Research Statement

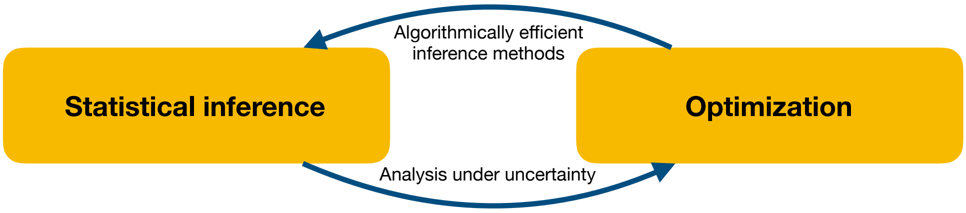
Dmitrii M. Ostrovskii

Over the past few years, I have been active in the areas of statistical learning (broadly construed) and optimization. I will first give a high-level outline some of the questions that spur my scientific curiosity. After that, I will describe the particular directions of my recent work in more detail, present my scientific accomplishments, and outline directions for future work.

Two overarching themes in my research are:

(i) statistical inference methods with **sharp performance guarantees** and efficient implementation;

(ii) construction of **efficient algorithms** for solving large-scale optimization and minimax problems.


Algorithmically efficient inference methods
Statistical inference
Optimization
Analysis under uncertainty

When working in both these directions, I aspire to understand the *fundamental limits* of how well certain statistical or optimization problem can be solved given limited information. For example, in statistical inference one might be interested in finding a sample-based prediction model with the best excess risk over the population distribution [39], or testing a hypothesis about the unknown distribution of the data [17] as effiently as possible in terms of the required sample size to reach a reliable conclusion. In optimization, one might want to approximate an exact solution (an optimal solution in a minimization problem, or a saddle-point in a minimax problem) as fast as possible in terms of the number of search points queried [26]. Establishing such theoretical guarantees usually requires to pass to *classes of problems* in which the specific features of a particular problem instance are abstracted out, yet the class is narrow enough to be mathematically interesting. While such an approach might seem restrictive to a practitioner, there is a deep practical motivation behind it: one focuses on the crucial properties of a particular class of problems, and this leads to procedures which can be broadly applied to a whole variety of problems. Of course, a procedure which is optimal or near-optimal from the theoretical viewpoint might have a very modest performance in specific problems one encounters in practice. As I am interested in the practical implications of my results, I try to avoid such outcomes by testing my theoretical predictions in numerical experiments. I should also mention that I find inspiration in exploring the *connections* between (i) and (ii). For example, optimal statistical procedures often turn out to be given by iterative algorithms. On the other hand, analysis of stochastic algorithms often relies on advanced statistical tools and sharp results.

Another theme of my work is the construction of **adaptive** statistical inference procedures. Oftentimes, it is relatively easy to construct a near-optimal estimator or test assuming the knowledge of certain structural parameter of the problem. Such parameter can be rather simple (e.g., the magnitude of the noise or cardinality of the ground-truth signal), or incapsulate the problem structure (say, it can be a linear subspace to which the signal belongs). After constructing such an "oracle" procedure, one might want to generalize it to the case where such a parameter or structure is unknown, ideally, preserving its favorable statistical properties. Usually, this is done by minimizing a certain data-based "complexity" criterion over the family of candidate procedures corresponding to the possible values of the unknown parameter. The guarantees come in the form of *oracle inequalities* that characterize the price of adaptation by relating the statistical performance of the selected procedure to that of the unavailable "oracle" procedure. The task of constructing adaptive estimators becomes especially interesting in the case of multi-dimensional structural parameters, where classical approaches do not lead to satisfactory results.

Yet another topic of my research, echoing the previous one, is **robustness** of statistical procedures. More specifically, this could mean robustness to *model misspecification*, where one aims to find out how the performance of a statistical procedure degrades when the model assumptions under which it is derived are violated. In a narrower sense, one may want to construct estimators or tests that are robust with respect to "heavy-tailed" observations, i.e., perform well under weak assumptions data-generating distribution.

Next I outline the specific directions of my work in more detail.

Current and past 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. Such problems arise in applications, for example in adversarially-robust statistical learning [21] and fair inference [3]. From a theoretical viewpoint, this problem class is interesting for a few reasons. First, it represents a transition between the classical convex-concave problems and the class of *nonconvex-nonconcave* problems, poorly studied up to date. Second, 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. Finally, nonconvex-concave minimax problems arise as the *max*-formulation of *non-smooth* non-convex minimization. In the recent work [32], I proposed a simple and intuitive algorithmic scheme with state-of-the-art performance guarantee for finding approximate FNE in nonconvex-concave minimax problems. An interesting observation is that the resulting first-order complexity estimate, in the classical sense of Nemirovski and Yudin ([14]), decomposes as the product of two terms that represent the complexities of the primal (nonconvex) and dual (concave) 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., [20]), and I am looking forward to finally tackling this problem, perhaps together with external collaborators.

Model discrimination with privacy guarantees. My recent work [33] introduces 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 violating the privacy of that platform.

Fast learning rates via self-concordance. In [29], I have investigated the connection of *generalized 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 [28] 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 [22], 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., [24, 12]).

Efficient algorithms for large-scale multiclass learning. I studied finite-sum optimization problems arising in the training of linear classifiers with a very large number of classes k, number of features d, and sample size n, via regularized empirical risk minimization. The focus here is on so-called Fenchel-Young losses [5] 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 primal-dual first-order algorithms such as mirror descent or Mirror Prox ([25, 27]) equipped with sampling techniques to reduce the time complexity of an iteration. In the regime of moderate k, existing variance reduction techniques allow for $O(d)$ or $O(d+n)$ complexity of an iteraion by sampling over the training examples ([37, 13, 40, 8]) in combination with sampling over the features [36, 41] which leads to $O(d+n)$ runtime but allows for more flexibility with regards to the problem geometry and better variance control. These complexity estimates are sublinear in the size of the problem input $O(dn)$. However, in the multi-class setup with a very large k, the iteration runtime for these approaches changes to $O(dk)$ or $O(dk+nk)$, and becomes prohibitively large – in fact, *linear* in the combined size of the primal and dual variables. In our work [1], this challenge is addressed through the design of ad-hoc bilevel sampling schemes in combination with a careful choice of proximal geometry. The resulting algorithmic scheme has time complexity $O(d+n+k)$ and favorable guarantees on the accuracy attained after a number of iterations. To the best of our knowledge, this is the first result of this kind for multiclass linear classification. Extending it to other losses, in particular to the multiclass logistic loss, is an interesting direction for further work.

Covariance estimation for heavy-tailed distributions. Recently, Wei and Minsker [45] 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 [34], 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 [23] 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 [9], Drori and Teboulle proposed a novel approach for the analysis of the worst-case performance of first-order proximal algorithms. This approach, called *performance estimation*, allows to explicitly obtain worst-case problem instances in a black-box complexity class (e.g., smooth and/or strongly convex objective) for a *specific* optimization algorithm. The worst-case instance is given as an optimal solution to certain convex program. In recent years, performance estimation techniques have found numerous applications in optimization theory (see, e.g., [19, 15, 43, 7, 42]). However, the existing techniques of performance estimation are restricted to Euclidean geometry, as such geometry is required to cast performance estimation as a semidefinite program (see [44]). Meanwhile, in optimization one often deals with non-Euclidean geometries that better "fit" the model assumptions. In ongoing work with colleagues at Inria Paris, I extend performance estimation to a family of non-Euclidean geometries.

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 [10, 35, 31], 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 [30], 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 random noise (see, e.g., [4, 11]).

Perspective future directions

Geometric statistical signal processing. Many signal processing problems involve data on Riemannian manifolds or graphs. For instance, in computer graphics and vision, 3D objects are modeled as manifolds endowed with properties such as color or texture, or alternatively, as graphs arising when discretizing these manifolds. Other relevant examples include the models of social networks [16], gene expression data, and dynamic models in neuroscience [38], in all of which one has to deal with multiple time-varying processes in the nodes of a large graph whose edges describe correlation between the processes. Exploiting the underlying structure is often vital in these applications, and the techniques of adaptive denoising and deconvolution, after proper generalization, can be capable of inferring this structure.

Post-selection inference. Linear regression is a simple and powerful statistical technique. Not only it allows to estimate the impacts of explanatory variables in the form of regression coefficients, but it also provides confidence intervals for these estimates. However, in modern datasets, the number of candidate variables is often much larger than the sample size, whereas only a small number of them are actually relevant. In these conditions, one would prefer first to select only (supposedly) relevant variables by means of some model selection procedure, and then to regress only on these variables. The problem with this approach is that the usual confidence intervals tend to be too narrow since the inference is now performed on a model which depends on the data and may prove to be wrong with non-vanishing probability. Current quantitative explanations of this phenomenon, see, e.g., [18] and [6], require some stringent assumptions and lack non-asymptotic results, so a lot can potentially be done in this direction.

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] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn. R\'enyi fair inference. *arXiv preprint arXiv:1906.12005*, 2019.

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

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

[6] V. Chernozhukov, C. Hansen, and M. Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7(1):649–688, 2015.

[7] E. De Klerk, F. Glineur, and A. B. Taylor. Worst-case convergence analysis of inexact gradient and newton methods through semidefinite programming performance estimation. *SIAM Journal on Optimization*, 30(3):2053–2082, 2020.

[8] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

[9] 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.

[10] 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.

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

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

[13] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

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

[15] D. Kim and J. A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, pages 1–28, 2020.

[16] D. Lazer et al. Life in the network: the coming age of computational social science. *Science*, 323(5915), 2009.

[17] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer. Springer, 1986.

[18] J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[19] F. Lieder. On the convergence rate of the halpern-iteration. *Optimization Letters*, pages 1–14, 2020.

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

[21] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[22] 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.

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

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

[25] A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[26] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

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

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

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

[30] 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.

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

[32] 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.

[33] D. M. Ostrovskii, M. Ndaoud, A. Javanmard, and M. Razaviyayn. Near-Optimal Model Discrimination with Non-Disclosure. *arXiv:2012.02901*, Dec. 2020.

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

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

[36] B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

[37] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

[38] M. Schwemmer, A. Fairhall, S. Denéve, and E. Shea-Brown. Constructing precisely computing networks with biophysical spiking neurons. *The Journal of Neuroscience*, 35(28):10112–10134, 2015.

[39] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[40] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

[41] Z. Shi, X. Zhang, and Y. Yu. Bregman divergence for stochastic variance reduction: saddle-point and adversarial prediction. In *Advances in Neural Information Processing Systems*, pages 6031–6041, 2017.

[42] A. Taylor and F. Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *COLT 2019-Conference on Learning Theory*, 2019.

[43] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017.

[44] 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.

[45] 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.