Research Statement

Dmitrii M. Ostrovskii

October 19, 2025

My work, broadly construed to be in the fields of statistical theory and optimization theory, tends to focus on some challenging problems arising at the interfaces of these two fields with information theory, approximation theory, and across each other. So far, I have managed to either solve or crucially advance the state of the art in several such open problems, working both in collaboration and "solo," that arise in such diverse areas as online learning, robust estimation, statistical estimation under structural constraints, and minimax optimization. Despite this variability of topics, the problems I tend to work on have a certain easily recognizable profile.

- First, they often combine minimalistic formulation with the challenge of getting a *near-optimal solution procedure*. In statistical applications, I am concerned with computational tractability of such procedures.
- Second, such procedures, if ever to be found, might rely on other areas of mathematics. In particular, I have brought tools from optimization theory, information theory, and approximation theory in my work.
- Third, there are similarities in terms of the underlying mathematical structures to appear, and tools to be used. Among these are self-concordant functions, Bregman divergences, localization method, the notions of affine invariance and equivariance, and the spectral properties of interpolation polynomials.

I am convinced that the excessive compartmentalization of modern academia is its deep flaw, particularly detrimental in mathematical research, and its widely accepted necessity is overestimated. In my work—both its research and teaching aspects—I strive to counteract this trend and "cross-pollinate" my areas of interest.

Optimal rates in statistical learning. Some of my results that have received significant traction concern the fundamental problem of *statistical learning*. Here, given some family $\mathcal{P} = \{P_\theta\}$ of probability distributions indexed by a d-dimensional parameter $\theta \in \Theta$, the learner receives a sample of independent observations $z_{1:n} := (z_1, \dots, z_n)$ whose distribution $P^* \in \mathcal{P}$ is unknown, and aims to infer P^* based on the sample. Formally, one would like to construct an *estimator* $\hat{\theta} = \hat{\theta}(z_{1:n}) \in \Theta$ such that the resulting distribution $\hat{P} = P_{\hat{\theta}}$ is as close as possible to the ground truth, e.g., in the sense of minimizing the Kullback-Leibler divergence of $\hat{P}$ from P^*. Classical statistical theory provides a reasonable universal choice of $\hat{\theta}$ is the *maximum likelihood estimator* (MLE), which amounts to maximizing the probability of observing the sample at hand; the most common example here is linear regression model, corresponding to the Gaussian family of distributions, with $\hat{\theta}$ reducing to the least-squares estimator.[1] Another tenet of this theory is the limiting rate d/n for the performance of MLE compared to the ground truth, as measured by the KL-divergence and other commonly used criteria, in the *asymptotic* large-sample regime $n \to \infty$ with fixed parameter dimension.

Meanwhile, in modern statistical theory, it is of interest to obtain *finite-sample* counterparts of such asymptotic results, showing the $O(d/n)$ rate as soon as n gets larger than some critical threshold polynomial–ideally, linear—in d. While readily available for linear regression models, such results are hard to come by in more general cases. In the paper [13], cited 75 times, I have managed to show the optimal $O(d/n)$ rate of excess risk for logistic regression, with an $O(d)$ critical sample size. This crucially improved over the earlier work of F. Bach [1], breaking the $O(d^2)$ critical sample size threshold. The key idea that led to this result was to connect the notion of *self-concordant* functions, coming from the theory of interior point methods in optimization [9], with the MLE localization process. These ideas and techniques, as well as those from the companion paper [7], have proliferated in the community and led to further progress; e.g., see [8, 4, 6, 2].

Estimation under structural constraints. My PhD work [3, 18, 11, 10] has been devoted to the problem of *estimation under shift-invariance*, introduced by Arkadi Nemirovski in the 1990s. In this problem, the parametric family $\mathcal{P}$ corresponds to the *unknown* subspace $\mathcal{S} \subset \mathbb{R}^d$ with dimension $s \ll d$, invariant under

[1] I am simplifying things to the point of vulgarity; this is inevitable given the format of this document and the intended scope.

the action of the shift operator. Such sets are ubiquitous in statistics, both in theory and in practice, and admit alternative description as solution sets of linear difference equations, that is exponential polynomials. It has long been known that one can build "adaptive" estimators, attaining the excess risk $\mathrm{poly}(s, \log d)n^{-1}$ without any information on $\mathcal{S}$ besides its dimension. In the recent work [12], I returned to this problem and succeeded in showing the near-optimal excess risk $s\,\mathrm{poly}(\log d)n^{-1}$. The crucial point here is that the rate is linear in s, same as if $\mathcal{S}$ were *known*. The key step leading to this improvement is a construction based on Lagrange interpolation of the estimated support polynomial, combined with Fourier-analytical techniques. In fact, the theory rests upon an approximation-theoretic result of independent interest, concerning the existence of compactly supported reproducing kernels with simultaneously sharp estimates on their spectral ℓ_p-norms.

One structural property that one might seek to impose upon an estimation procedure is its *equivariance* with respect to a group of transformations of the parameter, meaning that the estimate $\hat{\theta} = \hat{\theta}(z_{1:n})$ must transform in the same way as the true θ^* under the group action. Arguably, the most important special case is *affine equivariance*, where the group is $\mathrm{Aff}(\mathbb{R}^d)$, corresponding to arbitrary changes of basis and translations. While most classical estimation procedures have this property, modern theory spawned a plethora of those that lack it; in particular, such are estimators performing certain kind of truncation, thresholding, or pruning of observations based on their magnitude and/or closeness to a given point. Meanwhile, statistical optimality and affine equivariance go hand-in-hand: the former is known to imply the latter in the case of "unconstrained" estimation (i.e., with $\Theta = \mathbb{R}^d$). As such, it makes sense to have a procedure that, given a "reasonable" but non-affine-equivariant estimation procedure, converts it into an affine equivariant one, and preferably, in a tractable way. In a joint work with A. Rudi [16], I addressed this task in the scenario of covariance estimation of a heavy-tailed distribution. In a nutshell, we proposed a method of converting a "fixed-basis" estimator into a nearly affine-equivariant one, with the computational burden comparable to that of the initial estimator. Extending the method to other situations, e.g., classical compact groups such as $\mathrm{SO}(d)$, is an open challenge.·

Optimal and tractable online optimization. Online optimization is a framework for *sequential, adversarial* decision-making, formalized as a game played between the learner and the adversary. This framework can be contrasted with the statistical learning one: at each round $t \in \{1, \dots, n\}$, adversary generates new "observation" $z_t \in \mathcal{Z}$ *after* receving the decision $\theta_t \in \Theta$ of the learner; similarly, the learner selects the next θ_{t+1} only after observing z_t. The central object for both players is the *regret* against the best decision in hindsight,

$$\mathscr{R}_n(\theta_{1:n}|z_{1:n}) = \sum_{t\leqslant n} \ell(\theta_t, z_t) \;-\; \min_{\theta\in\Theta} \sum_{t\leqslant n} \ell(\theta, z_t),$$

with the respect to a given *loss function* $\ell(\cdot,\cdot)$ on $\Theta \times \mathcal{Z}$. Specifically, the learner seeks to minimize $\mathscr{R}_n$ in $\theta_{1:n}$, and adversary to maximize it in $z_{1:n}$, under the causal constraints above. The resulting optimal value, called the *minimax regret*, characterizes the information-theoretic complexity of a specific online learning problem.

My joint work [5] with R. Jézéquel and P. Gaillard addressed one of the most "iconic" online learning problems, *online portfolio selection*, corresponding to logarithmic losses $\ell(\theta, z) = -\log(\theta^\top z)$ and the positive orthant $\mathbb{R}^d_+$ in the role of both $\mathcal{Z}$ and Θ. It was introduced in the 1990s by Thomas M. Cover as a model for trading in adversarial environment, with guaranteed performance in the scenario of a market crash. Cover's proposed algorithm, Universal Portfolios, attaining the minimax regret of order $O(d \log n)$, yet challenging from the computational standpoint, with the fastest implementation running in $O(d^4 n^{14})$ per round. In [5], we proposed an alternative algorithm with the same regret guarantee, but running in only $O(d^2 n)$ per round. The proof of regret optimality, as well as the main idea of the construction, is based upon some notions from optimization theory, in particular, interior-point methods and self-concordant functions; on the other hand, we exploit information-theoretic tools to interpret the algorithm as an approximation of Universal Portfolios. The problem of finding a "realistic" alternative to Universal Portfolios with the matching regret was widely known in the learning theory community, with the open problem call [20] at COLT 2020, the top conference in the area; as the result, the work enjoyed some attention upon circulation (e.g., here). Subsequently, it spawned a follow-up work [19] from the colleagues at National Taiwan University, who generalized our algorithm to obtain the fastest-to-date regret-optimal algorithm for *online quantum tomography* with the logarithmic loss.

Amplitude maximization, and beyond. Recently, my interests expanded towards mathematics unrelated to data science – notably, extremal problems concerning exponential polynomials. In particular, in [17] I

solved a problem whose statement can be understood by an undergraduate student who took a first class on differential equations. Namely, consider the class of linear recurrent relations of order n, with characteristic roots of magnitude $r < 1$, and allow the initials $x_0, x_1, \dots, x_{n-1}$ to vary over the unit disk (or in a polydisk).

Given $t \geqslant n$, what configuration of the roots and initials gives the maximum possible amplitude $|x_t|$?

For the reader unconcerned with linear recurrences or difference equations, we note that the natural continuous-time counterpart of this problem is to maximize the magnitude of a solution to *unspecified* ordinatry differential equation (ODE) with characteristic roots in a vertical stripe, and bounded initial data, at a given time instant. As it turns out, for any triple t, n, r, the maximum is attained in the resonant configuration, i.e. with coinciding roots and cophase initials. This is striking, due to the fact that the statement holds for arbitrary t, despite the mismatch of phases across the fundamental solutions corresponding to different roots if these are present. It is also remarkable that our proof crucially relies on the notion of Schur positivity, coming from the theory of symmetric functions and partitions, and that this fact—along with its rich practical implications, e.g., in control theory, power grids, and motion planning—was overlooked by combinatorialists working in the area.

In the past couple of months, I have started to work on a related problem at the intersection of analysis and algebraic combinatorics, in a collaboration with Vinayak Dutta, a Master student from Cornell. We aim to find a complex-variable extension of the result of Khare and Tao, who showed that for any pair of partitions $\lambda \supseteq \mu$, the ratio of Schur polynomials S_λ / S_μ is increasing in each variable on the positive orthant.

Electric power networks. Since the past spring, I have been collaborating with the power network group at ECE led by Daniel Molzahn, in particular with Sam Talkington, a graduate student whom I have been effectively co-advising. Sam contacted me while attending my special topics class last Spring, and we started a project on minimizing the power losses in electrical grids via randomized first-order optimization algorithms. This resulted in a recent submission to AISTATS, a leading conference in statistics and data science, and another one at EPSR, the top venue for electrical power systems research; both are about to appear on arXiv.

Minimax optimization. During the first two years of my postdoc at USC, I have worked on minimax optimization problems, in particular those beyond the classical case of convex-concave saddle-point problems. *Minimax problems*, or *zero-sum games*, are central in optimization and game theory; these are of the form

$$\min_{x \in X} \max_{y \in Y} f(x, y),$$

where X, Y is a pair of convex sets; $f : X \times Y \to \mathbb{R}$ is the objective function from a certain class $\mathcal{F}$, specifying a problem instance. In classical optimization, one studies convex-concave problems, assuming that functions in $\mathcal{F}$ are convex in x, concave in y, and smooth in both variables. However, this framework falls short of capturing some modern applications, especially those arising in data-intensive domains (e.g., in deep learning). This motivated the study of minimax problems with *nonconvex-concave* and *nonconvex-nonconcave* objectives, with the adapted goals of exibiting either a locally optimal point or a first-order stationary point (FOSP).

In [15], we proposed a state-of-the-art algorithm for exhibiting an FOSP in the nonconvex-concave scenario. The resulting convergence rate is widely conjectured to be optimal, but proving this is a known open problem in this community. This work has enjoyed sustained attention, with 126 citations reported on Google Scholar.

In [14], I studied the more challenging *nonconvex-nonconcave scenario*. Such problems are known to be intractable in general, essentially since one cannot even evaluate the external objective $\varphi(x) = \max_{y \in Y} f(x, y)$. Positive results are known under a plethora of assumptions; however, most of these are either opaque or too restrictive, and tend to trivialize the problem. We introduced, arguably, the simplest and most natural simplifying assumption: small diameter $D = \mathrm{diam}(Y)$ of the maximization set, and proposed the natural approximation scheme, in which $f(x, y)$ is approximated with its k^{th}-order Taylor expansion $f_k(x, y)$ in y. As it turns out, for given target accuracy ε and approximation order k, there is a critical diameter $D^* = D^*(\varepsilon, k)$ allowing for this approximation, in the sense that any ε-approximate FOSP of f_k is $O(\varepsilon)$-approximate for f. When $k \leqslant 2$, this approach leads to efficient algorithms. Moreover, we exhibited problem instances showing the sharpness of these bounds in all ranges of parameters; these are given by certain bivariate polynomials.

References

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

[2] H. Chardon, M. Lerasle, and J. Mourtada. Finite-sample performance of the maximum likelihood estimator in logistic regression. *arXiv preprint arXiv:2411.02137*, 2024.

[3] 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.

[4] R. Jézéquel, P. Gaillard, and A. Rudi. Efficient improper learning for online logistic regression. In *Proceedings of the 33rd Conference On Learning Theory*, pages 2085–2108. PMLR, 2020.

[5] R. Jézéquel, D. M. Ostrovskii, and P. Gaillard. Efficient and near-optimal online portfolio selection. *Mathematics of Operations Research*, 2025.

[6] L. Liu, C. Cinelli, and Z. Harchaoui. Orthogonal statistical learning with self-concordant loss. *arXiv preprint arXiv:2205.00350*, 2022.

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

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

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

[10] 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.

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

[12] D. M. Ostrovskii. Near-optimal and tractable estimation under shift-invariance. *arXiv preprint arXiv:2411.03383*, 2024.

[13] D. M. Ostrovskii and F. Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1):326–391, 2021.

[14] D. M. Ostrovskii, B. Barazandeh, and M. Razaviyayn. Nonconvex-nonconcave min-max optimization with a small maximization domain. *arXiv preprint arXiv:2110.03950*, 2021.

[15] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021.

[16] D. M. Ostrovskii and A. Rudi. Affine invariant covariance estimation for heavy-tailed distributions. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99, pages 2531–2550. PMLR, 2019.

[17] D. M. Ostrovskii and P. S. Shcherbakov. Amplitude maximization in stable systems, Schur positivity, and some conjectures on polynomial interpolation. *arXiv preprint arXiv:2508.13554*, 2025.

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

[19] W.-F. Tseng, K.-C. Chen, Z.-H. Xiao, and Y.-H. Li. Online learning quantum states with the logarithmic loss via VB-FTRL. *arXiv preprint arXiv:2311.04237*, 2023.

[20] T. Van Erven, D. Van der Hoeven, W. Kotłowski, and W. M. Koolen. Open problem: Fast and optimal online portfolio selection. In *Proceedings of the 33rd Conference On Learning Theory*, pages 3864–3869. PMLR, 2020.