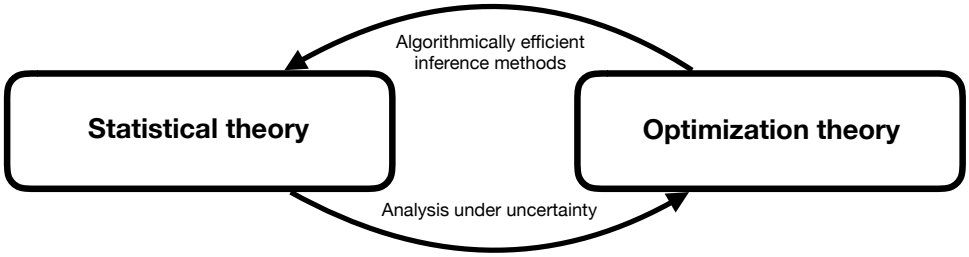
Research Statement

Dmitrii M. Ostrovskii

In the past few years, I have been active in the areas of statistical learning (broadly construed) and optimization theory. More specifically, my work tends to proceed in two following directions:

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

(b) **efficient algorithms** for solving large-scale optimization and min-max problems.



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 [60], or testing a hypothesis about the unknown distribution of the data [32] as effiently as possible in terms of the required sample size to reach a sound 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 [47]. 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 (a) and (b). 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 recurrent theme in my work is 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 interest of mine is **robustness** of statistical inference procedures. This could be 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 seek to construct estimators that are robust against "heavy-tailed" observations, i.e., perform well under weak assumptions about the data-generating distribution.

In what follows, I shall discuss the directions of my work in more detail. For reader's convenience, they are grouped into the two "streams" corresponding to optimization theory and mathematical statistics; as discussed previously, this division is rather speculative in most cases.

1 Mathematical statistics and learning theory

1.1 Fast learning rates with self-concordant losses

In [52], I have investigated the connection of *self-concordance* with the availability of fast rates for the excess risk of the corresponding M-estimator. Self-concordance was introduced by [48] in the context of interior-point algorithms; in a nutshell, a convex and smooth loss function is called self-concordant if its third derivative is upper-bounded in terms of the second. I demonstrated that self-concordance and its extension introduced in [4] 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 distribution tails at the target parameter value, similarly to the classical asymptotic theory. In the work [41], these results have been extended to the nonparametric scenario. The ideas and techniques introduced in these works have already proliferated in the community of statistical theorists, leading to some further progress; see, for example, [44, 27, 35].

1.2 Robust 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 [2] and mean estimation [11]; this has been followed by efficiently implementable truncation-based methods with sharper—but still suboptimal—statistical guarantees ([66, 56]) and by statistically optimal methods but extremely computationally-heavy methods based on the "tournament" technique of Lugosi and Mendelson ([39, 38, 37]). Yet, marrying robustness with tractability 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 [13].

Efficient and tractable robust covariance estimation. One specific task I am interested in the context of tractable robust estimation is that of obtaining a tractable and robust *covariance matrix* estimator under weak moment assumptions, namely the $L_4 - L_2$ moment equivalence condition [43]. This question has remained open for a few years now, after the progress obtained

by S. Hopkins and coauthors in the context of mean estimation. In fact, there are some reasons to believe that such an estimator does *not* exist, but actually proving such a negative result would require a combination of tools from robust statistics and computational complexity theory; as a result, this challenging problem has remained beyond the reach of either of the two communities.

Affine-equivariant robust covariance estimation. One property of a statistical estimator, desired in many applications, is *affine equivariance*: the estimator should preserve its functional form under affine reparametrizations of the data-generating distribution. For covariance estimation, this reduces to requiring that an estimate $\widehat{\Sigma}$ of Σ, the population covariance matrix, satisfies the positive-semidefinite inequality $(1-\varepsilon)\Sigma \preccurlyeq \widehat{\Sigma} \preccurlyeq (1+\varepsilon)\Sigma$ for some $\varepsilon \in (0,1)$, where $A \preccurlyeq B$ means that $B-A$ is positive-semidefinite. For the covariance estimators with proven performance guarantees, such as [66] or [43], such guarantees are not available due to their non-linearity in observations. In [56], I have proposed a nearly affine-equivariant covariance estimator with strong statistical guarantees under L_4-L_2 moment equivalence, and essentially the same computational cost as the sample covariance matrix. Interestingly, [42] proposes an alternative construction, leading to a *sharp* statistical guarantee, but at the price of losing the algorithmic tractability.

1.3 Model discrimination and two-sample testing

Consider the following setup: one is given a parameter value $\theta^* \in \mathbb{R}^d$ and two i.i.d. samples $Z^{(0)} = (Z_1^{(0)}, ..., Z_n^{(0)})$ and $Z^{(1)} = (Z_1^{(1)}, ..., Z_n^{(1)})$ one of which—and only one—is known to come from $\mathbb{P}_{\theta^*}$, while the other one comes from some unknown distribution $\mathbb{P}_{\bar{\theta}}$. One is then asked the question:

What is the sample complexity of recognizing the "right" sample (corresponding to $\mathbb{P}_{\theta^*}$)?

Assuming that $\mathbb{P}_{\theta^*}$ and $\mathbb{P}_{\bar{\theta}}$ belong to the same parametric family of distributions, we obtain a parametric version of this problem; the sample complexity should then be expected to depend on d—the parameter dimension, and Δ—a measure of separation between the two distributions.

This problem was introduced and studied in my joint work [55] with Mohamed Ndaoud (ESSEC). Our motivation came from applications: as it turns out, certain problems in statistical privacy and fairness can be reduced to solving such a hypothesis testing problem. For example, one can consider the task of determining which of the two (large) datasets a given statistical model was trained with, *without explicit knowledge of the datasets*. In this case, $Z^{(0)}$ and $Z^{(1)}$ are the (small) *subsamples* of the full datasets, and θ^* corresponds to the statistical model in question. In the case of random-design linear regression model, we have managed to resolve the problem, constructing a testing procedure with a worst-case near-optimal sample complexity. Similarly, we obtained sharp results for general parametric models in the appropriate asymptotic regime (with $\Delta \to 0$ as $n \to \infty$) and for generalized linear models with a small sample size. Our next step is to revisit the classical—and closely related—problem of *two-sample testing,* where the statistician is asked to understand whether two samples have the same distribution or not.

1.4 Structure-adaptive estimation

Consider estimation of a real- or complex-valued discrete-time *signal* $x := (x_\tau)$, $-n \le \tau \le n$, 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 [24, 57, 51], 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 [50], 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. I am now investigating the natural extension of the structure-adaptive estimation problem to the case of indirect observations, where the signal is fed through a linear time-invariant filter before being corrupted with the noise. Potentially, this work might lead to some progress in the classical problem of identification of a linear dynamical system from its noisy output; see, e.g., [9, 25].

2 Optimization theory

2.1 Tractable methods for online optimization

In a nutshell, online optimization is a framework for sequential and adversarial decision-making, formalized as a game played between the learner and her adversary. Learner makes a sequence of decisions over T rounds, formalized as a sequence $x_1, x_2, ...$ of elements from a certain set $X \subseteq \mathbb{R}^d$. In each round $t \in \{1, ..., T\}$, the decision x_t made by learner is revealed to the adversary, who then selects the next *instantaneous loss* $\ell_t(\cdot) : X \to \mathbb{R}$ from a certain functional class $\mathcal{F}$. The learner suffers the loss $\ell_t(x_t)$, but observes $\ell_t(\cdot)$ as a whole, and uses the information accumulated so far to make a selection x_{t+1} in the next round. Her strategy of generating $x_1, ..., x_T$ is evaluated against the best decision *in hindsight:* in other words, her goal is to have a small *regret*

$$\mathscr{R}_T(x_1, ..., x_T) \;:=\; \sum_{t=1}^{T} \ell_t(x_t) \;-\; \min_{x \in X} \sum_{t=1}^{T} \ell_t(x).$$

Here, the class of losses $\mathcal{F}$ is known to the learner, and the choice of $\mathcal{F}$ and X specifies a particular online optimization problem. If the set X and all functions in $\mathcal{F}$ are convex, one is dealing with *online convex optimization* (OCO). On the other hand, the strategy employed by the learner in order to make the decisions is known to the adversary, and specific strategies correspond to correspond to *online optimization algorithms.*

One direction of my recent work concerns a very natural online optimization setup called *online portfolio selection*, where X is the standard d-simplex, and $\mathcal{F}$ is comprised of all functions of the form $-\log(x^\top p)$ parametrized by $p \in \mathbb{R}^d_+$, a vector with nonnegative entries. This models *periodic investment*: in each round, the learner—or trader—invests her current capital by selecting a distribution x_t of her capital over d assets, after which the adversary—"the invisible hand"—selects a vector p_t of asset prices in this trading round. Then $x_t^\top p_t = \exp(-\ell_t(x_t))$ is the *return* on trader's investment, minimizing the regret corresponds to maximizing the "compound return" $\prod_{t=1}^T x_t^\top p_t$ after T investment rounds. In his seminal work [15], T. Cover simultaneously proposed this model and an algorithm, termed "Universal Portfolios," with an $O(d \log T)$ regret. Later on, this regret guarantee was shown to be optimal [12]. Unfortunately, Cover's algorithm is challenging from the computational standpoint: it involves computing certain d-dimensional integral, and in a practical implementation this integral has to be approximated. In the early 2000s, Kalai and Venpala [30] proposed an implementation of Universal Portfolios based on a customly-tailored Monte-Carlo sampling scheme; their implementation ran in $O(d^4(d+T)^{14})$ per round, essentially remaining prohibitive in any application scenario. On the other hand, a plethora of computationally feasible methods for online portfolio selection has been proposed

over the next two decades, but none of them could match the optimal regret of Cover's algorithm. Devising a truly practical alternative to Universal Portfolios is a long-standing open problem.

In my recent work [29], the above challenge has finally been addressed. Namely, I have proposed an algorithm with near-optimal regret— the same as for Universal portfolios up to a constant factor—and the runtime of only $O(d^2(T+d))$. In a nutshell, the algorithm reduces to minimizing the observed cumulative loss—the strategy known as *follow-the-leader*—regularized by the so-called *volumetric barrier* for the polytope specified by the vectors $\{p_1, ..., p_t\}$ of asset returns. In fact, this strategy arises as a natural variational approximation of the Universal Portfolios. It is in my future plans to expand this work in several directions, as discussed next.

Online portfolio selection with nearly constant runtime. In practical applications, it is preferable to have a procedure whose per-round runtime does not grow with the number of rounds, i.e. it should depend only on d but not on T. In fact, one may argue—heuristically—that the runtime for any procedure with near-optimal regret cannot be smaller than $O(d^3)$. Meanwhile, for the algorithm we have proposed in [29], the runtime is larger in the regime $T \gg d$. Finding a procedure that would attain the optimal runtime *and* regret, simultaneously, remains challenging.

Improved algorithms for quantum state estimation. Quantum state estimation is a fundamental problem at the intersection of information theory and quantum computing [1]. The goal is to build a reliable estimate of the state of a quantum system from linear measurements. Naturally, such a state can be described by a positive-semidefinite matrix with unit trace, and maximum likelihood estimation falls into the setting of online optimization with X being the set of all such matrices, and $\mathcal{F}$ consisting of functions of the form $-\log \operatorname{tr}\left(\mathsf{X}^\top \mathsf{P}\right)$ where P is a positive-semidefinite "observation matrix." Thus, quantum state estimation can be viewed as a natural "matrix generalization" of the online portfolio selection problem, and very similar mathematical tools can be applied to address it; this has recently been exploited in [67] to obtain tractable quantum-state estimation methods with improved performance. However, similarly to the case of online portfolio estimation, it is still unclear how far one can go in terms of reducing the computational burden of a *near-optimal* method. There are strong reasons to believe that adaptation of our approach in [29] will lead to a substantial progress in answering this question.

Tractable approximations of Online Mirror Descent in the space of measures. In a nutshell, Cover's Universal Portfolios algorithm is an instantiation of the general Online Mirror Descent (OMD) scheme [26] to the problem at hand, in a situation where this scheme turns out to be computationally burdensome because the search is performed over an infinite-dimensional space of probability measures. Similar situations arise elsewhere, notably in the context of computational Bayesian inference [16] and continuous Langevin dynamics [14]. Meanwhile, volumetric-barrier regularization can be understood in this more general context, as a generic method to get a computationally cheap approximation of OMD with a similar quality, via its variational formulation. Naturally, adapting this technique to other scenarios where Mirror Descent is computationally challenging is a promising direction of my future research endeavors.

2.2 Min-max optimization and saddle-point problems

In the past couple of years, I spent a significant 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 [40], fair inference [5], generalized adversarial networks [23], and reinforcement learning [17]. 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. Finally, some of my work has been concerned with the classical *convex-concave saddle-point* problems; here, I have adapted some known optimization methods to construct efficient estimation and prediction procedures.

First-order stationarity in nonconvex-concave problems. In [54], I proposed a simple and intuitive algorithm with state-of-the-art performance guarantees for the task of finding approximate first-order Nash equilibria and stationary points in nonconvex-concave and nonconvex-strongly concave min-max problems. These guarantees have later on been shown to be near-optimal in the strongly concave case [33], and are conjectured to be so in the non-strongly concave case. Closing this gap is a well-known open problem in the corresponding community.

Nonconvex-nonconcave optimization with a small maximization domain. In [53], I have studied 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 $\operatorname{diam}(Y) \leq \bar{\mathsf{D}}$ 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.

Efficient algorithms for large-scale multiclass learning. In [3], I applied the machinery of first-order methods for convex-concave saddle-point problems to the task of training multiclass linear classifiers with a large number of classes on massive datasets. More specifically, the goal was to efficiently train a linear classifier with a very large number of classes k, features d, and sample size n, via regularized empirical risk minimization, focusing 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 ([45, 49]) 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 ([59, 28, 61, 19]) in combination with sampling over the features [58, 62] 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 [3], I address this challenge by designing ad-hoc bilevel sampling schemes in combination with a careful choice of proximal geometry. In particular, for multiclass support vector machines (SVM) with ℓ_1 regularization, the resulting algorithm has runtime $O(d+n+k)$ per iteration, i.e., is *sublinear* in the input size. To the best of my knowledge, this result is the first of its kind for multiclass linear classification. Extending it to the case of multiclass logistic regression is an open problem.

2.3 Optimization and performance estimation with non-Euclidean geometry

In [22], 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 with a convex and smooth objective function) 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., [34, 31, 64, 18, 63]). 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 positive-semidefinite and Gram matrices [65].

However, 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, 36] 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 R. Dragomir (now at EPFL), which has formed a chapter of his PhD thesis [20, Chap. 5], we resolved this question in the positive, finding a convex programming formulation of performance estimation for entropy-smooth minimization.

Nesterov-type acceleration in the class of functions smooth with respect to entropy. Our next goal is to apply the framework to the question of whether it is possible to accelerate first-order methods "à la Nesterov" [46] for the corresponding functional class. This is an intriguing question: on one hand, Dragomir et al. [21] recently showed that acceleration is impossible under relative smoothness with respect to a certain *adversarially constructed* Bregman divergence. On the other hand, in the case of entropy one has additional algebraic structure lacking in the general case [7]. Overall, some preliminary heuristic considerations indicate that acceleration in this context might not be possible. In the future, I hope to resolve this question rigorously.

References

[1] S. Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2088):3089–3114, 2007.

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

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

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

[5] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn. Rényi fair inference. *arXiv preprint arXiv:1906.12005*, 2019.

[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] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[13] 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.

[14] L. Chizat. Mean-field Langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*, 2022.

[15] T. M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.

[16] B. Dai, N. He, H. Dai, and L. Song. Provable Bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pages 985–994. PMLR, 2016.

[17] 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.

[18] 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.

[19] 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.

[20] R.-A. Dragomir. *Bregman gradient methods for relatively-smooth optimization*. PhD thesis, l'Université Toulouse 1 Capitole, 2021.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[27] 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.

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

[29] R. Jézéquel, D. M. Ostrovskii, and P. Gaillard. Efficient and near-optimal online portfolio selection. *arXiv preprint arXiv:2209.13932*, 2022.

[30] A. T. Kalai and S. Vempala. Efficient algorithms for universal portfolios. *Journal of Machine Learning Research*, pages 423–440, 2002.

[31] 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.

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

[33] 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.

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

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

[36] 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.

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

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

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

[40] 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.

[41] 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.

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

[43] S. Mendelson and N. Zhivotovskiy. Robust covariance estimation under L_4–L_2 norm equivalence. *arXiv preprint arXiv:1809.10462*, 2018.

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

[45] 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.

[46] 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.

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

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

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

[50] 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.

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

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

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

[54] 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.

[55] D. M. Ostrovskii, M. Ndaoud, A. Javanmard, and M. Razaviyayn. Near-optimal procedures for model discrimination with non-disclosure properties. *arXiv preprint arXiv:2012.02901*, 2020.

[56] 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.

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

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

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

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

[61] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2):567–599, 2013.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] J. Zimmert, N. Agarwal, and S. Kale. Pushing the efficiency-regret Pareto frontier for online learning of portfolios and quantum states. *arXiv preprint arXiv:2202.02765*, 2022.