Research Statement

Dmitrii M. Ostrovskii

My work is focused on the interplay between statistical learning, numerical optimization, and signal processing. The overarching themes are construction of estimation and prediction procedures with sharp statistical guarantees, their efficient algorithmic implementation, and robustness with respect to unknown structure and parameters. My research interests tend to be influenced by classical results in parametric and nonparametric statistics. Next I review particular areas of my work and envisioned future directions.

Efficient search of approximate Nash equilibria In the past year, I have been working on the problem of the efficient search of first-order Nash equilibria (FNE) in nonconvex minimax problems. More specifically, my focus has been on finding a first-order Nash equilibrium point in problems of the form $\min_{x\in X}\max_{y\in Y} f(x,y)$, where the objective f is nonconvex in x, concave in y, and smooth in both variables, and X, Y is a pair of convex and compact sets. In addition to its practical importance, this problem class is interesting from a theoretical viewpoint. First, it bridges convex and non-convex smooth minimization through the "minimax interface", such that non-trivial interactions between the primal and dual variables may occur; second, it allows to model *non-smooth* non-convex problems. In the recent work [22], we proposed a simple and intuitive algorithmic scheme with state-of-the-art performance guarantee for finding approximate FNE in problems of the above class. An interesting observation is that the resulting first-order complexity estimate, in the classical sense of Nemirovski and Judin ([10]), decomposes as the product of two terms that represent the complexities of the primal (nonconvex) and dual (convex) problems with smoothness parameters modified to account for coupling between the variables. This strongly suggests that the obtained complexity estimate is tight. However, the matching lower bound remains elusive despite the best efforts of our group and several competitors (see, e.g., [13]), and I am looking forward to finally tackling this problem, perhaps together with external collaborators.

Model discrimination with privacy guarantees In the ongoing work with colleagues at University of Southern California (soon to appear on arXiv), we propose a hypothesis testing framework that allows to discriminate between two (or more) parametric models by observing the prediction of the target model θ^* on the datasets generated by the distributions $\mathbb{P}_0, \mathbb{P}_1$ corresponding to each candidate model $\theta \in \{\theta_0, \theta_1\}$. More precisely, we let θ_0, θ_1 be the population risk minimizers over some loss function $\ell(\theta, z)$, corresponding to the population distributions $\mathbb{P}_0, \mathbb{P}_1$ of z; the goal is to test the two hypotheses $H_0 : \theta^* = \theta_0$ and $H_1 : \theta^* = \theta_1$ about the target θ^* by observing the prediction of θ^* on the two datasets generated by sampling from $\mathbb{P}_0$ and $\mathbb{P}_1$. The high-level idea is to compare how well θ^* fits each sample by measuring a certain pointwise functional of the corresponding empirical risk at θ^*, and comparing the two measurements. The most intuitive choice of such functional is simply the value of empirical risk. Somewhat surprisingly, it does not lead to statistically optimal tests. In our work, we identify such functional as the *Newton decrement* of empirical risk, thus discovering the crucial role of second-order information in the model discrimination problem. The resulting testing procedures have rich practical applications in some tasts of privacy-aware machine learning. One such application is arbitrage for "the right to be forgotten", in which a user of a data collecting platform has to verify whether the platform removed their data upon their request, without breaking violating the privacy of that platform.

Fast learning rates via self-concordance In [19], I have investigated the connection of *quasi-self-concordance* of the loss with the availability of fast rates for the excess risk of the corresponding M-estimator. Self-concordance was introduced by [18] in the context of interior-point algorithms; a convex and sufficiently smooth loss is called quasi-self-concordant if its third derivative is upper-bounded with the $3/2$ power of the second. I demonstrated that self-concordance and its extension introduced in [2] in the context of logistic regression, is instrumental in quantifying the generalization properties of the associated M-estimators with random design, and allows to obtain fast rates. Essentially, it allows to

"sew together" the local quadratic approximations of the risk, resulting in similar generalization results as in the case of random-design linear regression. It is remarkable that the obtained results only require *local* assumptions about the loss derivatives at the optimal parameter value – similarly to the classical asymptotic theory. In the work [14], the framework has been extended to ℓ_2-regularized M-estimators. Since these works began circulating and highlighted the role of techniques based on self-concordance, such techniques have proliferated among statistical theorists. Notably, they have been applied to establishing fast rates for *improper estimators*, leading to sharp results (see, e.g., [16, 9]).

Efficient algorithms for large-scale multiclass learning Here we study finite-sum optimization problems arising as the empirical risk objective in linear classification with very large number of classes and dimensionality of the feature space. Our focus is on so-called Fenchel-Young losses [4] that can be represented as the maximum of a finite number of affine functions. This leads to well-structured *bilinear* saddle-point problems, which can be efficiently solved with certain stochastic primal-dual algorithms (see [17]) equipped with ad-hoc variance reduction techniques. Using this approach, in [1] we propose a sublinear algorithm to train multiclass support vector machines – to our best knowledge, the first algorithm of this kind for multiclass linear classification. Extending this result to other losses is a potential direction for further research.

Covariance estimation for heavy-tailed distributions Recently, Wei and Minsker [26] proposed an estimator of the covariance matrix Σ with a remarkable property: its deviations from the target, measured in the spectral norm, are subgaussian under weak moment assumptions on the underlying distribution. For the chosen criterion, this result is near-optimal. On the other hand, in some applications one is interested approximating Σ via $\widehat{\Sigma}$ in the positive-semidefinite sense, i.e., such that $(1-\varepsilon)\Sigma \preccurlyeq \widehat{\Sigma} \preccurlyeq (1+\varepsilon)\Sigma$ for some $0 < \varepsilon < 1$. Such guarantees for the estimator of Wei and Minsker are non-trivial to obtain due to its non-linearity in observations. In the recent work with Alessandro Rudi [23], we have proposed and estimator that admits such guarantees while having essentially the same computational cost as the sample covariance matrix, and considered its applications to noisy principal component analysis and random-design linear regression. Interestingly, the work [15] proposes an alternative approach which allows for even weaker moment assumptions at the expense of the tractability of the resulting estimator.

Non-Euclidean performance estimation In [6], Drori and Teboulle proposed a novel approach for the analysis of the worst-case performance of first-order proximal algorithms that allows to explicitly obtain worst-case problem instances over global complexity classes (such as those of smooth and/or strongly convex functions) for a particular optimization algorithm, as an optimal solution to certain convex program called *performance estimation program*. This could then be used to fine-tune the algorithm, and in some cases, improve over the existing black-box complexity bounds [25]. However, the existing techniques of performance estimation are restricted to Euclidean geometry, as such geometry is required to cast performance estimation programs as semi-definite programs which can then be efficiently solved. Extending performance estimation techniques to non-Euclidean geometries is an interesting open problem.

Structure-adaptive signal recovery Consider estimation of a real- or complex-valued discrete-time *signal* $x := (x_\tau)$, where $-n \le \tau \le n$, from noisy *observations* $y := (y_\tau)$ given by $y_\tau = x_\tau + \sigma\xi_\tau$, where the noise variables ξ_τ are i.i.d. standard Gaussian. More precisely, the goal can be to recover x on the integer points of the whole domain $[-n, n]$, a sub-domain, or in a single point. The classical approach is to assume that x belongs to a known set $\mathcal{X}$ with simple structure (e.g., an ℓ_p-ball). In such cases, estimators with near-optimal statistical performance can be obtained explicitly, and turned out to be linear in observations. Instead, my research has been focused on certain families of *non-linear* estimators with larger applicability and strong theoretical guarantees, applicable when the signal structure is unknown to the statistician. More precisely, such estimators usually take the form of the convolution of y with a time-invariant filter that itself depends on y, and is is obtained as an optimal solution to some convex optimization problem. In the series of papers [7, 24, 21], together with coauthors I have explored the statistical properties of adaptive convolution-type estimators, obtaining finite-sample high-probability oracle inequalities for the ℓ_p-norm error of such estimators, and proving tight lower bounds. In [20], I focused on efficient algorithmic implementation of adaptive convolution-type estimators. I devised first-order proximal algorithms adapted to the special structure of the associated optimization problems.

Currently I am investigating the natural extension of the structure-adaptive signal denoising problem to the case of indirect observations, where the signal is to a linear time-invariant filter before being corrupted with the random noise. Advances in this problem could result in some progress in the classical problem of identification of a linear dynamical system whose output is observed in the random noise (see, e.g., [3, 8]).

References

[1] D. Babichev, D. Ostrovskii, and F. Bach. Efficient primal-dual algorithms for large-scale multiclass classification. *arXiv:1902.03755*, 2019.

[2] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[3] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

[4] M. Blondel, A. F. Martins, and V. Niculae. Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms. *arXiv:1805.09717*, 2018.

[5] C. Butucea and F. Comte. Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli*, 15(1):69–98, 02 2009.

[6] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

[7] Z. Harchaoui, A. Juditsky, A. Nemirovski, and D. Ostrovsky. Adaptive recovery of signals by convex optimization. In *Proceedings of The 28th Conference on Learning Theory (COLT) 2015, Paris, France, July 3-6, 2015*, pages 929–955, 2015.

[8] M. Hardt, T. Ma, and B. Recht. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

[9] R. Jézéquel, P. Gaillard, and A. Rudi. Efficient improper learning for online logistic regression. *arXiv:2003.08109*, 2020.

[10] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28:681–712, 2000.

[11] A. Juditsky and A. Nemirovski. Nonparametric denoising of signals with unknown local structure, I: Oracle inequalities. *Appl. & Comput. Harmon. Anal.*, 27(2):157–179, 2009.

[12] A. Juditsky and A. Nemirovski. Nonparametric denoising of signals with unknown local structure, II: Nonparametric function recovery. *Appl. & Comput. Harmon. Anal.*, 29(3):354–367, 2010.

[13] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. *arXiv:2002.02417*, 2020.

[14] U. Marteau-Ferey, D. M. Ostrovskii, A. Rudi, and F. Bach. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. *To appear on arXiv*, 2019.

[15] S. Mendelson. Approximating the covariance ellipsoid. *arXiv preprint arXiv:1804.05402*, 2018.

[16] J. Mourtada and S. Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *arXiv:1912.10784*, 2019.

[17] Y. Nesterov and A. Nemirovski. On first-order algorithms for ℓ_1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 5 2013.

[18] Y. Nesterov and A. S. Nemirovski. *Interior-point Polynomial Algorithms in Convex Programming*. Society of Industrial and Applied Mathematics, 1994.

[19] D. Ostrovskii and F. Bach. Finite-sample Analysis of M-estimators using Self-concordance. *arXiv:1810.06838*, Oct. 2018.

[20] D. Ostrovskii and Z. Harchaoui. Efficient first-order algorithms for adaptive signal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3946–3955, 2018.

[21] D. Ostrovskii, Z. Harchaoui, A. Juditsky, and A. Nemirovski. Adaptive denoising of signals with shift-invariant structure. *arXiv:1806.04028*, June 2018.

[22] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *arXiv:2002.07919*, 2020.

[23] D. M. Ostrovskii and A. Rudi. Affine invariant covariance estimation for heavy-tailed distributions. *hal-02011464*, Feb 2019.

[24] D. Ostrovsky, Z. Harchaoui, A. Juditsky, and A. Nemirovski. Structure-blind signal recovery. In *Advances in Neural Information Processing Systems*, pages 4817–4825, 2016.

[25] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1):307–345, 2017.

[26] X. Wei and S. Minsker. Estimation of the covariance structure of heavy-tailed distributions. In *Advances in Neural Information Processing Systems*, pages 2859–2868, 2017.