Research statement

Dmitrii M. Ostrovskii

December 8, 2018

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 nonparametric and parametric statistics. Below I review particular areas of my current work, as well as envisioned future directions.

Current directions

Structure-adaptive signal denoising. 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, 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 $[0, n]$. In its general formulation, this *signal denoising* problem is classical in statistical estimation and signal processing communities [13, 22, 12, 32, 35]. The conventional approach is to assume that x comes from a known set $\mathcal{X}$ with a simple structure that can be exploited to construct the estimator. For example, one might consider signals belonging to linear subspaces § of signals whose spectral representation, as given by the Discrete Fourier or Discrete Wavelet transform, comes from a linearly transformed ℓ_p-ball [32, 14]. 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

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

here summation is over $\mathbb{Z}$ taking into account the boundaries; $*$ is the (non-circular) discrete convolution, and the *filter* φ 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 ℓ_p-norm residual $\|F_n[y - \varphi * y]\|_p$. regularized with the ℓ_1-norm $\|F_n[\varphi]\|_1$ of the Discrete Fourier transform of the filter. Conceptually, this is similar to *sparse recovery* procedures that learn decomposition of the signal in some fixed (overcomplete) dictionary, with φ in the role of the unknown parameter vector. Indeed, there is a formal connection with sparse harmonic recovery [3, 30], where the dictionary corresponds to all possible harmonic oscillations; I explored this connection, together with coauthors, in [10]. However, the analogy is not full, since

the matrix representing the dictionary is replaced with the convolution map which itself depends on the observations. As a result, adaptive convolution-type estimators are more powerful, from both the theoretical and practical perspectives, than the traditional denoising methods based on sparse recovery approaches such as Lasso or Dantzig Selector, allowing to circumvent the common *frequency separation* requirement as investigated in [10, 28, 27].

In the series of papers [10, 28, 27], I have explored, together with coauthors, the general statistical properties of adaptive convolution-type estimators. In particular, in [10] we studied ℓ_∞-fit estimators, demonstrating that such estimators allow to adapt to the unknown best linear filter – "linear oracle" – under the *recoverability* assumption (first introduced in [15]), which essentially states that the ℓ_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 ℓ_2-norm error.

In [28] and [27], I have studied ℓ_2-fit estimators, showing that they have better adaptation properties than the ℓ_∞-fit ones. In particular, oracle inequalities in the case of ℓ_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 ℓ_p-norm. From [15, 16] it has been known that the recoverability assumption is implied by the (exact) shift-invariance assumption. However, the known bounds for the ℓ_2-norm scaled exponentially with the subspace dimension $\dim(\mathcal{S})$. In [28], 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 [27]; in particular, the gap was closed for the special class of bilateral filters.

In [26], I have studied the question of efficient algorithmic implementation of adaptive filtering estimators. I have devised first-order proximal algorithms that take into account the geometric structure and statistical nature of the associated optimization problems. First-order algorithms have a special appeal in these problems, since computation of the gradient – usually the bottleneck of the overall computation – in this case can be reduced to convolutions, and implemented via the Fast Fourier transform. In [26], I advocated two particularly suitable methods for the computation of adaptive convolution-type estimators: one based on Nesterov's accelerated gradient algorithm [23] and best suited for ℓ_2-fit estimators, and another, based on the Mirror Prox algorithm, see [18], and best suited for the ℓ_∞-fit estimators. 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.

Structure-adaptive deconvolution. 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 [2, 11].

Fast rates in statistical learning. An important question in statistical learning theory is that of *fast rates*, see [33] and references therein. Whereas in the general setting the excess risk can be minimized at the rate $O(1/\sqrt{n})$, if some regularity conditions are imposed, one can obtain the faster rate $O(d/n)$ which corresponds to the asymptotic rate according to the

central limit theorem; here d is some measure of the dimensionality of the problem. The strongest such condition, strong convexity, rarely holds in practice. Commonly, one imposes weaker conditions on the loss, historically originating in online prediction theory, such as *exp-concavity* [21] and *mixability* [34]. In particular, mixability has recently been applied in [9] to construct an (improper) learning algorithm for logistic regression that achieves the $O(d/n)$ rate.

In [25], I have investigated another condition, *generalized self-concordance* of the loss, in connection with fast rates. Self-concordance was introduced by [24] 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. In [25], I showed that self-concordance, and its "irregular" version introduced in [1] 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. Together with colleagues, I am currently working on the extension of the theory to the non-parametric setup.

Efficient primal-dual algorithms for large-scale finite-sum optimization. In this line of work, I investigate finite-sum optimization problems arising in empirical risk minimization, with the goal of developing efficient proximal algorithms that can exploit inherent structure of the data. My 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 based on Mirror Descent and Mirror Prox (see [17, 18]), equipped with ad-hoc variance reduction techniques.

Future directions

Geometric statistical signal processing. Many signal processing problems involve data on non-Euclidean domains, such as 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 as triangulations of these manifolds. Other relevant examples include the models of social networks [19], gene expression data, and dynamic models in neuroscience [29], in all of which one has to deal with multiple time-varying processes in the nodes of a large graph, the edges of which govern the correlation between the processes. Exploiting the underlying low-rank structure is often vital in these applications, and our techniques, after a 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., [20] 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 [8], Drori and Teboulle proposed a novel approach of analyzing 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 complexity bounds from the black box complexity theory (in particular, on the level of constant factors) [7], [31]. However, the existing techniques of performance estimation are restricted to Euclidean geometry, as the Euclidean geometry is required to cast performance estimation programs as semi-definite programs which can then be efficiently solved. Extending performance estimation techniques to algorithms and complexity classes with non-Euclidean geometry is an interesting open problem.

References

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

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

[3] B. Bhaskar, G. Tang, and B. Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Trans. Signal Processing*, 61(23):5987–5999, 2013.

[4] M. Blondel, A. F. Martins, and V. Niculae. Learning classifiers with fenchel-young losses: Generalized entropies, margins, and algorithms. *arXiv preprint 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] E. de Klerk, F. Glineur, and A. B. Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7):1185–1199, Oct 2017.

[8] 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.

[9] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic regression: the importance of being improper. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 167–208, 2018.

[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] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 1991.

[13] I. Ibragimov and R. Khasminskii. *Statistical estimation. Asymptotic Theory*, volume 16 of *Applications of Mathematics*. Springer, 1981.

[14] I. Johnstone. *Gaussian estimation: sequence and multiresolution models*. Unpublished manuscript, 2011.

[15] 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.

[16] 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.

[17] A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, I: General purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011.

[18] A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, II: Utilizing problem structure. *Optimization for Machine Learning*, pages 149–183, 2011.

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

[20] 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.

[21] N. A. Mehta. Fast rates with high probability in exp-concave statistical learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1085–1093, 2017.

[22] A. Nemirovski. Topics in non-parametric statistics. *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXVIII-1998*, 28:85, 2000.

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

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

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

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

[27] D. Ostrovskii, Z. Harchaoui, A. Juditsky, and A. Nemirovski. Adaptive Denoising of Signals with Shift-Invariant Structure. *ArXiv e-prints*, page arXiv:1806.04028, June 2018.

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

[29] 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.

[30] G. Tang, B. Bhaskar, and B. Recht. Near minimax line spectral estimation. In *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, pages 1–6. IEEE, 2013.

[31] 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, Jan 2017.

[32] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

[33] T. Van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

[34] V. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

[35] L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.