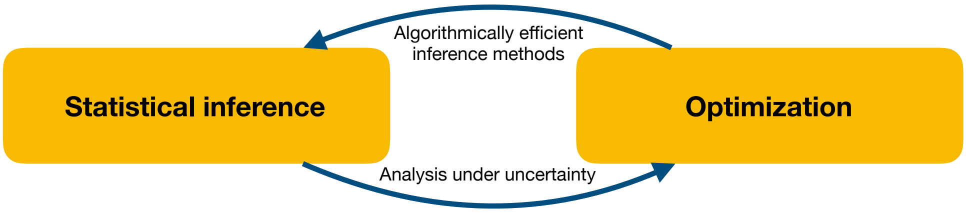
Research Statement

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.

My work tends to proceed in following two directions:

(i) statistical inference methods with **sharp performance guarantees** and efficient implementation;
(ii) construction of **efficient algorithms** for solving large-scale optimization and min-max 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 [52], or testing a hypothesis about the unknown distribution of the data [24] 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 min-max problem) as fast as possible in terms of the number of search points queried [38]. 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 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.

In what follows, I outline the specific directions of my work in more detail.

Nonconvex-(non)concave min-max optimization. In the past couple of years, I spent a sinificant amount of time working on min-max optimization problems beyond the classical convex-concave setup, with a particular focus on the tractability of first-order stationary points. Such problems are of the form

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

where the sets X, Y are convex; the objective f is smooth in both variables, nonconvex in x, and may or may not be concave in y. Such problems arise in many applications in modern machine learning: e.g., in adversarially-robust statistical learning [31], fair inference [4], generalized adversarial networks [18], or reinforcement learning [13]. The class of *nonconvex-concave* min-max problems are interesting from the theoretical viewpoint, as they highlight the challenges of studying the general *nonconvex-nonconcave* setup and give some insight for it. Moreover, nonconvex-concave objectives arise naturally as the maxima over parametrized families of (smooth) nonconvex functions, and thus are of independent interest.

In the recent work [45], I proposed a simple and intuitive algorithmic scheme with state-of-the-art performance guarantee for finding approximate first-order Nash equilibria and stationary points in nonconvex-concave and nonconvex-strongly concave min-max problems. The obtained convergence rates are near-optimal in the strongly concave case [25], and are conjectured to be so in the concave case. I am looking forward to closing this gap. In another recent work [44], I study nonconvex-nonconcave problems under the assumption that the maximization set Y has a small diameter–decreasing as some positive power of the target accuracy level ε. This regime is relevant for applications, in particular in the task of adversarially robust training. Specifically, under the assumption that the functions $f(x,\cdot)$ are k-order regular for each $x \in X$, I derive upper and lower bounds on the critical diameter $\bar{\mathsf{D}}$ such that having $\mathrm{diam}(Y) \le \bar{\mathsf{D}}$ allows to replace the task of searching for an ε-first-order stationary point of $f(x,y)$ with that of searching for an $O(\varepsilon)$-stationary point of its k^{th}-order Taylor approximation in y. This reduction leads to efficient algorithms for finding stationary points in nonconvex-nonconcave problems.

Model discrimination and two-sample testing. Think about a very natural question: given a vector θ^* of coefficients in linear regression and observing two i.i.d. samples $(X^{(0)}, Y^{(0)})$, $(X^{(1)}, Y^{(1)})$ one of which is generated using θ^* and the other one using some unknown θ^*, *what is the sample complexity of recognizing the right sample?* More generally, consider the following problem termed *model discrimination*: given a loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$, strictly convex on the domain $\Theta \subseteq \mathbb{R}^d$ of parameter θ, and the Nature chooses a pair of distributions $\mathbb{P}_0, \mathbb{P}_1$ of observation z with the corresponding population risk minimizers $\theta_0 \neq \theta_1 \in \mathbb{R}^d$, selects $k \in \{0,1\}$. The statistician is then given access to ℓ and to $\theta^* = \theta_k$, but not to the selection of k nor the distributions $\mathbb{P}_0, \mathbb{P}_1$; instead, the distributions are only accessed through the two i.i.d. samples $(z_1^{(0)}, ..., z_n^{(0)}) \sim \mathbb{P}_0^{\otimes n}$ and $(z_1^{(1)}, ..., z_n^{(1)}) \sim \mathbb{P}_1^{\otimes n}$. Given these data, what is the sample complexity of discriminating between the two hypotheses $\mathcal{H}_0 : \theta_0 = \theta^*$ against $\mathcal{H}_1 : \theta_1 = \theta^*$?

The problem described above was suggested and studied jointly with Mohamed Ndaoud in our recent work [46]. In the setup of two random-design linear regressions, we have managed to construct a testing procedure with a nearly minimax-optimal sample complexity. The procedure admits natural extensions beyond the linear regression setup; in particular, this allows to reach asymptotic optimality in the case of a general parametric family, and a sharp finite-sample results for generalized linear models. The framework has applications in statistical fairness and privacy, when tester's goal is to understand on which of two possible distributions a statistical prediction model θ^* has been learned, without full access to the distributions, and only with a *partial* access to the model (corresponding to the test statistic); see [46, Sections 1.1–1.2].

In this direction of research, my next step is to revisit the classical two-sample testing setup, where one has to test whether two i.i.d. samples come from the same distribution or different ones. The existing results due to [5] are restricted to the scenarios of mean regression or isotropic design, and I believe that the insights I gained while working on the model discrimination task will prove to be useful here as well.

Fast learning rates via self-concordance. In [41], 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 [40] 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 [3] 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 [32], 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., [35, 21]).

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 [10] 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 ([36, 39]) 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 ([51, 22, 53, 15]) in combination with sampling over the features [49, 54] 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 [2], 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.

Performance estimation for entropic geometry and beyond. In [17], Drori and Teboulle proposed a new technique for analyzing the worst-case performance of first-order proximal algorithms. The technique, called *performance estimation*, allows to explicitly obtain a worst-case instance in a particular problem class (such as, for example, the class of minimization problems on $\mathbb{R}^d$ with a convex and L-smooth objective) and 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., [26, 23, 56, 14, 55]). However, the existing techniques of performance estimation do not extend beyond the Euclidean case, i.e., require the distances and gradient magnitudes to be measured in ℓ_2-norm; this allows to cast performance estimation as a semidefinite program using the equivalence between the positive-semidefinite and Gram matrices [57].

Meanwhile, in optimization it is often beneficial to work with different geometries that better "fit" the model assumptions, such as ℓ_1-norm or the matrix nuclear norm. Usually, algorithms adapted to such non-Euclidean geometries are formulated and analysed in terms of the Bregman divergence adapted to the norm in question [8]—namely, with a potential function that is strongly convex with respect to the norm in question. Making a step further, the recent works [6, 27] introduced the notions of *relative smoothness and strong convexity*, where the corresponding properties are *defined* directly in terms of a Bregman divergence, bypassing the norm whatsoever. This includes the cases where there exists *no suitable norm*, as the Bregman potential is not strongly convex; however, even for the cases where a norm does exist (such as, e.g., entropy-type geometry on the probability simplex), relative smoothness and strong convexity lead to larger functional classes than those defined in terms of the corresponding norm, which results in gaps between the known upper and lower complexity estimates. Therefore, it is an interesting question whether performance estimation could be formulated in a computationally tractable form in such non-Euclidean situations. In a work with Radu-Alexandru Dragomir (UC Louvain), which has formed a chapter of his PhD thesis [50, Chap. 5], we found a convex programming formulation of performance estimation with entropic geometry. Our next goal is to apply the framework to the question of whether it is possible to accelerate first-order methods "à la Nesterov" [37] for the corresponding functional class. This is an intriguing question: on one hand, Dragomir et al. [16] recently showed that the "unaccelerated" $O(1/T)$ convergence rate cannot be improved uniformly over all Bregman divergences. On the other hand, in their construction a divergence is chosen adversarially; meanwhile, in the case of entropic geometry one could envision using some extra regularity lacking in the general Bregman case [7].

Covariance estimation for heavy-tailed distributions. Recently, Wei and Minsker [58] 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 [47], 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 [33] proposes an alternative approach which allows for even weaker moment assumptions at the expense of the tractability of the resulting estimator.

Efficient methods for robust covariance estimation, linear regression, and M-estimation. Theory of robust statistical estimation and prediction has experienced a renaissance in the past ten years, ignited by the seminal works of Catoni and Audibert on robust least-squares [1] and mean estimation [11]; this has been followed by efficiently implementable truncation-based methods with sharper—but still suboptimal—statistical guarantees ([58, 47]) and by statistically optimal methods but extremely computationally-heavy methods based on the "tournament" technique of Lugosi and Mendelson ([30, 29, 28]). Yet, marrying robustness with computational efficiency remains challenging: in fact, computationally efficient estimators with subgaussian confidence bounds under weak moment assumptions are only known for the basic problem of mean estimation ([12]). In particular, the task of obtaining a computationally efficient covariance estimator with subgaussian deviations under the $L_4 - L_2$ moment equivalence condition [34] remains open, and its solution would result in efficient robust algorithms for random-design linear regression and M-estimation with self-concordant losses.

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 [19, 48, 43], 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 [42], 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., [9, 20]).

References

[1] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Annals of Statistics*, 39(5):2766–2794, 2011.

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

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

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

[5] Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, pages 577–606, 2002.

[6] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[7] H. H. Bauschke and J. M. Borwein. Joint and separate convexity of the bregman distance. In *Studies in Computational Mathematics*, volume 8, pages 23–36. Elsevier, 2001.

[8] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. SIAM, 2001.

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

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

[11] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[12] Y. Cherapanamjeri, S. B. Hopkins, T. Kathuria, P. Raghavendra, and N. Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 601–609, 2020.

[13] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR, 2018.

[14] 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.

[15] 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.

[16] R.-A. Dragomir, A. B. Taylor, A. d'Aspremont, and J. Bolte. Optimal complexity and certification of bregman first-order methods. *Mathematical Programming*, pages 1–43, 2021.

[17] 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.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[19] 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.

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

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

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

[23] 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.

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

[25] H. Li, Y. Tian, J. Zhang, and A. Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *arXiv preprint arXiv:2104.08708*, 2021.

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

[27] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[28] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*, 2016.

[29] G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[30] G. Lugosi, S. Mendelson, et al. Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794, 2019.

[31] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[32] 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.

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

[34] S. Mendelson and N. Zhivotovskiy. Robust covariance estimation under $l\_4 - l\_2$ norm equivalence. *arXiv preprint arXiv:1809.10462*, 2018.

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

[36] 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.

[37] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[38] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2013.

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

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

[41] D. Ostrovskii and F. Bach. Finite-Sample Analysis of M-estimators using Self-Concordance. *Electronic Journal of Statistics*, 15(1):326–391, 2021.

[42] 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.

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

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

[45] 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, to appear*, 2021.

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

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

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

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

[50] D. Radu-Alexandru. *Bregman gradient methods for relatively-smooth optimization.* PhD thesis, l'Université Toulouse 1 Capitole, 2021.

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

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.