complexity The talk of Tomas was dealing with the fundamental statistical problem of regression without distributional assumption where the goal is to compete with the best linear predictor. In this context, bounding the test loss with the Rademacher complexity is suboptimal, and this can be improved with localized Rademacher complexity for convex hypothesis classes. However, if one wishes to use non-convex hypothesis class, a more recent applicable tool is the offset Rademacher complexity. Tomas presented that analysis of Audibert's star estimator with this tool.

Varun Kanade: Statistical Complexity of Mirror Descent Considering the mirror descent algorithm for least-squares regression, Varun proved that there exists an iteration number for which the estimator satisfies the condition to apply offset Rademacher complexity analysis. This leads to precise statistical bounds on the behavior of mirror descent with the entropy mirror map for sparse linear regression. This analysis is one of the first that captures the effect of early stopping for non-euclidean optimization geometries.

Damien Scieur: Average-case analysis in optimization For the final talk of this workshop, Damien presented a new kind of analysis in optimization that consist in looking at the best *average* run-time of algorithms (instead of their best *worst-case* run-time which is more standard). When specialized to the case of quadratic problems, this idea invokes tools from random matrix theory and orthogonal polynomials theory to derive the optimal algorithm given a distribution over optimization problems. It in particular sheds a new light on Polyak momentum which is shown to be optimal in certain contexts.

3.5 Broader impacts

Statistics and optimization in machine learning are employed in the service of creating large-scale systems with implications for the rest of society. An important question, therefore, is to understand the implications of the techniques and objectives used in machine learning when it is deployed in practice, and how it interacts with broader values and goal.

Lydia Liu: Social Dynamics of Machine Learning for Decision Making An underappreciated aspect of machine learning systems is that they are often used simultaneously across many contexts or by many strategic agents, and that their use is *dynamic* (i.e., agents interact with these systems on multiple occasions over time). As a result, optimality or efficiency guarantees that hold at the level of a single instance may fail to correctly capture the behavior of machine learning systems as actually deployed. Drawing on techniques from microeconomics and algorithmic game theory, Lydia presented several insights from this perspective, with a particular focus on long-term impact. Her work reveals that several strategies designed to address short-term failures of fairness in machine learning systems can provably have negative consequences for long-term welfare.

4 Recent Developments and Open Problems

The open problem session held on the second day of the meeting focused on the following topics.

- Find a class of statistical problems where the choice of a specific architecture matters on the optimization and statistical aspects. Anything that justifies certain choices of architectures and informs practice. Matus mentionned [1] as an example of work in this direction.
- Prove a non-asymptotic and quantitative version of the implicit bias result for wide two-layer neural networks in [2].
- Memory constrained first-order optimization: suppose one F be a convex and 1-Lipschitz function defined . this is an old-standing open question on which recent progress has been made which increased the lower bound [3].
- How small the initialization needs to be to obtain a certain statistical performance? (ϵ-loss, $\alpha(\epsilon)$).

5 Outcome of the Meeting

The Covid pandemic crisis led to a fragmentation of the mathematics, statistics, and computer science research communities, with far fewer opportunities for learning and engagement between disciplines. Moreover, in our experience, online talks and conferences barely solve this porblem and fail to promote genuine interaction. The organizers and participants are grateful to the support from BIRS, which allowed us to hold a vibrant and stimulating meeting.

Another important outcome in the meeting was the interaction between younger and more senior researchers. The workshop had a healthy mix of participants from different career levels, and several junior faculty members present at the workshop expressed gratitude for the opportunity to advertise their work and to have in-depth discussions with senior faculty members. Due to the Covid pandemic, several of the graduate student participants had never had the opportunity to attend an in-person workshop before, and their experience was very positive.

Comments on hybrid format We were fortunate to receive permission from the BIRS staff to have a larger in-person component of our workshop than had been authorized earlier in the pandemic. As a result, all of our speakers were able to present in-person, and a lively open problem session was held in the BIRS auditorium. Though substantially smaller than a typical, pre-Covid BIRS meeting, this critical mass of in-person participants was crucial to sparking the discussions and connections that occurred during the workshop. Moreover, the intimate size of the meeting meant that participants (from early-career students and postdocs to senior faculty) had the opportunity to meet and mingle informally at meals and at social events.

We also had a number of participants who joined virtually via zoom, though their overall number. At this point in the pandemic, we observed that interest in virtual workshops—no matter how well integrated—is waning, and that it was challenging to meaningfully integrate the few participants who elected to participate remotely.

References

[1] I. Safran, G. Vardi, and J.D. Lee, On the Effective Number of Linear Regions in Shallow Univariate ReLU Networks: Convergence Guarantees and Implicit Bias, *to appear in COLT* (2022).

[2] L. Chizat and F. Bach, Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss, *In Conference on Learning Theory*, (2020), 1305–1338.

[3] A. Marsden, V. Sharan, A. Sidford, and G. Valiant, Efficient Convex Optimization Requires Superlinear Memory, *arXiv preprint arXiv:2203.15260* (2022).