# Research statement

Dmitrii M. Ostrovskii

My work is focused on the interplay between numerical optimization, statistical learning, and signal processing. My research interests tend to be influenced by classical results in parametric and nonparametric statistics. In what follows, I review particular areas of my current work and envisioned future directions.

## Current directions

**Fast learning rates via self-concordance** In [17], I have investigated the connection of *generalized self-concordance* of the loss in connection with fast rates for the excess risk of the corresponding $M$-estimators. Self-concordance was introduced by [16] in the context of interior-point algorithms; a convex, and sufficiently smooth, loss is called 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 assymptotic theory. In our recent work [14], the framework has been extended to $\ell_2$-regularized $M$-estimators.

**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 $\xi_\tau$ are i.i.d. standard Gaussian, real or complex. More precisely, the goal can be to recover $x$ on the integer points of the whole domain $[-n, n]$ or some subdomain such as $[0, n]$. The classical approach to signal denoising to assume that $x$ comes from a known set $\mathcal{X}$ with a simple structure (such as an ellipsoid or a rectangle in $\ell_2$) that can be exploited to construct the estimator; In all these cases, estimators with near-optimal statistical performance can be computed explicitly, and correspond to linear functionals of $y$ – hence the name *linear estimators*.

My research has been focused on certain families of *non-linear* estimators with larger applicability and strong theoretical guarantees, applicable when the structure of the signal is unknown beforehand. Assuming for convenience that one must estimate $x_t$ only on $[0, n]$, these estimators can be expressed as

$$\widehat{x}_t^{\varphi} = [\varphi * y]_t := \sum_{\tau} \varphi_\tau y_{t-\tau} \quad 0 \le t \le n, \tag{1}$$

here summation is over $\mathbb{Z}$ taking into account the boundaries; $*$ is the (non-circular) discrete convolution, and the *filter* $\varphi$ is supported on $[0, n]$ which we write as $\varphi \in \mathbb{C}_n(\mathbb{Z})$. Non-linearity of the estimator is due to the fact that the filter is obtained as an optimal solution to some convex optimization problem. Such optimization problems rest upon a common principle – minimization of the Fourier-domain $\ell_p$-norm residual $\|F_n[y - \varphi * y]\|_p$, regularized with the Fourier-domain $\ell_1$-norm $\|F_n[\varphi]\|_1$ of the filter.

In the series of papers [8, 21, 19], I have explored, together with coauthors, the general statistical properties of adaptive convolution-type estimators. In particular, in [8] we studied $\ell_\infty$-fit estimators, demonstrating that such estimators allow to adapt to the unknown best linear filter – "linear oracle" – under the *recoverability* assumption, first introduced in [10], which essentially states that the $\ell_2$-norm of the oracle is much smaller than the sample size. These guarantees were stated in the form of finite-sample high-probability oracle inequalities for the pointwise and $\ell_2$-norm error.

In [21] and [19], I have studied $\ell_2$-fit estimators, showing that they enjoy better adaptation properties than the $\ell_\infty$-fit ones. In particular, oracle inequalities in the case of $\ell_2$ fit can be made *sharp*, i.e. hold with the unit leading constant, under the *approximate shift-invariance* asssumption that states that the

extension of $x$ to $\mathbb{Z}$ belongs to arbitrary, and unknown, shift-invariant linear subspace $\mathcal{S}$ of $\mathbb{C}_\infty(\mathbb{Z})$ with small dimension, or is (locally) close to such a subspace in $\ell_p$-norm. From [10, 11] it has been known that the recoverability assumption is implied by the (exact) shift-invariance assumption. However, the known bounds for the $\ell_2$-norm scaled exponentially with the subspace dimension $\dim(\mathcal{S})$. In [21], we proved a polynomial in $\dim(\mathcal{S})$ upper bound on the oracle norm, with the lower bound of $O(\sqrt{\dim(\mathcal{S})})$. This result was improved in [19]; in particular, the gap was closed for the special class of bilateral filters.

In [18], I have studied the question of efficient algorithmic implementation of adaptive filtering estimators. I have devised first-order proximal algorithms adapted to the special structure of the associated optimization problems. First-order algorithms have a special appeal in these problems, since in this case evaluation of the gradient can be efficiently implemented via the Fast Fourier transform. In [18], I advocated two particularly suitable families of algorithms for the computation of adaptive convolution-type estimators, based on Nesterov's Accelerated Gradient and Mirror Prox algorithms [15]. Besides, I have rigorously established the "statistical complexity" of the proposed algorithms – the number of iterations sufficient to match the statistical performance of the precise estimator.

Currently, I am investigating the natural extension of the structure-adaptive denoising problem to the case of indirect observations of the form $y_\tau = [a * x]_\tau + \sigma\xi_\tau$. Here $a \in \mathbb{C}_m(\mathbb{Z})$ is a given *observation filter*, and the goal is still to recover $x$, i.e., perform deconvolution from noisy observations. Compared to the case of direct observations, statistical assumptions of existing methods are often too restrictive [5], and even some of the basic questions are beyond their grasp. As such, it would be interesting to extend the adaptive filtering techniques to this more general scenario. Potentially, this could lead to some progress in the classical problem of identification of a linear dynamical system observing its output in the noise [3, 9].

**Efficient algorithms for large-scale multiclass learning** Together with coauthors, I 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 [15]) 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 [24] proposed an estimator $\widehat{\Sigma}_{\mathrm{WM}}$ of the covariance matrix $\Sigma$ with a remarkable property: its deviations from the target, measured in the spectral norm, are subgaussian under extremely 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 in relative error bounds, i.e., approximating $\Sigma$ via $\widehat{\Sigma}$, up to a multiplicative factor close to one, in the positive-semidefinite sense. Such guarantees for the estimator $\widehat{\Sigma}_{\mathrm{WM}}$ are non-trivial to obtain due to its non-linearity in observations. In the recent work [20], I 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.

# 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 [12], gene expression data, and dynamic models in neuroscience [22], 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., [13] and [6], require some stringent assumptions and lack non-asymptotic results, so a lot can potentially be done in this direction.

**Non-Euclidean performance estimation** In [7], 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 [23]. 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.

# References


[1] D. Babichev, D. M. Ostrovskii, and F. Bach. Efficient primal-dual algorithms for large-scale multiclass classification. *To appear on arXiv*, Feb 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] 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] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

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

[15] Y. Nesterov and A. Nemirovski. On first-order algorithms for $\ell_1$/nuclear norm minimization. *Acta Numerica*, 22:509–575, 5 2013.

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

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

[18] 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.

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

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

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

[22] 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.

[23] 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.

[24] 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.