Moscow institute of physics and technology

as a manuscript

Pavel Evgenievich Dvurechenskii

Numerical methods in large-scale optimization: inexact oracle and primal-dual analysis

Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Sciences in Computer Science

Moscow - 2020

The dissertation was prepared at Moscow Institute of Physics and Technology (National Research University).

Scientific Consultant:

Alexander Vladimirovich Gasnikov, Doctor of Sciences in Mathematical Modelling, Numerical Methods and Software Complexes, Associate Professor at Mathematical Foundations of Control chair, Moscow Institute of Physics and Technology.

1 Introduction

Numerical optimization remains an active area of research since 1980's, motivated by a vast range of applications, e.g. operations research, optimal control. Starting with the works [1, 2] one of the main areas of research in numerical optimization became interior-point methods. These methods combine Newton steps with penalty approach and allow to solve a very general class of convex problems in polynomial-time, which is justified both theoretically and practically. The new century introduced new challenges for numerical methods in optimization. Thanks to increasing amount of available data and more powerful computational resources, machine learning became an area of intensive research. A cornerstone optimization problem in machine learning is the empirical risk minimization with the key aspect being large dimension of the decision variable and large number of components used in the objective function. In this setting the Newton iteration becomes expensive in general since it requires matrix inversion. This motivated a sacrifice of logarithmic dependence on the accuracy to a cheap iteration and the use of first-order methods to solve such problems. Another reason was that the data is usually noisy and there is no need to solve the optimization problem to a high accuracy in this setting. Another main application for first-order methods is signal processing and image analysis, where the goal is to reconstruct a high-dimensional signal from high-dimensional data, e.g. noisy images.

Yet, known already for a long time [3, 4, 5], first-order methods entered their renaissance in 2000's. Some important facts on these methods were already known for 15 years. In particular, the concept of black-box oracle [6] allowed to obtain lower worst-case complexity bounds for different classes of problems and methods. In particular, a gap was discovered between the lower bound $O(1/k^2)$ and an upper bound $O(1/k)$ for the convergence rate of the gradient method for minimizing convex smooth functions. Here k is the iteration counter. This gap led to an important phenomenon of acceleration for first-order methods and accelerated gradient method [7]. In the new century many extensions of this algorithm were proposed motivated by image processing problems and machine learning, including composite versions [8, 9], accelerated stochastic gradient method [10], accelerated variance reduction methods [11, 12, 13, 14, 15]. In addition to accelerated stochastic gradient methods for finite-sum problems, which use a random choice of the gradient of the component, acceleration was introduced for other randomized methods such as random coordinate descent [16] and random gradient-free optimization [17]. The latter is motivated by problems, in which only zero-order oracle is available, e.g. when the objective is given as a solution of some auxiliary problem. For this setting, it is important to analyze zero-order methods with inexact function values since this auxiliary problem may be possible to solve only inexactly. In the setting of first-order methods inexactness may also be encountered in practice. Accelerated gradient method with inexact gradients was analyzed in [18], and an important framework of inexact first-order oracle was introduced in [19]. Another important extension of accelerated first-order methods are accelerated methods for problems with linear constraints, which was proposed in [20], yet with a non-optimal rate $O(1/k)$ for the constraints feasibility.

Object and goals of the dissertation. The goal of the dissertation is twofold. The first goal is to further extend the existing first and zero-order methods for problems with inexactness in function and gradient values, the inexactness being deterministic or stochastic. The second goal is to construct new primal-dual first-order methods, which allow to solve simultaneously the primal and dual problem with optimal convergence rates. A particular focus is made on problems with linear constraints and the application of the proposed methods to optimal transport distance and barycenter problems.

The obtained results:

1. We propose a stochastic intermediate gradient method for convex problems with stochastic inexact oracle.

2. We develop a gradient method with inexact oracle for deterministic non-convex optimization.

3. We develop gradient-free method with inexact oracle for deterministic convex optimization.

4. We develop a method to calculate the derivative of the pagerank vector and in combination with the above two methods propose gradient-based and gradient-free optimization methods for learning supervised pagerank model.

5. We develop a concept of inexact oracle for the methods which use directional derivatives and propose accelerated directional derivative method for smooth stochastic convex optimization. We also develop an accelerated and non-accelerated directional derivative method for strongly convex smooth stochastic optimization.

6. We develop primal-dual methods for solving infinite-dimensional games in convex-concave and strongly convex-concave setting.

7. We develop non-adaptive and adaptive accelerated primal-dual gradient method for strongly convex minimization problems with linear equality and inequality constraints.

8. We apply this algorithm to the optimal transport problem and obtain new complexity estimates for this problem, which in some regime are better than the ones for the Sinkhorn's algorithm.

9. We propose a stochastic primal-dual accelerated gradient method for problems with linear constraints and apply it to the problem of approximation of Wasserstein barycenter.

10. We propose a primal-dual extension of accelerated methods which use line-search to define the stepsize and to be adaptive to the Lipschitz constant of the gradient.

Author's contribution includes the development of the listed above optimization methods, proving convergence rates and complexity result theorems for these methods and their applications to optimal transport problems and learning problem for a supervised pagerank model.

Novelties. The proposed versions of accelerated first and zero-order methods for convex optimization under different types of inexactness are novel. The proposed primal-dual methods for the listed setups are also novel, and allow to obtain new methods for optimal transport problems. In particular, we obtain new complexity results for non-regularized optimal transport problem and a new distributed algorithm for approximating Wasserstein barycenter of a set of measures using samples from these measures.

As a result of the work on this dissertation, 10 papers were published:

First-tier publications:

1. Dvurechensky, P., and Gasnikov, A. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. Journal of Optimization Theory and Applications 171, 1 (2016), 121–145, Scopus Q1 (main co-author; the author of this thesis proposed main algorithms, formulated and proved convergence rate theorems for the proposed methods).

2. Gasnikov, A. V., and Dvurechensky, P. E. Stochastic intermediate gradient method for convex optimization problems. Doklady Mathematics 93, 2 (2016), 148–151, Scopus Q2 (main co-author; the author of this thesis proposed main algorithms, formulated and proved convergence rate theorems for the proposed methods).

3. Bogolubsky, L., Dvurechensky, P., Gasnikov, A., Gusev, G., Nesterov, Y., Raigorodskii, A. M., Tikhonov, A., and Zhukovskii, M. Learning supervised pagerank with gradient-based and gradient-free optimization methods. In Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 4914–4922, CORE A* (the author of this thesis proposed general gradient-free (Algorithm 1,2) and gradient (Algorithm 3,4) methods with inexact oracle, proposed a method for approximating the derivative of the pagerank vector, formulated and proved convergence rate theorems for the proposed methods: Lemma 1,2, Theorem 1-4).

4. Dvurechensky, P., Gorbunov, E., and Gasnikov, A. An accelerated directional derivative method for smooth stochastic convex optimization. European Journal of Operational Research (2020), `https://doi.org/10.1016/j.ejor.2020.08.027`, Scopus Q1 (main co-author; the author of this thesis proposed a concept of inexact oracle for directional derivatives in stochastic convex optimization, proved (in inseparable cooperation with E. Gorbunov) convergence rate Theorem 1 for the accelerated directional derivative method, proved convergence rate Theorems 3,4 for strongly convex problems).

5. Dvurechensky, P., Nesterov, Y., and Spokoiny, V. Primal-dual methods for solving in infinite-dimensional games. Journal of Optimization Theory and Applications 166, 1 (2015), 23–51, Scopus Q1 (main co-author; the author of this thesis developed main algorithms and proved convergence rate theorems).

6. Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In Proceedings of the 5th International Conference on Machine Learning (2018), J. Dy and A. Krause, Eds., vol. 80 of Proceedings of Machine Learning Research, pp. 1367–1376, CORE A* (main co-author; the author of this thesis proposed general primal-dual adaptive accelerated gradient method (Algorithm 3) for problems with linear constraints, proved convergence rate Theorem 3, proposed an algorithm for approximating optimal transport (OT) distance (Algorithm 4), obtained complexity bound for approximating OT distance (Theorem 4), performed numerical experiments for comparison of this method with the Sinkhorn's method).

7. Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., and Nedić, A. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In Advances in Neural Information Processing Systems 31 (2018), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., NeurIPS 2018, Curran Associates, Inc., pp. 10783–10793, CORE A* (main co-author; the author of this thesis proposed the general idea of the paper, general primal-dual accelerated stochastic gradient method (Algorithm 2) for problems with linear constraints, proved convergence rate Theorem 2, proposed an algorithm for approximating Wasserstein barycenter (Algorithm 4), proved (in inseparable cooperation with D. Dvinskikh) its complexity Theorem 3).

8. Guminov, S. V., Nesterov, Y. E., Dvurechensky, P. E., and Gasnikov, A. V. Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems. Doklady

Mathematics 99, 2 (2019), 125-128, Scopus Q2 (the author of this thesis proposed a primal-dual variant of the accelerated gradient method with linesearch for problems with linear constraints, proved convergence rate Theorem 3).

9. Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. Optimization Methods and Software (2020), `https://doi.org/10.1080/10556788.2020.1731747`, Scopus Q1 (the author of this thesis proposed a primal-dual variant of the universal accelerated gradient method with small-dimensional relaxation (Algorithm 7) for problems with linear constraints, proved its convergence rate Theorem 4.1).

Second-tier publications:

1. Chernov, A., Dvurechensky, P., and Gasnikov, A. Fast primal-dual gradient method for strongly convex minimization problems with linear constraints. In Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings (2016), Y. Kochetov, M. Khachay, V. Beresnev, E. Nurminski, and P. Pardalos, Eds., Springer International Publishing, pp. 391–403, Web of Science and Scopus (main co-author; the author of this thesis developed main algorithm and proved convergence rate theorem).

Reports at conferences and seminars:

1. International Workshop "Advances in Optimization and Statistics", Berlin, 15.05.2014–16.05.2014, "Stochastic Intermediate Gradient Method for Convex Problems with Inexact Stochastic Oracle".

2. Seminar "Modern Methods in Applied Stochastics and Nonparametric Statistics", Berlin, 03.06.2014, "Gradient methods for convex problems with stochastic inexact oracle".

3. V International Conference on Optimization Methods and Applications (OPTIMA-2014), Petrovac, Montenegro, 28.09.2014–04.10.2014, "Gradient-free optimization methods with ball randomization".

4. VI traditional school for young scientists "Control, information, optimization", Moscow, 22.06.2014-29.06.2014, "Gradient methods for convex problems with stochastic inexact oracle".

5. 38-th conference-school of IITP RAS "Information technologies and systems", Nizhnii Novgorod, 01.09.2014–05.09.2014, "Stochastic Intermediate Gradient Method for Convex Problems with Inexact Stochastic Oracle".

6. Workshop “Frontiers of High Dimensional Statistics, Optimization, and Econometrics”, Moscow, 26.02.2015–27.02.2015, "Random gradient-free methods for random walk based web page ranking functions learning".
7. VII traditional school for young scientists "Control, information, optimization", Moscow, 14.06.2014-20.06.2014, "Semi-Supervised PageRank Model Learning with Gradient-Free Optimization Methods".
8. 29-th conference-school of IITP RAS "Information technologies and systems", Sochi, 07.09.2014–11.09.2015, "Stochastic Intermediate Gradient Method: convex and strongly-convex case".
9. 30th annual conference of Belgian Operational Research Society (ORBEL 30), Louvain-la-Neuve, Belgium, 28.01.2016–29.01.2016, "Random gradient-free methods for ranking algorithm learning".
10. Workshop on Modern Statistics and Optimization, Moscow, 23.02.2016–24.02.2016, "Gradient and gradient-free methods for pagerank algorithm learning".
11. VII International Conference Optimization and Applications (OPTIMA 2016), Petrovac, Montenegro, 25.09.2016–02.10.2016, "Accelerated Primal-Dual Gradient Method for Linearly Constrained Minimization Problems".
12. VIII Moscow International Conference on Operations Research (ORM 2016), Moscow, 17.10.2016–22.10.2016, "Accelerated Primal-Dual Gradient Method for Composite Optimization with Unknown Smoothness Parameter"
13. **Conference on Neural Information Processing Systems (NIPS 2016)**, Barcelona, 05.12.2016–10.12.2016, "Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods".
14. Workshop Shape, Images and Optimization, Münster, Germany, 28.02.2017 –03.03.2017, "Gradient Method With Inexact Oracle for Composite Non-Convex Optimization".
15. Optimization and Statistical Learning, Les Houches, France, 10.04.2017 – 14.04.2017, "Gradient Method With Inexact Oracle for Composite Non-Convex Optimization".
16. Foundations of Computational Mathematics, Barcelona, Spain, 10.07.2017 – 19.07.2017, "Gradient Method With Inexact Oracle for Composite Non-Convex Optimization".

17. Co-Evolution of Nature and Society Modelling, Problems & Experience. Devoted to Academician Nikita Moiseev centenary (Moiseev-100), Moscow, 07.11.2017 – 10.11.2017, "Adaptive Similar Triangles Method: a Stable Alternative to Sinkhorn's Algorithm for Regularized Optimal Transport".

18. 18th French-German-Italian Conference on Optimization, Germany, 25.09.2017 – 28.09.2017, Paderborn, Germany, "Gradient method with inexact oracle for composite non-convex optimization"

19. 3. International Matheon Conference on Compressed Sensing and its Applications, Berlin, 04.12.2017 – 08.12.2017, "Adaptive Similar Triangles Method: a Stable Alternative to Sinkhorn's Algorithm for Regularized Optimal Transport".

20. Games, Dynamics and Optimization (GDO2018), Vienna, Austria, 13.03.2018 – 15.03.2018, "Primal-Dual Methods for Solving Infinite -Dimensional Games".

21. **International Conference on Machine Learning (ICML 2018)**, Stockholm, Sweden, 10.07.2018 – 15.07.2018, "Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm".

22. 23rd International Symposium on Mathematical Programming, Bordeaux, France, 01.07.2018 – 06.07.2018, "Computational Optimal Transport: Accelerated Gradient Descent vs Sinkhorn".

23. Grenoble Optimization Days 2018: Optimization algorithms and applications in statistical learning, Grenoble, France, 28.06.2018 – 29.06.2018, "Faster algorithms for (regularized) optimal transport".

24. Statistical Optimal Transport Conference, Moscow, 24.07.2018 – 25.07.2018, "Computational Optimal Transport: Accelerated Gradient Descent vs Sinkhorn's Algorithm".

25. **Conference on Neural Information Processing Systems (NIPS 2018)**, Montreal, Canada, 02.12.2018 – 08.12.2018, "Decentralize and randomize: Faster algorithm for Wasserstein barycenters".

26. Optimization and Statistical Learning, Les Houches, France, 24.03.2019 – 29.03.2019, "Distributed optimization for Wasserstein barycenter".

27. **International Conference on Machine Learning (ICML 2019)**, Long Beach, USA, 09.06.2019 – 15.06.2019, "On the Complexity of Approximating Wasserstein Barycenters".

28. International Conference on Continuous Optimization (ICCOPT 2019), Berlin, Germany, 03.08.2019 – 08.08.2019, "A Unifying Framework for Accelerated Randomized Optimization Methods".

29. Workshop on optimization and applications, Moscow, 27.09.2019, "Accelerated Alternating Minimization for Optimal Transport".

30. Recent advances in mass transportation, Moscow, 23.09.2019 - 27.09.2019, "On the complexity of optimal transport problems".

31. Workshop by the GAMM Activity Group on Computational and Mathematical Methods in Data Science, Berlin, Germany, 24.10.2019 – 25.10.2019, "On the complexity of optimal transport problems".

32. HSE-Yandex autumn school on generative models, Moscow, 26.11.2019 – 29.11.2019, "Optimization methods for optimal transport".

33. Workshop on Mathematics of Deep Learning 2019, Berlin, Germany, 03.12.2019 – 05.12.2019, "On the complexity of optimal transport problems".

34. Workshop on PDE Constrained Optimization under Uncertainty and Mean Field Games, Berlin, Germany, 28.01.2020 – 30.01.2020, "Distributed optimization for Wasserstein barycenters".

2 Optimization with inexact oracle

In this section we briefly describe the methods and their convergence properties for optimization problems under inexact information. We consider first-order methods and directional derivative methods.

2.1 Stochastic intermediate gradient method for convex problems with stochastic inexact oracle

The results of this subsection are published in [21, 22].

Let E be a finite-dimensional real vector space and E^* be its dual. We denote the value of a linear function $g \in E^*$ at $x \in E$ by $\langle g, x \rangle$. Let $\| \cdot \|$ be some norm on E. We denote by $\| \cdot \|_*$ the dual norm for $\| \cdot \|_E$, i.e. $\|g\|_* = \sup_{y \in E}\{\langle g, y \rangle : \|y\|_E \leq 1\}$. By $\partial f(x)$ we denote the subdifferential of the function $f(x)$ at a point x. In this subsection, we consider the *composite optimization* problem of the form

$$\min_{x \in Q}\{\varphi(x) := f(x) + h(x)\}, \tag{1}$$

where $Q \subset E$ is a closed and convex set, $h(x)$ is a simple convex function, $f(x)$ is a convex function with *stochastic inexact oracle* [23]. This means that, for every $x \in Q$, there exist $f_{\delta,L}(x) \in \mathbb{R}$ and $g_{\delta,L}(x) \in E^*$, such that

$$0 \leq f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x \rangle \leq \frac{L}{2}\|x - y\|^2 + \delta, \quad \forall y \in Q, \qquad (2)$$

and also that, instead of $(f_{\delta,L}(x), g_{\delta,L}(x))$ (we will call this pair a (δ, L)-oracle), we use their stochastic approximations $(F_{\delta,L}(x,\xi), G_{\delta,L}(x,\xi))$. The latter means that, for any point $x \in Q$, we associate with x a random variable ξ whose probability distribution is supported on a set $\Xi \subset \mathbb{R}$ and such that $\mathbb{E}_\xi F_{\delta,L}(x,\xi) = f_{\delta,L}(x)$, $\mathbb{E}_\xi G_{\delta,L}(x,\xi) = g_{\delta,L}(x)$ and $\mathbb{E}_\xi(\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*)^2 \leq \sigma^2$.

To deal with such problems we will need a *prox-function* $d(x)$, which is differentiable and strongly convex with parameter 1 on Q with respect to $\|\cdot\|$. Let x_0 be the minimizer of $d(x)$ on Q. By translating and scaling $d(x)$, if necessary, we can always ensure that $d(x_0) = 0$, $d(x) \geq \frac{1}{2}\|x - x_0\|^2$, $\forall x \in Q$. We define also the corresponding *Bregman distance*: $V(x,z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$. Let $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0} \subset \mathbb{R}$ be three sequences of coefficients satisfying

$$\alpha_0 \in]0,1], \quad \beta_{i+1} \geq \beta_i > L, \quad \forall i \geq 0, \qquad (3)$$

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \qquad (4)$$

$$\alpha_k^2 \beta_k \leq B_k \beta_{k-1} \leq \left(\sum_{i=0}^{k} \alpha_i\right) \beta_{k-1}, \quad \forall k \geq 1. \qquad (5)$$

$$A_k := \sum_{i=0}^{k} \alpha_i, \quad \tau_i := \frac{\alpha_{i+1}}{B_{i+1}} \qquad (6)$$

The Stochastic Intermediate Gradient Method (SIGM) is described below as Algorithm 1. Let $a \geq 1$ and $b \geq 0$ be some parameters. Let us assume that we know a number R such that $\sqrt{2d(x^*)} \leq R$. We set for $p \in [1,2]$

$$\alpha_i = \frac{1}{a}\left(\frac{i+p}{p}\right)^{p-1}, \quad \forall i \geq 0, \qquad (7)$$

$$\beta_i = L + \frac{b\sigma}{R}(i+p+1)^{\frac{2p-1}{2}}, \quad \forall i \geq 0, \qquad (8)$$

$$B_i = a\alpha_i^2 = \frac{1}{a}\left(\frac{i+p}{p}\right)^{2p-2}, \quad \forall i \geq 0. \qquad (9)$$

Theorem 2.1. *If the sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ are chosen according to* (7), (8), (9) *with $a = 2^{\frac{2p-1}{2}}$ and $b = 2^{\frac{5-2p}{4}} p^{\frac{1-2p}{2}}$, then the sequence*

y_k *generated by the SIGM satisfies*

$$\mathbb{E}_{\xi_0,\ldots,\xi_k}\varphi(y_k) - \varphi^* \leq \frac{LR^2p^p2^{\frac{2p-3}{2}}}{(k+p)^p} + \frac{\sigma R 2^{\frac{3+2p}{4}}\sqrt{p}(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p} +$$
$$+ 2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1}+1\right)\delta \leq \frac{C_1LR^2}{k^p} + \frac{C_2\sigma R}{\sqrt{k}} + C_3k^{p-1}\delta =$$
$$= \Theta\left(\frac{LR^2}{k^p} + \frac{\sigma R}{\sqrt{k}} + k^{p-1}\delta\right),$$

where $C_1 = 4\sqrt{2}$*,* $C_2 = 16\sqrt{2}$*,* $C_3 = 48$*.*

Algorithm 1 Stochastic Intermediate Gradient Method (SIGM)

Require: The sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$, functions $d(x)$, $V(x,z)$.
Ensure: The point y_k.
1: Compute $x_0 := \arg\min_{x\in Q}\{d(x)\}$. Let ξ_0 be a realization of the random variable ξ. Calculate $G_{\delta,L}(x_0,\xi_0)$. Set $k=0$.
2: $y_0 := \arg\min_{x\in Q}\{\beta_0 d(x) + \alpha_0\langle G_{\delta,L}(x_0,\xi_0), x - x_0\rangle + \alpha_0 h(x)\}$.
3: **repeat**
4: $z_k := \arg\min_{x\in Q}\{\beta_k d(x) + \sum_{i=0}^{k}\alpha_i\langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle + A_k h(x)\}$.
5: $x_{k+1} := \tau_k z_k + (1-\tau_k)y_k$.
6: Let ξ_{k+1} be a realization of the random variable ξ. Calculate $G_{\delta,L}(x_{k+1},\xi_{k+1})$.
7: $\hat{x}_{k+1} := \arg\min_{x\in Q}\{\beta_k V(x,z_k) + \alpha_{k+1}\langle G_{\delta,L}(x_{k+1},\xi_{k+1}), x - z_k\rangle + \alpha_{k+1}h(x).\}$.
8: $w_{k+1} := \tau_k\hat{x}_{k+1} + (1-\tau_k)y_k$.
9: $y_{k+1} := \frac{A_{k+1}-B_{k+1}}{A_{k+1}}y_k + \frac{B_{k+1}}{A_{k+1}}w_{k+1}$.
10: **until**

It is possible to obtain an upper bound on the probability of large deviations for the $\varphi(y_k) - \varphi^*$. To do that, we make the following additional assumptions.

1. $\xi_0,\ldots,\xi_k$ are i.i.d random variables.

2. $G_{\delta,L}(x,\xi)$ satisfies the light-tail condition

$$\mathbb{E}_\xi\left[\exp\left(\frac{\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*^2}{\sigma^2}\right)\right] \leq \exp(1).$$

3. Set Q is bounded, and we know a number $D > 0$, such that $\max_{x,y\in Q}\|x-y\| \leq D$.

Theorem 2.2. *If the sequences* $\{\alpha_i\}_{i\geq 0}$*,* $\{\beta_i\}_{i\geq 0}$*,* $\{B_i\}_{i\geq 0}$ *are chosen according to* (7), (8), (9) *with* $a = 2^{\frac{2p-1}{2}}$ *and* $b = 2^{\frac{5-2p}{4}} p^{\frac{1-2p}{2}}$*, then the sequence* y_k *generated by the SIGM satisfies*

$$\mathbb{P}\left(\varphi(y_k) - \varphi^* > \frac{C_1 L R^2}{k^p} + \frac{C_2(1+\Omega)\sigma R}{\sqrt{k}} + C_3 k^{p-1}\delta + \frac{C_4 D\sigma\sqrt{\Omega}}{\sqrt{k}}\right)$$

$$\leq \mathbb{P}\left(\varphi(y_k) - \varphi^* > \frac{LR^2 p^p 2^{\frac{2p-3}{2}}}{(k+p)^p} + \frac{(1+\Omega)\sigma R 2^{\frac{3+2p}{4}}\sqrt{p}(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p}\right.$$

$$\left. + 2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1} + 1\right)\delta + \frac{2D\sigma\sqrt{6\Omega p}}{\sqrt{k+p}}\right) \leq 3\exp(-\Omega),$$

where $C_1 = 4\sqrt{2}$*,* $C_2 = 16\sqrt{2}$*,* $C_3 = 48$*,* $C_4 = 4\sqrt{3}$*.*

Next, we consider two modifications of the SIGM for strongly convex problems. For the first modification, we obtain the rate of convergence in terms of the non-optimality gap expectation and for the second we bound the probability of large deviations from this rate. We additionally assume that E is a Euclidean space with scalar product $\langle \cdot, \cdot \rangle$ and norm $\|x\| := \sqrt{\langle x, Hx \rangle}$, where H is a symmetric positive definite matrix. Without loss of generality, we assume that the function $d(x)$ satisfies conditions $0 = \arg\min_{x\in Q} d(x)$ and $d(0) = 0$. Also we assume that the function $\varphi(x)$ is strongly convex, i.e. $\frac{\mu}{2}\|x - y\|^2 \leq \varphi(y) - \varphi(x) - \langle g(x), y - x\rangle$ for all $x, y \in Q, g(x) \in \partial\varphi(x)$. As a corollary, we have

$$\varphi(x) - \varphi(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2, \quad \forall x \in Q, \tag{10}$$

where x^* is the solution of the problem (1). We also assume that $d(x)$ satisfies the following property. If x_0 is a random vector such that $\mathbb{E}_{x_0}\|x - x_0\|^2 \leq R_0^2$ for some fixed point x and number R_0, then, for some $V > 0$,

$$\mathbb{E}_{x_0} d\left(\frac{x - x_0}{R_0}\right) \leq \frac{V^2}{2}. \tag{11}$$

Theorem 2.3. *After* $k \geq 1$ *outer iterations of Algorithm 2, we have*

$$\mathbb{E}\varphi(u_k) - \varphi^* \leq \frac{\mu R_0^2}{2}\mathrm{e}^{-k} + \frac{C_3 \mathrm{e} 2^{p-1}}{\mathrm{e}-1}\left(\frac{4\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\delta, \tag{15}$$

$$\mathbb{E}\|u_k - x^*\|^2 \leq R_0^2 \mathrm{e}^{-k} + \frac{C_3 \mathrm{e} 2^p}{\mu(\mathrm{e}-1)}\left(\frac{4\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\delta. \tag{16}$$

As a consequence, if we choose the error δ *of the oracle satisfying*

$$\delta \leq \frac{\varepsilon(\mathrm{e}-1)}{2^p C_3 \mathrm{e}}\left(\frac{4\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{1-p}{p}}, \tag{17}$$

Algorithm 2 Stochastic Intermediate Gradient Method for Strongly Convex Problems

Require: The function $d(x)$, point u_0, number R_0 such that $\|u_0 - x^*\| \le R_0$, number $p \in [1,2]$.

Ensure: The point u_{k+1}.

1: Set $k = 0$.

2: Calculate

$$N_k := \left\lceil \left(\frac{4\mathrm{e}C_1 LV^2}{\mu}\right)^{\frac{1}{p}} \right\rceil. \tag{12}$$

3: **repeat**

4: Calculate

$$m_k := \max\left\{1, \left\lceil \frac{16\mathrm{e}^{k+2}C_2^2\sigma^2V^2}{\mu^2R_0^2N_k} \right\rceil\right\}, \tag{13}$$

$$R_k^2 := R_0^2\mathrm{e}^{-k} + \frac{2^p\mathrm{e}C_3\delta}{\mu(\mathrm{e}-1)}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-k}\right). \tag{14}$$

5: Run Algorithm 1 with $x_0 = u_k$ and prox-function $d\left(\frac{x-u_k}{R_k}\right)$ for N_k steps, using oracle $\tilde{G}^k_{\delta,L}(x) := \frac{1}{m_k}\sum_{i=1}^{m_k} G_{\delta,L}(x,\xi^i)$, where ξ^i, $i = 1,...,m_k$ are i.i.d, on each step and sequences $\{\alpha_i\}_{i\ge0}$, $\{\beta_i\}_{i\ge0}$, $\{B_i\}_{i\ge0}$ defined in Theorem 2.1.

6: Set $u_{k+1} = y_{N_k}$, $k = k+1$.

7: **until**

then we need $N = \left\lceil \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right\rceil$ *outer iterations and no more than*

$$\left(1 + \left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{1}{p}}\right)\left(1+\ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right) + \frac{16\mathrm{e}^3C_2^2\sigma^2V^2}{\mu\varepsilon(\mathrm{e}-1)}$$

oracle calls to guarantee that $\mathbb{E}\varphi(u_N) - \varphi^* \le \varepsilon$.

To obtain complexity in terms of large deviations probability, we assume that the prox-function has quadratic growth with parameter V^2 with respect to the chosen norm, i.e.

$$d(x) \le \frac{V^2}{2}\|x\|^2, \quad \forall x \in \mathbb{R}^n. \tag{18}$$

Now we present a modification of Algorithm 2 and a theorem with a bound for the probability of large deviations for the non-optimality gap of this algorithm.

Algorithm 3 Stochastic Intermediate Gradient Method for Strongly Convex Problems 2

Require: The function $d(x)$, point u_0, number R_0 such that $\|u_0 - x^*\| \leq R_0$, number $p \in [1, 2]$, number $N \geq 1$ of outer iterations, confidence level Λ.

Ensure: The point u_N.

Set $k = 0$.

Calculate

$$N_k := \left\lceil \left(\frac{6\mathrm{e}C_1 L V^2}{\mu} \right)^{\frac{1}{p}} \right\rceil. \tag{19}$$

repeat

Calculate

$$m_k := \max \left\{ 1, \left\lceil \frac{36\mathrm{e}^{k+2} C_2^2 \sigma^2 V^2 \left(1 + \ln\left(\frac{3N}{\Lambda}\right)\right)^2}{\mu^2 R_0^2 N_k} \right\rceil, \left\lceil \frac{144\mathrm{e}^{k+2} C_4^2 \sigma^2 \ln\left(\frac{3N}{\Lambda}\right)}{\mu^2 R_0^2 N_k} \right\rceil \right\}, \tag{20}$$

$$R_k^2 := R_0^2 \mathrm{e}^{-k} + \frac{2^p \mathrm{e} C_3 \delta}{\mu(\mathrm{e}-1)} \left(\frac{6\mathrm{e}C_1 L V^2}{\mu} \right)^{\frac{p-1}{p}} \left(1 - \mathrm{e}^{-k}\right), \tag{21}$$

$$Q_k := \left\{ x \in Q : \|x - u_k\|^2 \leq R_k^2 \right\}. \tag{22}$$

Run Algorithm 1 applied to the problem $\min_{x \in Q_k} \varphi(x)$ with $x_0 = u_k$ and prox-function $d\left(\frac{x - u_k}{R_k}\right)$ for N_k steps using oracle $\tilde{G}_{\delta,L}^k(x) := \frac{1}{m_k} \sum_{i=1}^{m_k} G_{\delta,L}(x, \xi^i)$, where ξ^i, $i = 1, ..., m_k$ are i.i.d, on each step and sequences $\{\alpha_i\}_{i \geq 0}$, $\{\beta_i\}_{i \geq 0}$, $\{B_i\}_{i \geq 0}$ defined in Theorem 2.1.

Set $u_{k+1} = y_{N_k}$, $k = k + 1$.

until $k = N - 1$

Theorem 2.4. *After N outer iterations of Algorithm 3, we have*

$$\mathbb{P} \left\{ \varphi(u_N) - \varphi^* > \frac{\mu R_0^2}{2} \mathrm{e}^{-N} + \frac{2^{p-1} \mathrm{e} C_3 \delta}{(\mathrm{e}-1)} \left(\frac{6\mathrm{e}C_1 L V^2}{\mu} \right)^{\frac{p-1}{p}} \delta \right\} \leq \Lambda. \tag{23}$$

As a consequence, if we choose error of the oracle δ satisfying

$$\delta \leq \frac{\varepsilon(\mathrm{e}-1)}{2^p C_3 \mathrm{e}} \left(\frac{6\mathrm{e}C_1 L V^2}{\mu} \right)^{\frac{1-p}{p}}, \tag{24}$$

then we need no more than $N = \left\lceil \ln\left(\frac{\mu R_0^2}{\varepsilon}\right) \right\rceil$ outer iterations and no more *than*

$$
\left(1+\left(\frac{6\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{1}{p}}\right)\left(1+\ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)+
+\frac{36\mathrm{e}^3C_2^2\sigma^2V^2}{\mu(\mathrm{e}-1)\varepsilon}\left(1+\ln\left(\frac{3}{\Lambda}\left(1+\ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)\right)\right)^2+
+\frac{144\mathrm{e}^3C_4^2\sigma^2}{\mu\varepsilon(\mathrm{e}-1)}\ln\left(\frac{3}{\Lambda}\left(1+\ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)\right) \tag{25}
$$

oracle calls to guarantee that $\mathbb{P}\{\varphi(u_N)-\varphi^*>\varepsilon\}\le\Lambda$.

2.2 Learning supervised pagerank with gradient-based and gradient-free optimization methods.

In this subsection we consider a parametric model for web-page ranking and learning the parameters of this model in a supervised setting. The results of this subsection are published in [24].

2.2.1 Loss-minimization problem statement

We consider minimization of the following loss function

$$
f(\varphi)=\frac{1}{|Q|}\sum_{q=1}^{|Q|}\|(A_q\pi_q(\varphi))_+\|_2^2 \tag{26}
$$

as a function of $\varphi\in\mathbb{R}^m$ over some set of feasible values Φ, where vector x_+ has components $[x_+]_i=\max\{x_i,0\}$, the numbers q,r_q and matrices $A_q\in\mathbb{R}^{r_q\times p_q}, q\in Q$ are given. We denote $r=\max_{q\in Q}r_q$. Moreover, the probability vectors $\pi_q(\varphi)\in\mathbb{R}^{p_q}$ are the solutions of the equation

$$
\pi=\alpha\pi_q^0(\varphi)+(1-\alpha)P_q^T(\varphi)\pi, \tag{27}
$$

where $\pi_q^0(\varphi)\in\mathbb{R}^{p_q}$ is a given differentiable vector-function with first n_q non-zero components and all the rest being equal to zero, $P_q(\varphi)\in\mathbb{R}^{p_q\times p_q}$ is a given differentiable matrix-valued function. We denote $p=\max_{q\in Q}p_q$, $n=\max_{q\in Q}n_q$, $s=\max_{q\in Q}s_q$, where s_q is the maximum number of non-zero components in the rows of P_q.

We choose some $\hat\varphi$ and $R>0$ such that the set Φ defined as $\Phi=\{\varphi\in\mathbb{R}^m:\|\varphi-\hat\varphi\|_2\le R\}$ lies in the set of vectors with positive components $\mathbb{R}^m_{++}$. The loss-minimization problem which we solve is as follows

$$
\min_{\varphi\in\Phi}f(\varphi),\Phi=\{\varphi\in\mathbb{R}^m:\|\varphi-\hat\varphi\|_2\le R\}. \tag{28}
$$

The method [25] for approximation of $\pi_q(\varphi)$ for any fixed $q \in Q$ constructs a sequence π_k and the output $\tilde{\pi}_q(\varphi, N)$ (for some fixed non-negative integer N) by the following rule

$$\pi_0 = \pi_q^0(\varphi), \quad \pi_{k+1} = P_q^T(\varphi)\pi_k, \quad \tilde{\pi}_q(\varphi, N) = \frac{\alpha}{1-(1-\alpha)^{N+1}} \sum_{k=0}^{N} (1-\alpha)^k \pi_k. \tag{29}$$

Lemma 2.1. *Assume that for some $\delta_1 > 0$ Method (29) with $N = \left\lceil \frac{1}{\alpha} \ln \frac{8r}{\delta_1} \right\rceil - 1$ is used to calculate the vector $\tilde{\pi}_q(\varphi, N)$ for every $q \in Q$. Then*

$$\widetilde{f}(\varphi, \delta_1) = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \|(A_q \tilde{\pi}_q(\varphi, N))_+\|_2^2 \tag{30}$$

satisfies

$$|\widetilde{f}(\varphi, \delta_1) - f(\varphi)| \le \delta_1. \tag{31}$$

Moreover, the calculation of $\widetilde{f}(\varphi, \delta_1)$ requires not more than $|Q|(3mps + 3psN + 6r)$ a.o. and not more than $3ps$ memory items.

Our generalization of the method [25] for calculation of $\frac{d\pi_q(\varphi)}{d\varphi^T}$ for any $q \in Q$ is the following. Choose some non-negative integer N_1 and calculate $\tilde{\pi}_q(\varphi, N_1)$ using (29). Calculate a sequence Π_k

$$\Pi_0 = \alpha \frac{d\pi_q^0(\varphi)}{d\varphi^T} + (1-\alpha) \sum_{i=1}^{p_q} \frac{dp_i(\varphi)}{d\varphi^T} [\tilde{\pi}_q(\varphi, N_1)]_i, \quad \Pi_{k+1} = P_q^T(\varphi)\Pi_k. \tag{32}$$

The output is (for some fixed non-negative integer N_2)

$$\tilde{\Pi}_q(\varphi, N_2) = \frac{1}{1-(1-\alpha)^{N_2+1}} \sum_{k=0}^{N_2} (1-\alpha)^k \Pi_k. \tag{33}$$

In what follows, we use the following norm on the space of matrices $A \in \mathbb{R}^{n_1 \times n_2}$: $\|A\|_1 = \max_{j=1,\dots,n_2} \sum_{i=1}^{n_1} |a_{ij}|$.

Lemma 2.2. *Let β_1 be a number (explicitly computable, see [24]) such that for all $\varphi \in \Phi$*

$$\alpha \left\| \frac{d\pi_q^0(\varphi)}{d\varphi^T} \right\|_1 + (1-\alpha) \sum_{i=1}^{p_q} \left\| \frac{dp_i(\varphi)}{d\varphi^T} \right\|_1 \le \beta_1. \tag{34}$$

Assume that Method (29) with $N_1 = \left\lceil \frac{1}{\alpha} \ln \frac{24\beta_1 r}{\alpha \delta_2} \right\rceil - 1$ is used for every $q \in Q$ to calculate the vector $\tilde{\pi}_q(\varphi, N_1)$ and Method (32), (33) with $N_2 =$

$\left\lceil \frac{1}{\alpha} \ln \frac{8\beta_1 r}{\alpha \delta_2} \right\rceil - 1$ *is used for every* $q \in Q$ *to calculate the matrix* $\tilde{\Pi}_q(\varphi, N_2)$ (33). *Then the vector*

$$\tilde{g}(\varphi, \delta_2) = \frac{2}{|Q|} \sum_{q=1}^{|Q|} \left(\tilde{\Pi}_q(\varphi, N_2) \right)^T A_q^T (A_q \tilde{\pi}_q(\varphi, N_1))_+ \tag{35}$$

satisfies

$$\|\tilde{g}(\varphi, \delta_2) - \nabla f(\varphi)\|_\infty \le \delta_2. \tag{36}$$

Moreover the calculation of $\tilde{g}(\varphi, \delta_2)$ *requires no more than* $|Q|(10mps + 3psN_1 + 3mpsN_2 + 7r)$ *a.o. and not more than* $4ps + 4mp + r$ *memory items.*

As we see, there is an inexact oracle available for the considered supervised learning problem. Thus, in the next subsections, we consider a general problem with intexact oracle and solve it by zero-order and first-order methods.

2.2.2 Solving the learning problem by zero-order method

First, we consider a general zero-order method with inexact function evaluations and then we apply it to solve the learning problem. Let $\mathcal{E}$ be an m-dimensional vector space. First, we consider a general function $f(\cdot) : \mathcal{E} \to \mathbb{R}$ and denote its argument by x or y to avoid confusion with the above text. We denote the value of linear function $g \in \mathcal{E}^*$ at $x \in \mathcal{E}$ by $\langle g, x \rangle$. We choose some norm $\|\cdot\|$ in $\mathcal{E}$ and say that $f \in C_L^{1,1}(\|\cdot\|)$ iff

$$|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \le \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in \mathcal{E}. \tag{37}$$

The problem of our interest is to find $\min_{x \in X} f(x)$, where $f \in C_L^{1,1}(\|\cdot\|)$, X is a closed convex set and there exists a number $D \in (0, +\infty)$ such that $\mathrm{diam} X := \max_{x,y \in X} \|x - y\| \le D$. Also we assume that the inexact zero-order oracle for $f(x)$ returns a value $\widetilde{f}(x, \delta) = f(x) + \tilde{\delta}(x)$, where $\tilde{\delta}(x)$ is the error satisfying for some $\delta > 0$ (which is known) $|\tilde{\delta}(x)| \le \delta$ for all $x \in X$. Let $x^* \in \arg\min_{x \in X} f(x)$. Denote $f^* = \min_{x \in X} f(x)$.

Unlike [17], we define the biased gradient-free oracle $g_\tau(x, \delta) = \frac{m}{\tau}(\widetilde{f}(x + \tau\xi, \delta) - \widetilde{f}(x, \delta))\xi$, where ξ is a random vector uniformly distributed over the unit sphere $\mathcal{S} = \{t \in \mathbb{R}^m : \|t\|_2 = 1\}$, τ is a smoothing parameter.

Algorithm 4 below is the variation of the projected gradient descent method. Here $\Pi_X(x)$ denotes the Euclidean projection of a point x onto the set X.

Next theorem gives the convergence rate of Algorithm 4. Denote by $\Xi_k = (\xi_0, \dots, \xi_k)$ the history of realizations of the vector ξ generated on each iteration of the algorithm.

Algorithm 4 Gradient-type method

Input: Point $x_0 \in X$, stepsize $h > 0$, number of steps M.
Set $k = 0$.
repeat
  Generate ξ_k and calculate corresponding $g_\tau(x_k, \delta)$.
  Calculate $x_{k+1} = \Pi_X(x_k - h g_\tau(x_k, \delta))$.
  Set $k = k + 1$.
until $k > M$
Output: The point $y_M = \arg\min_x \{f(x) : x \in \{x_0, \dots, x_M\}\}$.

Theorem 2.5. *Let $f \in C_L^{1,1}(\|\cdot\|_2)$ and convex. Assume that $x^* \in \mathrm{int}X$, and the sequence x_k is generated by Algorithm 4 with $h = \frac{1}{8mL}$. Then for any $M \geq 0$, we have*

$$\mathbb{E}_{\Xi_{M-1}} f(y_M) - f^* \leq \frac{8mLD^2}{M+1} + \frac{\tau^2 L(m+8)}{8} + \frac{\delta m D}{4\tau} + \frac{\delta^2 m}{L\tau^2}. \tag{38}$$

Algorithm 5 Gradient-free method for Problem (28)

Input: Point $\varphi_0 \in \Phi$, L – Lipschitz constant for the function $f(\varphi)$ on Φ, accuracy $\varepsilon > 0$.
Define $M = \left\lceil 128m\frac{LR^2}{\varepsilon}\right\rceil$, $\delta = \frac{\varepsilon^{\frac{3}{2}}\sqrt{2}}{16mR\sqrt{L(m+8)}}$, $\tau = \sqrt{\frac{2\varepsilon}{L(m+8)}}$.
Set $k = 0$.
repeat
  Generate random vector ξ_k uniformly distributed over a unit Euclidean sphere $\mathcal{S}$ in R^m.
  Calculate $\widetilde{f}(\varphi_k + \tau\xi_k, \delta)$, $\widetilde{f}(\varphi_k, \delta)$ using Lemma 2.1 with $\delta_1 = \delta$.
  Calculate $g_\tau(\varphi_k, \delta) = \frac{m}{\tau}(\widetilde{f}(\varphi_k + \tau\xi_k, \delta) - \widetilde{f}(\varphi_k, \delta))\xi_k$.
  Calculate $\varphi_{k+1} = \Pi_\Phi\left(\varphi_k - \frac{1}{8mL} g_\tau(\varphi_k, \delta)\right)$.
  Set $k = k + 1$.
until $k > M$
Output: The point $\hat{\varphi}_M = \arg\min_\varphi \{f(\varphi) : \varphi \in \{\varphi_0, \dots, \varphi_M\}\}$.

Next, we apply the above method to solve the learning problem (28). The resulting algorithm is listed as Algorithm 5.

Theorem 2.6. *Assume that the set Φ in (28) is chosen in a way such that $f(\varphi)$ is convex on Φ and some $\varphi^* \in \arg\min_{\varphi\in\Phi} f(\varphi)$ belongs also to $\mathrm{int}\Phi$. Then the mean total number of arithmetic operations of the Algorithm 5 for the accuracy ε (i.e. for the inequality $\mathbb{E}_{\Xi_{M-1}} f(\hat{\varphi}_M) - f(\varphi^*) \leq \varepsilon$ to hold) is no more than*

$$768mps|Q|\frac{LR^2}{\varepsilon}\left(m + \frac{1}{\alpha}\ln\frac{128mrR\sqrt{L(m+8)}}{\varepsilon^{3/2}\sqrt{2}} + 6r\right).$$

2.2.3 Solving the learning problem by first-order method

First we consider a general first-order method with inexact function values and inexact gradient, and then we apply it to solve the learning problem. Let $\mathcal{E}$ be a finite-dimensional real vector space and $\mathcal{E}^*$ be its dual. We denote the value of linear function $g \in \mathcal{E}^*$ at $x \in \mathcal{E}$ by $\langle g, x \rangle$. Let $\|\cdot\|$ be some norm on $\mathcal{E}$, $\|\cdot\|_*$ be its dual. Our problem of interest in this subsection is a *composite optimization* problem of the form

$$\min_{x \in X} \{\psi(x) := f(x) + h(x)\}, \tag{39}$$

where $X \subset \mathcal{E}$ is a closed convex set, $h(x)$ is a simple convex function, e.g. $\|x\|_1$. We assume that $f(x)$ is a general function endowed with an inexact first-order oracle in the following sense. There exists a number $L \in (0, +\infty)$ such that for any $\delta \geq 0$ and any $x \in X$ one can calculate $\widetilde{f}(x, \delta) \in \mathbb{R}$ and $\tilde{g}(x, \delta) \in \mathcal{E}^*$ satisfying

$$|f(y) - (\widetilde{f}(x, \delta) - \langle \tilde{g}(x, \delta), y - x \rangle)| \leq \frac{L}{2}\|x - y\|^2 + \delta. \tag{40}$$

for all $y \in X$. The constant L can be considered as "Lipschitz constant" because for the exact first-order oracle for a function $f \in C_L^{1,1}(\|\cdot\|)$ (40) holds with $\delta = 0$. This is a generalization of the concept of (δ, L)-oracle considered in [26] for convex problems.

We choose a *prox-function* $d(x)$ which is continuously differentiable and 1-strongly convex on X with respect to $\|\cdot\|$. This means that for any $x, y \in X$ $d(y) - d(x) - \langle \nabla d(x), y - x \rangle \geq \frac{1}{2}\|y - x\|^2$. We define also the corresponding *Bregman distance* $V(x, z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$.

Theorem 2.7. *Assume that $f(x)$ is endowed with the inexact first-order oracle in the sense of* (40) *and that there exists a number $\psi^* > -\infty$ such that $\psi(x) \geq \psi^*$ for all $x \in X$. Then after M iterations of Algorithm* (6) *it holds that*

$$\|M_K(x_K - x_{K+1})\|^2 \leq \frac{4L(\psi(x_0) - \psi^*)}{M + 1} + \frac{\varepsilon}{2}. \tag{41}$$

Moreover, the total number of inner steps is no more than $M + \log_2 \frac{2L}{L_0}$.

Next we apply the general method to the learning problem. We set $\mathcal{E} = R^m$ and $\|\cdot\| = \|\cdot\|_2$, choose the prox-function $d(\varphi) = \frac{1}{2}\|\varphi\|_2^2$ and $V(\varphi, \omega) = \frac{1}{2}\|\varphi - \omega\|_2^2$. Algorithm 7 is a formal record of the algorithm.

Theorem 2.8. *The total number of arithmetic operations in Algorithm 7 for the accuracy ε (i.e. for the inequality $\|M_K(\varphi_K - \varphi_{K+1})\|_2^2 \leq \varepsilon$ to hold) is no more than*

$$\left(\frac{8L(f(\varphi_0) - f^*)}{\varepsilon} + \log_2 \frac{2L}{L_0}\right) \cdot \left(7r|Q| + \frac{6mps|Q|}{\alpha} \ln \frac{1024\beta_1 rRL\sqrt{m}}{\alpha\varepsilon}\right).$$

Algorithm 6 Adaptive projected gradient algorithm

Input: Point $x_0 \in X$, number $L_0 > 0$.
Set $k = 0$, $z = +\infty$.
repeat
  Set $M_k = L_k$, flag $= 0$.
  repeat
    Set $\delta = \frac{\varepsilon}{16 M_k}$. Calculate $\widetilde{f}(x_k, \delta)$ and $\tilde{g}(x_k, \delta)$.
    $w_k = \arg\min_{x \in Q} \left\{ \langle \tilde{g}(x_k, \delta), x \rangle + M_k V(x, x_k) + h(x) \right\}$
    If the inequality

$$\widetilde{f}(w_k, \delta) \leq \widetilde{f}(x_k, \delta) + \langle \tilde{g}(x_k, \delta), w_k - x_k \rangle + \frac{M_k}{2} \|w_k - x_k\|^2 + \frac{\varepsilon}{8 M_k}$$

    holds, set flag $= 1$. Otherwise set $M_k = 2M_k$.
  until flag $= 1$
  Set $x_{k+1} = w_k$, $L_{k+1} = \frac{M_k}{2}$.
  If $\|M_k(x_k - x_{k+1})\| < z$, set $z = \|M_k(x_k - x_{k+1})\|$, $K = k$.
  Set $k = k + 1$.
until $z \leq \varepsilon$
Output: The point x_{K+1}.

2.3 An accelerated directional derivative method for smooth stochastic convex optimization.

In this section we consider directional derivatives methods with inexact oracle for stochastic convex optimization. The results of this subsection are published in [27]. We consider the following optimization problem

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) := \mathbb{E}_\xi [F(x, \xi)] = \int_{\mathcal{X}} F(x, \xi) dP(x) \right\}, \tag{42}$$

where ξ is a random vector with probability distribution $P(\xi)$, $\xi \in \mathcal{X}$, and for P-almost every $\xi \in \mathcal{X}$, the function $F(x, \xi)$ is closed and convex. Moreover, we assume that, for P almost every ξ, the function $F(x, \xi)$ has gradient $g(x, \xi)$, which is $L(\xi)$-Lipschitz continuous with respect to the Euclidean norm and there exists $L_2 \geqslant 0$ such that $\sqrt{\mathbb{E}_\xi L(\xi)^2} \leqslant L_2 < +\infty$. Under this assumptions, $\mathbb{E}_\xi g(x, \xi) = \nabla f(x)$ and f has L_2-Lipschitz continuous gradient with respect to the Euclidean norm. Also we assume that

$$\mathbb{E}_\xi [\|g(x, \xi) - \nabla f(x)\|_2^2] \leqslant \sigma^2, \tag{43}$$

where $\| \cdot \|_2$ is the Euclidean norm.

Finally, we assume that an optimization procedure, given a point $x \in \mathbb{R}^n$, direction $e \in S_2(1)$ and ξ independently drawn from P, can obtain a noisy

Algorithm 7 Adaptive gradient method for Problem (28)

Input: Point $\varphi_0 \in \Phi$, number $L_0 > 0$, accuracy $\varepsilon > 0$.
Set $k = 0$, $z = +\infty$.
repeat
  Set $M_k = L_k$, flag $= 0$.
  repeat
    Set $\delta_1 = \frac{\varepsilon}{32M_k}$, $\delta_2 = \frac{\varepsilon}{64M_kR\sqrt{m}}$.
    Calculate $\widetilde{f}(\varphi_k, \delta_1)$ using Lemma 2.1 and $\tilde{g}(\varphi_k, \delta_2)$ using Lemma 2.2.
    Find
$$\omega_k = \arg\min_{\varphi\in\Phi} \left\{ \langle \tilde{g}(\varphi_k, \delta_2), \varphi\rangle + \frac{M_k}{2}\|\varphi - \varphi_k\|_2^2. \right\}$$
    Calculate $\widetilde{f}(\omega_k, \delta_1)$ using Lemma 2.1.
    If the inequality
$$\widetilde{f}(\omega_k, \delta_1) \leq \widetilde{f}(\varphi_k, \delta_1) + \langle \tilde{g}(\varphi_k, \delta_2), \omega_k - \varphi_k\rangle + \frac{M_k}{2}\|\omega_k - \varphi_k\|_2^2 + \frac{\varepsilon}{8M_k}$$
    holds, set flag $= 1$. Otherwise set $M_k = 2M_k$.
  until flag $= 1$
  Set $\varphi_{k+1} = \omega_k$, $L_{k+1} = \frac{M_k}{2}$, .
  If $\|M_k(\varphi_k - \varphi_{k+1})\|_2 < z$, set $z = \|M_k(\varphi_k - \varphi_{k+1})\|_2$, $K = k$.
  Set $k = k + 1$.
until $z \leq \varepsilon$
Output: The point φ_{K+1}.

stochastic approximation $\widetilde{f}'(x, \xi, e)$ for the directional derivative $\langle g(x,\xi), e\rangle$:

$$\begin{aligned} \widetilde{f}'(x,\xi,e) &= \langle g(x,\xi), e\rangle + \zeta(x,\xi,e) + \eta(x,\xi,e), \\ \mathbb{E}_\xi(\zeta(x,\xi,e))^2 &\leqslant \Delta_\zeta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \\ |\eta(x,\xi,e)| &\leqslant \Delta_\eta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \text{ a.s. in } \xi, \end{aligned} \tag{44}$$

where $S_2(1)$ is the Euclidean sphere of radius one with the center at the point zero and the values Δ_ζ, Δ_η are controlled and can be made as small as it is desired. Note that we use the smoothness of $F(\cdot,\xi)$ to write the directional derivative as $\langle g(x,\xi), e\rangle$, but we *do not assume* that the whole stochastic gradient $g(x,\xi)$ is available. We choose a *prox-function* $d(x)$ which is continuous, convex on $\mathbb{R}^n$ and is 1-strongly convex on $\mathbb{R}^n$ with respect to $\|\cdot\|_p$, i.e., for any $x, y \in \mathbb{R}^n$ $d(y) - d(x) - \langle \nabla d(x), y - x\rangle \geq \frac{1}{2}\|y - x\|_p^2$. Without loss of generality, we assume that $\min_{x\in\mathbb{R}^n} d(x) = 0$. We define also the corresponding *Bregman divergence* $V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z\rangle$,

$x, z \in \mathbb{R}^n$. For the case $p = 1$, we choose the following prox-function [28]

$$d(x) = \frac{\mathrm{e}n^{(\kappa-1)(2-\kappa)/\kappa}\ln n}{2}\|x\|_\kappa^2, \quad \kappa = 1 + \frac{1}{\ln n} \tag{45}$$

and, for the case $p = 2$, we choose the prox-function to be the squared Euclidean norm $d(x) = \frac{1}{2}\|x\|_2^2$.

Based on the noisy stochastic observations (44) of the directional derivative, we form the following stochastic approximation of $\nabla f(x)$

$$\widetilde{\nabla}^m f(x) = \frac{1}{m}\sum_{i=1}^{m} \widetilde{f}'(x, \xi_i, e)e, \tag{46}$$

where $e \in RS_2(1)$, ξ_i, $i = 1, ..., m$ are independent realizations of ξ, m is the *batch size*.

2.3.1 Algorithms and main results for convex problems

The Accelerated Randomized Directional Derivative (ARDD) method is listed as Algorithm 8.

Algorithm 8 Accelerated Randomized Directional Derivative (ARDD) method

---

Require: x_0 —starting point; $N \geqslant 1$ — number of iterations; $m \geqslant 1$ — batch size.

Ensure: point y_N.

1: $y_0 \leftarrow x_0$, $z_0 \leftarrow x_0$.
2: **for** $k = 0, \ldots, N-1$. **do**
3: $\alpha_{k+1} \leftarrow \frac{k+2}{96n^2\rho_n L_2}$, $\tau_k \leftarrow \frac{1}{48\alpha_{k+1}n^2\rho_n L_2} = \frac{2}{k+2}$.
4: Generate $e_{k+1} \in RS_2(1)$ independently from previous iterations and ξ_i, $i = 1, ..., m$ – independent realizations of ξ.
5: $\widetilde{\nabla}^m f(x_{k+1}) = \frac{1}{m}\sum\limits_{i=1}^{m} \widetilde{f}'(x_{k+1}, \xi_i, e)e$.
6: $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k)y_k$.
7: $y_{k+1} \leftarrow x_{k+1} - \frac{1}{2L_2}\widetilde{\nabla}^m f(x_{k+1})$.
8: $z_{k+1} \leftarrow \operatorname*{argmin}\limits_{z\in\mathbb{R}^n}\left\{\alpha_{k+1}n\left\langle \widetilde{\nabla}^m f(x_{k+1}),\, z - z_k\right\rangle + V[z_k](z)\right\}$.
9: **end for**
10: **return** y_N

Theorem 2.9. *Let ARDD method be applied to solve problem* (42). *Then*

$$\begin{aligned}\mathbb{E}[f(y_N)] - f(x^*) \leqslant &\frac{384\Theta_p n^2\rho_n L_2}{N^2} + \frac{4N}{nL_2}\cdot\frac{\sigma^2}{m} + \frac{61N}{24L_2}\Delta_\zeta + \frac{122N}{3L_2}\Delta_\eta^2 \\ &+ \frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2,\end{aligned} \tag{47}$$

where $\Theta_p = V[z_0](x^*)$ *is defined by the chosen proximal setup and* $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\ldots,e_N,\xi_{1,1},\ldots,\xi_{N,m}}[\cdot]$.

The appropriate choice of the ARDD method parameters is given in Table 1.

	$p=1$	$p=2$
N	$\sqrt{\frac{n \ln n L_2 \Theta_1}{\varepsilon}}$	$\sqrt{\frac{n^2 L_2 \Theta_2}{\varepsilon}}$
m	$\max\left\{1, \sqrt{\frac{\ln n}{n}} \cdot \frac{\sigma^2}{\varepsilon^{3/2}} \cdot \sqrt{\frac{\Theta_1}{L_2}}\right\}$	$\max\left\{1, \frac{\sigma^2}{\varepsilon^{3/2}} \cdot \sqrt{\frac{\Theta_2}{L_2}}\right\}$
Δ_ζ	$\min\left\{n(\ln n)^2 L_2^2 \Theta_1, \frac{\varepsilon^2}{n\Theta_1}, \frac{\varepsilon^{\frac{3}{2}}}{\sqrt{n \ln n}} \cdot \sqrt{\frac{L_2}{\Theta_1}}\right\}$	$\min\left\{n^3 L_2^2 \Theta_2, \frac{\varepsilon^2}{n\Theta_2}, \frac{\varepsilon^{\frac{3}{2}}}{n} \cdot \sqrt{\frac{L_2}{\Theta_2}}\right\}$
Δ_η	$\min\left\{\sqrt{n} \ln n L_2 \sqrt{\Theta_1}, \frac{\varepsilon}{\sqrt{n\Theta_1}}, \frac{\varepsilon^{\frac{3}{4}}}{\sqrt[4]{n \ln n}} \cdot \sqrt[4]{\frac{L_2}{\Theta_1}}\right\}$	$\min\left\{n^{\frac{3}{2}} L_2 \sqrt{\Theta_2}, \frac{\varepsilon}{\sqrt{n\Theta_2}}, \frac{\varepsilon^{\frac{3}{4}}}{\sqrt{n}} \cdot \sqrt[4]{\frac{L_2}{\Theta_2}}\right\}$
Calls	$\max\left\{\sqrt{\frac{n \ln n L_2 \Theta_1}{\varepsilon}}, \frac{\sigma^2 \Theta_1 \ln n}{\varepsilon^2}\right\}$	$\max\left\{\sqrt{\frac{n^2 L_2 \Theta_2}{\varepsilon}}, \frac{\sigma^2 \Theta_2 n}{\varepsilon^2}\right\})$

Table 1: Algorithm 8 parameters for the cases $p = 1$ and $p = 2$.

The Randomized Directional Derivative (RDD) method is listed as Algorithm 9.

Algorithm 9 Randomized Directional Derivative (RDD) method

Require: x_0 —starting point; $N \geqslant 1$ — number of iterations; $m \geqslant 1$ — batch size.

Ensure: point $\bar{x}_N$.

for $k = 0, \ldots, N-1$. **do**

  $\alpha \leftarrow \frac{1}{48 n \rho_n L_2}$.

  Generate $e_{k+1} \in RS_2(1)$ independently from previous iterations and ξ_i, $i = 1, ..., m$ – independent realizations of ξ.

  $\widetilde{\nabla}^m f(x_k) = \frac{1}{m} \sum\limits_{i=1}^{m} \widetilde{f}'(x_k, \xi_i, e) e.$

  $x_{k+1} \leftarrow \operatorname*{argmin}\limits_{x \in \mathbb{R}^n} \left\{\alpha n \left\langle \widetilde{\nabla}^m f(x_k), x - x_k \right\rangle + V[x_k](x)\right\}.$

end for

return $\bar{x}_N \leftarrow \frac{1}{N} \sum\limits_{k=0}^{N-1} x_k$

Theorem 2.10. *Let RDD method be applied to solve problem* (42). *Then*

$$\mathbb{E}[f(\bar{x}_N)] - f(x_*) \leqslant \frac{384n\rho_n L_2 \Theta_p}{N} + \frac{2}{L_2}\frac{\sigma^2}{m} + \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2$$
$$+ \frac{8\sqrt{2n\Theta_p}}{N}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2, \qquad (48)$$

where $\Theta_p = V[z_0](x^*)$ *is defined by the chosen proximal setup and* $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\dots,e_N,\xi_{1,1},\dots,\xi_{N,m}}[\cdot]$.

The appropriate choice of the RDD method parameters is given in Table 2.

	$p=1$	$p=2$
N	$\frac{L_2\Theta_1 \ln n}{\varepsilon}$	$\frac{nL_2\Theta_2}{\varepsilon}$
m	$\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}$	$\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}$
Δ_ζ	$\min\left\{\frac{(\ln n)^2}{n}L_2^2\Theta_1, \frac{\varepsilon^2}{n\Theta_1}, \frac{\varepsilon L_2}{n}\right\}$	$\min\left\{nL_2^2\Theta_2, \frac{\varepsilon^2}{n\Theta_2}, \frac{\varepsilon L_2}{n}\right\}$
Δ_η	$\min\left\{\frac{\ln n}{\sqrt{n}}L_2\sqrt{\Theta_1}, \frac{\varepsilon}{\sqrt{n\Theta_1}}, \sqrt{\frac{\varepsilon L_2}{n}}\right\}$	$\min\left\{\sqrt{n}L_2\sqrt{\Theta_2}, \frac{\varepsilon}{\sqrt{n\Theta_2}}, \sqrt{\frac{\varepsilon L_2}{n}}\right\}$
Nm	$\max\left\{\frac{L_2\Theta_1 \ln n}{\varepsilon}, \frac{\sigma^2\Theta_1 \ln n}{\varepsilon^2}\right\}$	$\max\left\{\frac{nL_2\Theta_2}{\varepsilon}, \frac{n\sigma^2\Theta_2}{\varepsilon^2}\right\}$

Table 2: Algorithm 9 parameters for the cases $p=1$ and $p=2$.

2.3.2 Algorithms and main results for strongly convex problems.

To obtain faster rates, we assume additionally that f is μ_p-strongly convex w.r.t. p-norm. Our algorithms and proofs rely on the following assumption. Let x_* be some fixed point and x be a random point such that $\mathbb{E}_x\left[\|x - x_*\|_p^2\right] \leqslant R_p^2$, then

$$\mathbb{E}_x d\left(\frac{x - x_*}{R_p}\right) \leqslant \frac{\Omega_p}{2}, \qquad (49)$$

where $\mathbb{E}_x$ denotes the expectation with respect to random vector x and Ω_p is defined as follows. For $p=1$ and our choice of the prox-function (45), $\Omega_p = en^{(\kappa-1)(2-\kappa)/\kappa}\ln n = O(\ln n)$ with $\kappa = 1 + \frac{1}{\ln n}$, see [6, 29]. For $p=2$ and our choice of the prox-function, $\Omega_p = 1$. Our Accelerated Randomized Directional Derivative method for strongly convex problems (ARDDsc) is listed as Algorithm 10.

Algorithm 10 Accelerated Randomized Directional Derivative method for strongly convex functions (ARDDsc)

Require: x_0 —starting point s.t. $\|x_0 - x_*\|_p^2 \le R_p^2$; $K \geqslant 1$ — number of iterations; μ_p – strong convexity parameter.

Ensure: point u_K.

Set $N_0 = \left\lceil \sqrt{\frac{8aL_2\Omega_p}{\mu_p}} \right\rceil$, where $a = 384n^2\rho_n$.

for $k = 0, \dots, K-1$ **do**

$m_k := \max\left\{1, \left\lceil \frac{32\sigma^2 N_0 2^k}{nL_2\mu_p R_p^2} \right\rceil \right\}, \quad R_k^2 := R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right),$

Set $d_k(x) = R_k^2 d\left(\frac{x-u_k}{R_k}\right)$.

Run ARDD with starting point u_k and prox-function $d_k(x)$ for N_0 steps with batch size m_k.

Set $u_{k+1} = y_{N_0}$, $k = k+1$.

end for

return u_K

Theorem 2.11. *Let f in problem (42) be μ_p-strongly convex and ARDDsc method be applied to solve this problem. Then*

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta. \tag{50}$$

where $\Delta = \frac{61N_0}{24L_2}\Delta_\zeta + \frac{122N_0}{3L_2}\Delta_\eta^2 + \frac{12\sqrt{2nR_p^2\Omega_p}}{N_0^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2.$
Moreover, under an appropriate choice of Δ_ζ and Δ_η s.t. $2\Delta \leqslant \varepsilon/2$, the oracle complexity to achieve ε-accuracy of the solution is

$$\widetilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2\Omega_p}{\mu_p}}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

The appropriate choice of the ARDDsc method parameters is given in Table 3.

	$p = 1$	$p = 2$
Δ_ζ	$\min\left\{\varepsilon\sqrt{\frac{L_2\mu_1}{n\ln n\Omega_1}}, \varepsilon^2\frac{n(\ln n)^2L_2^2\Omega_1}{R_1^2\mu_1^2}, \varepsilon\cdot\frac{\mu_1}{n\Omega_1}\right\}$	$\min\left\{\varepsilon\sqrt{\frac{L_2\mu_2}{n^2\Omega_2}}, \varepsilon^2\frac{n^3L_2^2\Omega_2}{R_2^2\mu_2^2}, \varepsilon\cdot\frac{\mu_2}{n\Omega_2}\right\}$
Δ_η	$\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_1}{n\ln n\Omega_1}}, \varepsilon\frac{\sqrt{n}\ln nL_2\sqrt{\Omega_1}}{R_1\mu_1}, \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_1}{n\Omega_1}}\right\}$	$\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_2}{n^2\Omega_2}}, \varepsilon\frac{\sqrt{n^3}L_2\sqrt{\Omega_2}}{R_2\mu_2}, \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_2}{n\Omega_2}}\right\}$
Calls	$\max\left\{\sqrt{\frac{n\ln nL_2\Omega_1}{\mu_1}}\log_2\frac{\mu_1R_1^2}{\varepsilon}, \frac{\sigma^2\Omega_1\ln n}{\mu_1\varepsilon}\right\}$	$\max\left\{n\sqrt{\frac{L_2\Omega_2}{\mu_2}}\log_2\frac{\mu_2R_2^2}{\varepsilon}, \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}$

Table 3: Algorithm 10 parameters for the cases $p = 1$ and $p = 2$.

Our Randomized Directional Derivative method for strongly convex problems (RDDsc) is listed as Algorithm 11.

Algorithm 11 Randomized Directional Derivative method for strongly convex functions (RDDsc)

Require: x_0 —starting point s.t. $\|x_0 - x_*\|_p^2 \le R_p^2$; $K \geqslant 1$ — number of iterations; μ_p – strong convexity parameter.

Ensure: point u_K.

1: Set $N_0 = \left\lceil \frac{8aL_2\Omega_p}{\mu_p} \right\rceil$, where $a = 384n\rho_n$.

2: **for** $k = 0, \ldots, K-1$ **do**

3: $\quad m_k := \max\left\{1, \left\lceil \frac{16\sigma^2 2^k}{L_2\mu_p R_p^2} \right\rceil\right\}, \quad R_k^2 := R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right),$

4: $\quad$ Set $d_k(x) = R_k^2 d\left(\frac{x-u_k}{R_k}\right)$.

5: $\quad$ Run RDD with starting point u_k and prox-function $d_k(x)$ for N_0 steps with batch size m_k.

6: $\quad$ Set $u_{k+1} = y_{N_0}$, $k = k+1$.

7: **end for**

8: **return** u_K

Theorem 2.12. *Let f in problem (42) be μ_p-strongly convex and RDDsc method be applied to solve this problem. Then*

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta. \tag{51}$$

where $\Delta = \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 + \frac{8\sqrt{2nR_p^2\Omega_p}}{N_0}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2.$
Moreover, under an appropriate choice of Δ_ζ and Δ_η s.t. $2\Delta \leqslant \varepsilon/2$, the oracle complexity to achieve ε-accuracy of the solution is

$$\widetilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2\Omega_p}{\mu_p}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

The appropriate choice of the RDDsc method parameters is given in Table 4.

	$p=1$	$p=2$
Δ_ζ	$\min\left\{\frac{\varepsilon L_2}{n}, \varepsilon^2\frac{(\ln n)^2 L_2^2}{nR_1^2\mu_1^2}, \varepsilon\frac{\mu_1}{n\Omega_1}\right\}$	$\min\left\{\frac{\varepsilon L_2}{n}, \varepsilon^2\frac{nL_2^2}{R_2^2\mu_2^2}, \varepsilon\frac{\mu_2}{n\Omega_2}\right\}$
Δ_η	$\min\left\{\sqrt{\frac{\varepsilon L_2}{n}}, \varepsilon\frac{\ln nL_2}{\sqrt{n}R_1\mu_1}, \sqrt{\varepsilon\frac{\mu_1}{n\Omega_1}}\right\}$	$\min\left\{\sqrt{\frac{\varepsilon L_2}{n}}, \varepsilon\frac{\sqrt{n}L_2}{R_2\mu_2}, \sqrt{\varepsilon\frac{\mu_2}{n\Omega_2}}\right\}$
Calls	$\max\left\{\frac{L_2\Omega_1\ln n}{\mu_1}\log_2\frac{\mu_1 R_1^2}{\varepsilon}, \frac{\sigma^2\Omega_1}{\mu_1\varepsilon}\right\}$	$\max\left\{\frac{nL_2\Omega_2}{\mu_2}\log_2\frac{\mu_2 R_2^2}{\varepsilon}, \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}$

Table 4: Algorithm 11 parameters for the cases $p=1$ and $p=2$.

3 Primal-dual methods

In this section, we focus on the developed primal-dual first-order methods for convex problems with linear constraints.

3.1 Primal-dual methods for solving infinite-dimensional games

The results of this subsection are published in [30]. Consider two moving objects with dynamics given by the following equations:

$$\dot{x}(t) = A_x(t)x(t) + B(t)u(t), \dot{y}(t) = A_y(t)y(t) + C(t)v(t),$$
$$(x(0), y(0)) = (x_0, y_0). \tag{52}$$

Here $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^m$ are the phase vectors of these objects, $u(t)$ is the control of the first object (pursuer), and $v(t)$ is the control of the second object (evader). Matrices $A_x(t), A_y(t), B(t)$, and $C(t)$ are continuous and have appropriate sizes. The system is considered on the time interval $[0, \theta]$. Controls are restricted in the following way $u(t) \in P \subseteq \mathbb{R}^p$, $v(t) \in Q \subseteq \mathbb{R}^q \quad \forall t \in [0, \theta]$. We assume that P, Q are closed, convex sets.

The goal of the pursuer is to minimize the value of the functional:

$$F(u, v) + \Phi(x(\theta), y(\theta)) := \int_0^\theta \tilde{F}(\tau, u(\tau), v(\tau))d\tau + \Phi(x(\theta), y(\theta)). \tag{53}$$

The goal of the evader is the opposite. We need to find an optimal guaranteed result for each object, which leads to the problem of finding the saddle point of the above functional. We assume the following:

- $u(\cdot) \in L^2([0, \theta], \mathbb{R}^p)$, and $v(\cdot) \in L^2([0, \theta], \mathbb{R}^q)$ (for the notation simplification we denote $L^2([0, \theta], \mathbb{R}^p)$ by L^2_p and $L^2([0, \theta], \mathbb{R}^q)$ by L^2_q),
- the saddle point in this class of strategies exists,
- the function $F(u, v)$ is upper semi-continuous in v and lower semi-continuous in u,
- $\Phi(x, y)$ is continuous.

Denote by $V_x(t, \tau)$ the transition matrix of the first system in (52). It is the unique solution of the following matrix Cauchy problem

$$\frac{dV_x(t, \tau)}{dt} = A_x(t)V_x(t, \tau), \quad t \geq \tau, \quad V_x(\tau, \tau) = E.$$

Here E is the identity matrix. If the matrix $A_x(t)$ is constant, then $V_x(t, \tau) = e^{(t-\tau)A}$.

If we solve the first differential equation in (52), then we can express $x(\theta)$ as a result of the application of the linear operator $\mathcal{B}: L_p^2 \to \mathbb{R}^n$:

$$x(\theta) = V_x(\theta, 0)x_0 + \int_0^\theta V_x(\theta, \tau)B(\tau)u(\tau)d\tau := \tilde{x}_0 + \mathcal{B}u. \tag{54}$$

Below, we will use the conjugate operator $\mathcal{B}^*$ for the operator $\mathcal{B}$. Let us find it explicitly. Let μ be a n-dimensional vector. Then

$$\langle \mu, \mathcal{B}u \rangle = \langle \mu, \int_0^\theta V_x(\theta, \tau)B(\tau)u(\tau)d\tau \rangle = \int_0^\theta \langle \mu, V_x(\theta, \tau)B(\tau)u(\tau) \rangle d\tau =$$

$$= \int_0^\theta \langle B^T(\tau)V_x^T(\theta, \tau)\mu, u(\tau) \rangle d\tau = \langle \mathcal{B}^* \mu, u \rangle.$$

Note that the vector $\zeta(t) = V_x^T(\theta, t)\mu$ is the solution of the following Cauchy problem:

$$\dot{\zeta}(t) = -A_x^T(t)\zeta(t), \quad \zeta(\theta) = \mu, \quad t \in [0, \theta].$$

So we can solve this ODE and find $\mathcal{B}^*\mu$ using the obtained solution $\zeta(t)$ as $\mathcal{B}^*\mu(t) = B^T(t)\zeta(t)$.

In the same way, we introduce the transition matrix $V_y(t, \tau)$ of the second system in (52), the operator $\mathcal{C}: L_q^2 \to \mathbb{R}^m$ defined by the formula $\mathcal{C}v := \int_0^\theta V_y(\theta, \tau)C(\tau)v(\tau)d\tau$, and the vector $\tilde{y}_0 := V_y(\theta, 0)y_0$. The adjoint operator $\mathcal{C}^*$ also can be computed using the solution of some ODE.

So below we study differential game problem in the following form:

$$\min_{u \in \mathcal{U}} \left[\max_{v \in \mathcal{V}} \{F(u, v) + \Phi(x, y) : y = \tilde{y}_0 + \mathcal{C}v\} : x = \tilde{x}_0 + \mathcal{B}u \right], \tag{55}$$

where

$$\mathcal{U} := \{u(\cdot) \in L_p^2 : u(t) \in P \quad \forall t \in [0, \theta]\}, \mathcal{V} := \{v(\cdot) \in L_q^2 : v(t) \in Q \quad \forall t \in [0, \theta]\}$$

are sets of admissible strategies of the players and $u \in \mathcal{U}$, $v \in \mathcal{V}$ mean $u(\cdot) \in \mathcal{U}$, $v(\cdot) \in \mathcal{V}$. Our goal is to introduce a computational method for finding an approximate solution of the problem (55).

First, we consider the problem (55) under two assumptions.

A1 The sets P and Q are bounded.

A2 In (53) the functional $F(\cdot, v)$ is convex for any fixed v, $F(u, \cdot)$ is concave for any fixed u, $\Phi(\cdot, y)$ is convex for any fixed y, and $\Phi(x, \cdot)$ is concave for any fixed x.

From **A1**, since the norms of the operators $\mathcal{B}, \mathcal{C}$ are bounded, $x(\theta), y(\theta)$ are also bounded and we can equivalently reformulate the problem (55) in the following way:

$$\min_{u \in \mathcal{U}, x \in X} \left[\max_{v \in \mathcal{V}, y \in Y} \{F(u, v) + \Phi(x, y) : y = \tilde{y}_0 + \mathcal{C}v\} : x = \tilde{x}_0 + \mathcal{B}u \right] =$$

$$\max_{v \in \mathcal{V}, y \in Y} \left[\min_{u \in \mathcal{U}, x \in X} \{F(u, v) + \Phi(x, y) : x = \tilde{x}_0 + \mathcal{B}u\} : y = \tilde{y}_0 + \mathcal{C}v \right], \tag{56}$$

where the sets X and Y are closed, convex and bounded. Let us introduce the spaces of dual variables $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$ corresponding to the linear constraints in the problem (56), and some norms $\|\cdot\|_\lambda$ and $\|\cdot\|_\mu$ in these spaces. We define the norms in the dual space in the standard way

$$\|s_\lambda\|_{\lambda,*} := \max\{\langle s_\lambda, \lambda\rangle : \|\lambda\|_\lambda \le 1\}, \quad \|s_\mu\|_{\mu,*} := \max\{\langle s_\mu, \mu\rangle : \|\mu\|_\mu \le 1\}.$$

Lemma 3.1. *Let the Assumptions* **A1**, **A2** *hold. Also assume that the function $F(u,v)$ is upper semi-continuous in v and lower semi-continuous in u, the function $\Phi(x,y)$ is continuous, and that the sets P and Q are convex and closed. Then the problem* (56) *is equivalent to the problem*

$$\begin{aligned} &\min\nolimits_\lambda \max\nolimits_\mu \{\min\nolimits_{u\in\mathcal{U}} \max\nolimits_{v\in\mathcal{V}} [F(u,v) - \langle \mu, \mathcal{B}u\rangle + \langle \lambda, \mathcal{C}v\rangle] \\ &+ \min\nolimits_{x\in X} \max\nolimits_{y\in Y} [\Phi(x,y) + \langle \mu, x\rangle - \langle \lambda, y\rangle] - \langle \mu, \tilde{x}_0\rangle + \langle \lambda, \tilde{y}_0\rangle\}, \end{aligned} \tag{57}$$

which we call the conjugate problem to (56).

We denote by $\psi(\lambda,\mu)$ the function, for which the goal in (57) is to find its saddle point.

3.1.1 Algorithm for convex-concave problem

We assume that we are given some prox-function $d_\lambda(\lambda)$ with prox-center λ_0, which is strongly convex with convexity parameter σ_λ in the given norm $\|\cdot\|_\lambda$. For μ we introduce the similar assumptions. Since (λ^*,μ^*) is the saddle point, (λ^*,μ^*) is a weak solution to the following variational inequality $\langle g(\lambda,\mu), (\lambda-\lambda^*, \mu-\mu^*)\rangle \ge 0, \quad \forall \lambda,\mu$, where $g(\lambda,\mu) := (\psi'_\lambda(\lambda,\mu), -\psi'_\mu(\lambda,\mu))$. We apply the method of Simple Dual Averages (SDA) from [31] for finding an approximate solution of the finite-dimensional problem (57). Let us choose some $\kappa \in]0,1[$. We consider a space of $z := (\lambda,\mu)$ with the norm

$$\|z\|_z := \sqrt{\kappa\sigma_\lambda \|\lambda\|_\lambda^2 + (1-\kappa)\sigma_\mu \|\mu\|_\mu^2}, \tag{58}$$

an oracle $g(z) := (g_\lambda(z), -g_\mu(z))$, a new prox-function $d(z) := \kappa d_\lambda(\lambda) + (1-\kappa)\, d_\mu(\mu)$, which is strongly convex with constant $\sigma_0 = 1$ with respect to the norm (58). We define $W := \mathbb{R}^m \times \mathbb{R}^n$. The conjugate norm for (58) is $\|g\|_{z,*} := \sqrt{\frac{1}{\kappa\sigma_\lambda}\|g_\lambda\|_{\lambda,*}^2 + \frac{1}{(1-\kappa)\sigma_\mu}\|g_\mu\|_{\mu,*}^2}$. So we have a uniform upper bound for the answers of the oracle $\|g(\lambda,\mu)\|_{z,*}^2 \le L^2 := \frac{L_\lambda^2}{\kappa\sigma_\lambda} + \frac{L_\mu^2}{(1-\kappa)\sigma_\mu}$, where $L_\lambda := \sqrt{\theta}\,\|\mathcal{C}\|_{\lambda,L_q^2} \operatorname{diam}_2 Q + \operatorname{diam}_{\lambda,*} Y + \|\tilde{y}_0\|_{\lambda,*}$ and $L_\mu := \sqrt{\theta}\,\|\mathcal{B}\|_{\mu,L_p^2} \operatorname{diam}_2 P + \operatorname{diam}_{\mu,*} X + \|\tilde{x}_0\|_{\mu,*}$.

The SDA method for solving (57) is the following

1. Initialization: Set $s_0 = 0$. Choose z_0, $\gamma > 0$.

2. Iteration ($k \geq 0$):

$$\begin{array}{ll} \text{Compute } g_k = g(z_k). \text{ Set } s_{k+1} = s_k + g_k. & \\ \beta_{k+1} = \gamma \hat{\beta}_{k+1}. \text{ Set } z_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1}). & \end{array} \tag{M1}$$

Here the sequence $\hat{\beta}_{k+1}$ is defined by relations $\hat{\beta}_0 = \hat{\beta}_1 = 1$, $\hat{\beta}_{i+1} = \hat{\beta}_i + \frac{1}{\hat{\beta}_i}$, for $i \geq 1$. The mapping $\pi_\beta(s)$ is defined in the following way $\pi_\beta(s) := \arg\min_{z \in W} \{-\langle s, z\rangle + \beta d(z)\}$.

We choose D_λ, D_μ such that $d_\lambda(\lambda_i) \leq D_\lambda$, $d_\mu(\mu_i) \leq D_\mu$ for all $i \geq 0$ and also, the pair (λ^*, μ^*) is an interior solution: $\mathfrak{B}^\lambda_{r/\sqrt{\kappa\sigma_\lambda}}(\lambda^*) \subseteq W_\lambda := \{\lambda : d_\lambda(\lambda) \leq D_\lambda\}$, and $\mathfrak{B}^\mu_{r/\sqrt{(1-\kappa)\sigma_\mu}}(\mu^*) \subseteq W_\mu := \{\mu : d_\mu(\mu) \leq D_\mu\}$ for some $r > 0$. Then we have $z^* := (\lambda^*, \mu^*) \in \mathcal{F}_D := \{z \in W : d(z) \leq D\}$ with $D := \kappa D_\lambda + (1-\kappa) D_\mu$ and $\mathfrak{B}^z_r(z^*) \subseteq \mathcal{F}_D$.

Let us introduce a gap function

$$\delta_k(D) := \max_z \left\{ \sum_{i=0}^{k} \langle g_i, z_i - z\rangle : z \in \mathcal{F}_D \right\}. \tag{59}$$

From the Theorem 2 in [31] we have

$$\frac{1}{k+1}\delta_k(D) \leq \frac{\hat{\beta}_{k+1}}{k+1}\left(\gamma D + \frac{L^2}{2\gamma}\right). \tag{60}$$

Denote

$$(\hat{u}_{k+1}, \hat{v}_{k+1}, \hat{x}_{k+1}, \hat{y}_{k+1}) := \frac{1}{k+1}\sum_{i=0}^{k}(u_i, v_i, x_i, y_i), \tag{61}$$

where (u_i, v_i), (x_i, y_i) are the saddle points at the point (λ_i, μ_i) in (57). We define a function

$$\begin{array}{l} \phi(u,x,v,y) := \min_\lambda \max_\mu \{F(u,v) + \Phi(x,y) + \langle \mu, x - \tilde{x}_0 - \mathcal{B}u\rangle + \\ +\langle \lambda, \mathcal{C}v + \tilde{y}_0 - y\rangle : d_\lambda(\lambda) \leq D_\lambda, d_\mu(\mu) \leq D_\mu\}. \end{array} \tag{62}$$

Since $d_\lambda(\lambda^*) \leq D_\lambda$, $d_\mu(\mu^*) \leq D_\mu$, and the conjugate problem is equivalent to the initial one, we conclude that the initial problem is equivalent to the problem

$$\min_{u \in \mathcal{U}, x \in X} \max_{v \in \mathcal{V}, y \in Y} \phi(u,x,v,y). \tag{63}$$

Let us introduce two auxiliary functions:

$$\xi(u,x) := \max_{v \in \mathcal{V}, y \in Y} \phi(u,x,v,y), \tag{64}$$

$$\eta(v,y) := \min_{u \in \mathcal{U}, x \in X} \phi(u,x,v,y). \tag{65}$$

Note that $\xi(u,x)$ is convex, $\eta(v,y)$ is concave, and $\xi(u,x) \geq \phi(u^*,x^*,v^*,y^*) \geq \eta(v,y)$ for all $u \in \mathcal{U}, v \in \mathcal{V}, x \in X, y \in Y$, where $\phi(u^*,x^*,v^*,y^*)$ is the solution to (63).

Theorem 3.1. *Let the assumptions* **A1** *and* **A2** *be true. Then the points* (61) *generated by the method* (M1) *satisfy:*

$$\xi(\hat{u}_{k+1}, \hat{x}_{k+1}) - \eta(\hat{v}_{k+1}, \hat{y}_{k+1}) \leq \frac{\hat{\beta}_{k+1}}{k+1}\left(\gamma D + \frac{L^2}{2\gamma}\right), \tag{66}$$

$$\begin{aligned} \|\tilde{x}_0 + \mathcal{B}\hat{u}_{k+1} - \hat{x}_{k+1}\|_{\mu,*} &\leq \tfrac{\hat{\beta}_{k+1}\sqrt{\sigma_\mu}}{r(k+1)}\left(\gamma D + \tfrac{L^2}{2\gamma}\right), \\ \|\tilde{y}_0 + \mathcal{C}\hat{v}_{k+1} - \hat{y}_{k+1}\|_{\lambda,*} &\leq \tfrac{\hat{\beta}_{k+1}\sqrt{\sigma_\lambda}}{r(k+1)}\left(\gamma D + \tfrac{L^2}{2\gamma}\right). \end{aligned} \tag{67}$$

3.1.2 Algorithm for strongly convex-concave problem

In this subsection, we consider the problem (55), under stronger assumptions and obtain faster convergence rates.

A3 The function $F(\cdot, v)$ is strongly convex for any fixed v with constant σ_{F_u} which does not depend on v, and function $F(u, \cdot)$ is strongly concave for any fixed u with constant σ_{F_v} which does not depend on u. Assume that:

$$\|\nabla_u F(u, v_1) - \nabla_u F(u, v_2)\|_{L_p^2} \leq L_{uv}\|v_1 - v_2\|_{L_q^2}, \tag{68}$$

$$\|\nabla_v F(u_1, v) - \nabla_v F(u_2, v)\|_{L_q^2} \leq L_{vu}\|u_1 - u_2\|_{L_p^2}. \tag{69}$$

A4 $\Phi(\cdot, y)$ is strongly convex for any fixed y with respect to the norm $\|\cdot\|_{\mu,*}$ with constant $\sigma_{\Phi x}$ which doesn't depend on y and $\Phi(x, \cdot)$ is strongly concave for any fixed x with respect to the norm $\|\cdot\|_{\lambda,*}$ with constant $\sigma_{\Phi y}$ which doesn't depend on x. Also we assume that:

$$\|\nabla_x\Phi(x, y_1) - \nabla_x\Phi(x, y_2)\|_\mu \leq L_{xy}\|y_1 - y_2\|_{\lambda,*}, \tag{70}$$

$$\|\nabla_y\Phi(x_1, y) - \nabla_y\Phi(x_2, y)\|_\lambda \leq L_{yx}\|x_1 - x_2\|_{\mu,*}, \tag{71}$$

$$\|\nabla_x\Phi(x_1, y) - \nabla_x\Phi(x_2, y)\|_\mu \leq L_{xx}\|x_1 - x_2\|_{\mu,*}, \tag{72}$$

$$\|\nabla_y\Phi(x, y_1) - \nabla_y\Phi(x, y_2)\|_\lambda \leq L_{yy}\|y_1 - y_2\|_{\lambda,*}. \tag{73}$$

Similarly to Lemma 3.1, we get that the conjugate problem for (55) is

$$\begin{aligned} \min_\lambda \max_\mu \{ \quad & \min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} [F(u, v) - \langle\mu, \mathcal{B}u\rangle + \langle\lambda, \mathcal{C}v\rangle] \\ & + \min_x \max_y [\Phi(x, y) + \langle\mu, x\rangle - \langle\lambda, y\rangle] - \langle\mu, \tilde{x}_0\rangle + \langle\lambda, \tilde{y}_0\rangle \ \}. \end{aligned} \tag{74}$$

We assume that the norms $\|\cdot\|_\lambda$ and $\|\cdot\|_\mu$ are Euclidian. Let us introduce the prox-function $d_\lambda(\lambda) := \frac{\sigma_\lambda}{2}\|\lambda\|_\lambda^2$. The function $d_\lambda(\lambda)$ is strongly convex in this norm with the convexity parameter σ_λ. For the variable μ we introduce the prox-function $d_\mu(\mu) := \frac{\sigma_\mu}{2}\|\mu\|_\mu^2$, which is strongly convex with the convexity parameter σ_μ with respect to the norm $\|\cdot\|_\mu$.

For any $\lambda_1, \lambda_2 \in \mathbb{R}^m$ we can define the Bregman distance:

$$\omega_\lambda(\lambda_1, \lambda_2) := d_\lambda(\lambda_2) - d_\lambda(\lambda_1) - \langle\nabla d_\lambda(\lambda_1), \lambda_2 - \lambda_1\rangle.$$

Using the explicit expression for $d_\lambda(\lambda)$, we get $\omega_\lambda(\lambda_1, \lambda_2) = \frac{\sigma_\lambda}{2} \|\lambda_1 - \lambda_2\|^2$. Let us choose $\bar{\lambda} = 0$ as the center of the space $\mathbb{R}^m$. Then we have $\omega_\lambda(\bar{\lambda}, \lambda) = d_\lambda(\lambda)$. For μ we introduce the similar settings.

Finding the saddle point (λ^*, μ^*) for the conjugate problem (74) is equivalent to solving the variational inequality

$$\langle g(\lambda, \mu), (\lambda - \lambda^*, \mu - \mu^*) \rangle \geq 0, \quad \forall \lambda, \mu, \tag{75}$$

$$\text{where} \quad g(\lambda, \mu) := (\nabla_\lambda \psi(\lambda, \mu), -\nabla_\mu \psi(\lambda, \mu)). \tag{76}$$

Let us choose some $\kappa \in]0, 1[$. Consider a space of $z := (\lambda, \mu)$ with the norm

$$\|z\|_z := \sqrt{\kappa \sigma_\lambda \|\lambda\|_\lambda^2 + (1 - \kappa)\sigma_\mu \|\mu\|_\mu^2},$$

an oracle $g(z) := (\nabla_\lambda \psi(\lambda, \mu), -\nabla_\mu \psi(\lambda, \mu))$, a new prox-function

$$d(z) := \kappa d_\lambda(\lambda) + (1 - \kappa) d_\mu(\mu)$$

which is strongly convex with constant $\sigma_0 = 1$. We define $W := \mathbb{R}^m \times \mathbb{R}^n$, the Bregman distance

$$\omega(z_1, z_2) := \kappa \omega_\lambda(\lambda_1, \lambda_2) + (1 - \kappa)\omega_\lambda(\mu_2, \mu_2)$$

which has an explicit form of $\omega(z_1, z_2) = d(z_1 - z_2)$, and center $\bar{z} = (0, 0)$. Then, $\omega(\bar{z}, z) = d(z)$. Note that the norm in the dual space is defined as

$$\|g\|_{z,*} := \sqrt{\frac{1}{\kappa \sigma_\lambda} \|g_\lambda\|_{\lambda,*}^2 + \frac{1}{(1 - \kappa)\sigma_\mu} \|g_\mu\|_{\mu,*}^2}.$$

In accordance to [32] for solving (75), we can use the following method:

1. Initialization: Fix $\beta = L$ (Lipshitz constant of g) . Set $s_{-1} = 0$.
2. Iteration ($k \geq 0$):

$$\begin{aligned} &\text{Compute } x_k = T_\beta(\bar{z}, s_{k-1}), \\ &\text{Compute } z_k = T_\beta(x_k, -g(x_k)), \\ &\text{Set } s_k = s_{k-1} - g(z_k). \end{aligned} \tag{M2}$$

Here $T_\beta(z, s) := \arg\max_{x \in W} \{\langle s, x - z \rangle - \beta \omega(z, x)\}$.

Similarly to [31], we can prove that the method (M2) generates a bounded sequence $\{z_i\}_{i \geq 0}$. Hence the sequences $\{\lambda_i\}_{i \geq 0}, \{\mu_i\}_{i \geq 0}$ are also bounded. Also, since the saddle point in the problem (55) exists, there exists a saddle point (λ^*, μ^*) for the conjugate problem (74). These arguments allow us to choose D_λ, D_μ such that $d_\lambda(\lambda_i) \le D_\lambda$, $d_\mu(\mu_i) \le D_\mu$ for all $i \ge 0$, which also ensure that (λ^*, μ^*) is an interior solution:

$$\mathfrak{B}^\lambda_{r/\sqrt{\kappa\sigma_\lambda}}(\lambda^*) \subseteq W_\lambda := \{\lambda : d_\lambda(\lambda) \le D_\lambda\},$$

$$\mathfrak{B}^\mu_{r/\sqrt{(1-\kappa)\sigma_\mu}}(\mu^*) \subseteq W_\mu := \{\mu : d_\mu(\mu) \le D_\mu\}$$

for some $r > 0$. Then we have $z^* := (\lambda^*, \mu^*) \in \mathcal{F}_D := \{z \in W : d(z) \le D\}$ with $D := \kappa D_\lambda + (1-\kappa)D_\mu$ and $\mathfrak{B}^z_r(z^*) \subseteq \mathcal{F}_D$.

Theorem 3.2. *Let the Assumptions* **A3** *and* **A4** *be true,* $\kappa = \frac{\sigma_\mu}{\sigma_\mu+\sigma_\lambda}$, *and*

$$L = \tfrac{\sigma_\lambda+\sigma_\mu}{\sigma_\mu\sigma_\lambda}\sqrt{2\left(\frac{\|\mathcal{C}\|^2_{\lambda,L_q^2}}{\sigma_{F_v}} + \frac{1}{\sigma_{\Phi_y}} + \frac{\|\mathcal{B}\|_{\mu,L_p^2}\|\mathcal{C}\|_{\lambda,L_q^2}L_{vu}}{\sigma_{F_u}\sigma_{F_v}} + \frac{L_{yx}}{\sigma_{\Phi_x}\sigma_{\Phi_y}}\right)}$$
$$\sqrt{\left(\frac{\|\mathcal{B}\|_{\mu,L_p^2}\|\mathcal{C}\|_{\lambda,L_q^2}L_{uv}}{\sigma_{F_u}\sigma_{F_v}} + \frac{L_{xy}}{\sigma_{\Phi_x}\sigma_{\Phi_y}} + \frac{\|\mathcal{B}\|^2_{\mu,L_p^2}}{\sigma_{F_u}} + \frac{1}{\sigma_{\Phi_x}}\right)}. \tag{77}$$

Let the points $z_i = (\lambda_i, \mu_i), i \ge 0$ *be generated by the method* (M2). *Let the points in* (61) *be defined by points* (u_i, v_i), (x_i, y_i) *which are the saddle points at the points* (λ_i, μ_i) *in* (74). *Then for functions* $\xi(u,x), \eta(v,y)$ *defined in* (64) *and* (65) *we have:*

$$\xi(\hat{u}_{k+1}, \hat{x}_{k+1}) - \eta(\hat{v}_{k+1}, \hat{y}_{k+1}) \le \frac{LD}{k+1}. \tag{78}$$

Also the following is true:

$$\|\mathcal{B}\hat{u}_{k+1} + \tilde{x}_0 - \hat{x}_{k+1}\|_{\mu,*} \le \frac{LD\sqrt{\sigma_\mu}}{r(k+1)}, \quad \|\mathcal{C}\hat{v}_{k+1} + \tilde{y}_0 - \hat{y}_{k+1}\|_{\lambda,*} \le \frac{LD\sqrt{\sigma_\lambda}}{r(k+1)}.$$

3.2 Accelerated primal-dual gradient method for strongly convex problems with linear constraints

The results of this subsection are published in [33, 34].

The main motivation for the algorithms in this subsection is approximating the optimal transport (OT) distance, which amounts to solving the *OT problem* [35]:

$$\min_{X \in \mathcal{U}(r,c)} \langle C, X\rangle,$$
$$\mathcal{U}(r,c) := \{X \in \mathbb{R}^{n\times n}_+ : X\mathbf{1} = r,\ X^T\mathbf{1} = c\}, \tag{79}$$

where X is *transportation plan*, $C \in \mathbb{R}^{n\times n}_+$ is a given ground cost matrix, $r, c \in \mathbb{R}^n$ are given vectors from the probability simplex Δ^n, $\mathbf{1}$ is the vector of all ones. The *regularized OT problem* is

$$\min_{X \in \mathcal{U}(r,c)} \langle C, X\rangle + \gamma\mathcal{R}(X), \tag{80}$$

where $\gamma > 0$ is the *regularization parameter* and $\mathcal{R}(X)$ is a strongly convex *regularizer*, e.g. negative entropy or squared Euclidean norm. Our goal is to find $\widehat{X} \in \mathcal{U}(r,c)$ such that

$$\langle C, \widehat{X} \rangle \leq \min_{X \in \mathcal{U}(r,c)} \langle C, X \rangle + \varepsilon. \tag{81}$$

In this case, $\langle C, \widehat{X} \rangle$ is an ε-approximation for the OT distance and $\widehat{X}$ is an approximation for the transportation plan.

Let us introduce some notation. For a general finite-dimensional real vector space E, we denote by E^* its dual, given by linear pairing $\langle g, x \rangle$, $x \in E$, $g \in E^*$; by $\| \cdot \|_E$ the norm in E and by $\| \cdot \|_{E,*}$ the norm in E^*, which is dual to $\| \cdot \|_E$. For a linear operator $A : E \to H$, we define its norm as $\|A\|_{E \to H} = \max_{x \in E, u \in H^*} \{ \langle u, Ax \rangle : \|x\|_E = 1, \|u\|_{H,*} = 1 \}$. We say that a function $f : E \to \mathbb{R}$ is γ-strongly convex on a set $Q \subseteq E$ w.r.t. a norm in E iff, for any $x, y \in Q$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\gamma}{2} \|x - y\|_E^2$, where $\nabla f(x)$ is any subgradient of $f(x)$ at x.

For a matrix A and a vector a, we denote e^A, e^a, $\ln A$, $\ln a$ their entrywise exponents and natural logarithms respectively. For a vector $a \in \mathbb{R}^n$, we denote by $\|a\|_1$ the sum of absolute values of its elements, and by $\|a\|_2$ its Euclidean norm, and by $\|a\|_\infty$ the maximum absolute value of its elements. Given a matrix $A \in \mathbb{R}^{n \times n}$, we denote by $\mathrm{vec}(A)$ the vector in $\mathbb{R}^{n^2}$, which is obtained from A by writing its columns one below another. For a matrix $A \in \mathbb{R}^{n \times n}$, we denote $\|A\|_1 = \|\mathrm{vec}(A)\|_1$ and $\|A\|_\infty = \|\mathrm{vec}(A)\|_\infty$. Further, we define the entropy of a matrix $X \in \mathbb{R}_+^{n \times n}$ by

$$H(X) := -\sum_{i,j=1}^{n} X^{ij} \ln X^{ij}. \tag{82}$$

For two matrices A, B, we denote their Frobenius inner product by $\langle A, B \rangle$. We denote by $\Delta^n := \{a \in \mathbb{R}_+^n : a^T \mathbf{1} = 1\}$ the probability simplex in $\mathbb{R}^n$.

We start by the following template minimization problem

$$\min_{x \in Q \subseteq E} \{ f(x) : Ax = b \}, \tag{83}$$

where E is a finite-dimensional real vector space, Q is a simple closed convex set, A is a given linear operator from E to some finite-dimensional real vector space H, $b \in H$ is given, $f(x)$ is a γ-strongly convex function on Q with respect to some chosen norm $\| \cdot \|_E$ on E.
The Lagrange dual problem for (83), written as a minimization problem, is

$$\min_{\lambda \in H^*} \left\{ \varphi(\lambda) := \langle \lambda, b \rangle + \max_{x \in Q} \left(-f(x) - \langle A^T \lambda, x \rangle \right) \right\}. \tag{84}$$

Note that $\nabla \varphi(\lambda) = b - Ax(\lambda)$ is Lipschitz-continuous [36]

$$\|\nabla \varphi(\lambda_1) - \nabla \varphi(\lambda_1)\|_H \leq L \|\lambda_1 - \lambda_2\|_{H,*},$$

Algorithm 12 Adaptive Primal-Dual Accelerated Gradient Descent (APDAGD)

Require: Accuracy $\varepsilon_f, \varepsilon_{eq} > 0$, initial estimate L_0 s.t. $0 < L_0 < 2L$.

1: Set $i_0 = k = 0$, $M_{-1} = L_0$, $\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
2: **repeat** {Main iterate}
3: **repeat** {Line search}
4: Set $M_k = 2^{i_k - 1} M_k$, find α_{k+1} s.t. $\beta_{k+1} := \beta_k + \alpha_{k+1} = M_k \alpha_{k+1}^2$. Set $\tau_k = \alpha_{k+1}/\beta_{k+1}$.
5: $\lambda_{k+1} = \tau_k \zeta_k + (1 - \tau_k)\eta_k$.
6: $\zeta_{k+1} = \zeta_k - \alpha_{k+1} \nabla \varphi(\lambda_{k+1})$.
7: $\eta_{k+1} = \tau_k \zeta_{k+1} + (1 - \tau_k)\eta_k$.
8: **until**

$$\varphi(\eta_{k+1}) \leq \varphi(\lambda_{k+1}) + \langle \nabla \varphi(\lambda_{k+1}), \eta_{k+1} - \lambda_{k+1} \rangle + \frac{M_k}{2} \|\eta_{k+1} - \lambda_{k+1}\|_2^2.$$

9: $\hat{x}_{k+1} = \tau_k x(\lambda_{k+1}) + (1 - \tau_k)\hat{x}_k$.
10: Set $i_{k+1} = 0$, $k = k + 1$.
11: **until** $f(\hat{x}_{k+1}) + \varphi(\eta_{k+1}) \leq \varepsilon_f$, $\|A\hat{x}_{k+1} - b\|_2 \leq \varepsilon_{eq}$.

Ensure: $\hat{x}_{k+1}$, η_{k+1}.

where $x(\lambda) := \arg\min_{x \in Q} \left(-f(x) - \langle A^T \lambda, x \rangle \right)$ and $L \leq \frac{\|A\|_{E \to H}^2}{\gamma}$. This estimate can be pessimistic and our algorithm does not use it and adapts automatically to the local value of the Lipschitz constant.

We assume that the dual problem (84) has a solution and there exists some $R > 0$ such that $\|\lambda^*\|_2 \leq R < +\infty$, where λ^* is the solution to (84) with minimum value of $\|\lambda^*\|_2$.

Theorem 3.3. *Assume that the objective in the primal problem* (83) *is γ-strongly convex and that the dual solution λ^* satisfies $\|\lambda^*\|_2 \leq R$. Then, for $k \geq 1$, the points $\hat{x}_k$, η_k in Algorithm 12 satisfy*

$$f(\hat{x}_k) - f^* \leq f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{16\|A\|_{E \to H}^2 R^2}{\gamma k^2}, \tag{85}$$

$$\|A\hat{x}_k - b\|_2 \leq \frac{16\|A\|_{E \to H}^2 R}{\gamma k^2}, \tag{86}$$

$$\|\hat{x}_k - x^*\|_E \leq \frac{8}{k} \frac{\|A\|_{E \to H} R}{\gamma}, \tag{87}$$

where x^* and f^* are respectively an optimal solution and the optimal value in (83)*. Moreover, the stopping criterion in step 11 is correctly defined.*

Now we apply the general method to derive a complexity estimate for finding $\widehat{X} \in \mathcal{U}(r, c)$ satisfying (81). We use entropic regularization of problem

Algorithm 13 Approximate OT by APDAGD

Require: Accuracy ε.

Set $\gamma = \frac{\varepsilon}{3\ln n}$.
for $k = 1, 2, ...$ **do**
  Make step of APDAGD and calculate $\widehat{X}_k$ and η_k.
  Find $\widehat{X}$ as the projection of $\widehat{X}_k$ on $\mathcal{U}(r,c)$ by Algorithm 2 in [37].
  if $\langle C, \widehat{X} - \widehat{X}_k\rangle \leq \frac{\varepsilon}{6}$ and $f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{\varepsilon}{6}$ **then**
    Return $\widehat{X}$.
  else
    $k = k + 1$ and continue.
  end if
end for

(79) and consider the regularized problem (80) with the regularizer $\mathcal{R}(X) = -H(X)$, where $H(X)$ is given in (82). We define $E = \mathbb{R}^{n^2}$, $\|\cdot\|_E = \|\cdot\|_1$, and variable $x = \mathrm{vec}(X) \in \mathbb{R}^{n^2}$ to be the vector obtained from a matrix X by writing each column of X below the previous column. Also we set $f(x) = \langle C, X\rangle - \gamma H(X)$, $Q = \Delta^{n^2}$, $b^T = (r^T, c^T)$ and $A : \mathbb{R}^{n^2} \to \mathbb{R}^{2n}$ defined by the identity $(A\,\mathrm{vec}(X))^T = ((X\mathbf{1})^T, (X^T\mathbf{1})^T)$. With this setting, we solve problem (83) by our APDAGD. Let $\widehat{X}_k$ be defined by identity $\mathrm{vec}(\widehat{X}_k) = \hat{x}_k$, where $\hat{x}_k$ is generated by APDAGD. We also define $\widehat{X} \in \mathcal{U}(r,c)$ to be the projection of $\widehat{X}_k$ onto $\mathcal{U}(r,c)$ constructed by Algorithm 2 in [37]. The pseudocode of our procedure for approximating the OT distance is listed as Algorithm 13.

Theorem 3.4. *Algorithm 13 outputs $\widehat{X} \in \mathcal{U}(r,c)$ satisfying* (81) *in*

$$O\left(\min\left\{\frac{n^{9/4}\sqrt{R\|C\|_\infty \ln n}}{\varepsilon}, \frac{n^2 R\|C\|_\infty \ln n}{\varepsilon^2}\right\}\right) \tag{88}$$

arithmetic operations.

3.3 Distributed primal-dual accelerated stochastic gradient method

The results of this subsection are published in [38].

We start with some notation. We define $\mathcal{M}_+^1(\mathcal{X})$ – the set of positive Radon probability measures on a metric space $\mathcal{X}$, and $S_1(n) = \{a \in \mathbb{R}_+^n \mid \sum_{l=1}^n a_l = 1\}$ the probability simplex. We use $\mathcal{C}(\mathcal{X})$ as the space of continuous functions on $\mathcal{X}$. We denote by $\delta(x)$ the Dirac measure at point x. We refer to $\lambda_{\max}(W)$ as the maximum eigenvalue of matrix W. We also use bold symbols for stacked vectors $\mathtt{p} = [p_1^T, \cdots, p_m^T]^T \in \mathbb{R}^{mn}$, where $p_1, ..., p_m \in \mathbb{R}^n$. In this case $[\mathtt{p}]_i = p_i$ – the i-th block of $\mathtt{p}$. For a vector $\lambda \in \mathbb{R}^n$, we denote by $[\lambda]_l$ its l-th component. We refer to the Euclidean norm of a vector $\|p\|_2 := \sqrt{\sum_{l=1}^{n}([p]_l)^2}$ as 2-norm.

Following the line of work started by [39], we consider entropic regularization for the optimal transport problem. Assume that we are given a positive Radon probability measure μ with density $q(y)$ on a metric space $\mathcal{Y}$, and a discrete probability measure $\nu = \sum_{i=1}^{n} p_i \delta(z_i)$ with weights p and finite support given by points $z_1, \ldots, z_n \in \mathcal{Z}$ from a metric space $\mathcal{Z}$. The regularized Wasserstein distance in semi-discrete setting between μ and ν is defined as

$$\mathcal{W}_\gamma(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \left\{ \sum_{i=1}^{n} \int_{\mathcal{Y}} c_i(y)\pi_i(y)dy + \gamma KL(\pi|\xi) \right\},$$

where $c_i(y) = c(z_i, y)$ is a cost function for transportation of a unit of mass from point z_i to point y, ξ is the uniform distribution on $\mathcal{Y} \times \mathcal{Z}$, $KL(\pi|\xi) = \sum_{i=1}^{n} \int_{\mathcal{Y}} \pi_i(y) \log\left(\frac{\pi_i(y)}{\xi}\right) dy$, and the set of admissible coupling measures π is defined as follows

$$\Pi(\mu, \nu) = \left\{ \pi \in \mathcal{M}_+^1(\mathcal{Y}) \times S_1(n) : \sum_{i=1}^{n} \pi_i(y) = q(y), y \in \mathcal{Y}, \int_{\mathcal{Y}} \pi_i(y)dy = p_i, \forall\, i = 1, \ldots, n \right\}.$$

For a set of positive Radon probability measures $(\mu_1, \ldots, \mu_m)$ the regularized Wasserstein barycenter in the semi-discrete setting is defined as the solution p to the problem

$$\min_{p \in S_1(n)} \sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}(p) = \min_{\substack{p_1 = \cdots = p_m \\ p_1, \ldots, p_m \in S_1(n)}} \sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}(p_i), \tag{89}$$

where we fixed the support $z_1, \ldots, z_n \in \mathcal{Z}$ of the barycenter ν and characterize it by the vector $p \in S_n(1)$, i.e., $\nu = \sum_{i=1}^{n} p_i \delta(z_i)$ and $\mathcal{W}_{\gamma,\mu}(p) := \mathcal{W}_\gamma(\mu, \nu)$.

We now describe the distributed optimization setting for solving the second problem in (89). We assume that each measure μ_i is held by an agent i on a network and this agent can sample from this measure. We model such a network as a fixed *connected undirected graph* $\mathcal{G} = (V, E)$, where V is the set of m nodes and E is the set of edges. We assume that the graph $\mathcal{G}$ does not have self-loops. The network structure imposes information constraints, specifically, each node i has access to μ_i only and a node can exchange information only with its immediate neighbors, i.e., a node i can communicate with node j if and only if $(i, j) \in E$.

We represent the communication constraints imposed by the network by introducing a single equality constraint instead of the constraints $p_1 = \cdots = p_m$ in (89). To do so, we define the Laplacian matrix $\bar{W} \in \mathbb{R}^{m \times m}$ of the graph $\mathcal{G}$ such that a) $[\bar{W}]_{ij} = -1$ if $(i, j) \in E$, b) $[\bar{W}]_{ij} = \deg(i)$ if $i = j$, c) $[\bar{W}]_{ij} = 0$ otherwise. Here $\deg(i)$ is the degree of the node i, i.e., the number of neighbors of the node. Finally, define the communication matrix (also referred to as an interaction matrix) by $W := \bar{W} \otimes I_n$.

In this setting, $\sqrt{W}\mathtt{p} = 0$ if and only if $p_1 = \cdots = p_m$. Using this fact, we equivalently rewrite problem (89) as the maximization problem with linear equality constraint

$$\max_{\substack{p_1,\dots,p_m \in S_1(n)\\ \sqrt{W}\mathtt{p}=0}} \quad -\sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}(p_i). \tag{90}$$

Given that problem (90) is an optimization problem with linear constraints, we introduce a stacked vector of dual variables $\boldsymbol{\lambda} = [\lambda_1^T, \cdots, \lambda_m^T]^T \in \mathbb{R}^{mn}$ for the constraints $\sqrt{W}\mathtt{p} = 0$ in (90). Then, the Lagrangian dual problem for (90) is

$$\begin{aligned} &\min_{\boldsymbol{\lambda}\in\mathbb{R}^{mn}} \max_{p_1,\dots,p_m\in S_1(n)} \left\{ \sum_{i=1}^{m} \langle \lambda_i, [\sqrt{W}\mathtt{p}]_i\rangle - \mathcal{W}_{\gamma,\mu_i}(p_i) \right\} \\ &= \min_{\boldsymbol{\lambda}\in\mathbb{R}^{mn}} \sum_{i=1}^{m} \mathcal{W}^*_{\gamma,\mu_i}([\sqrt{W}\boldsymbol{\lambda}]_i), \end{aligned} \tag{91}$$

where $[\sqrt{W}\mathtt{p}]_i$ and $[\sqrt{W}\boldsymbol{\lambda}]_i$ denote the i-th n-dimensional block of vectors $\sqrt{W}\mathtt{p}$ and $\sqrt{W}\boldsymbol{\lambda}$ respectively, and $\mathcal{W}^*_{\gamma,\mu_i}(\cdot)$ is the Fenchel-Legendre transform of $\mathcal{W}_{\gamma,\mu_i}(p_i)$.

Next, we consider a general smooth stochastic convex optimization problem which is dual to some optimization problem with linear equality constraints. For any finite-dimensional real vector space E, we denote by E^* its dual. Let $\|\cdot\|_E$ denote some norm on E and $\|\cdot\|_{E,*}$ denote the norm on E^* which is dual to $\|\cdot\|_E$ $\|\lambda\|_{E,*} = \max_{\|x\|_E\le 1}\langle \lambda, x\rangle$. For a linear operator $A : E_1 \to E_2$, we define the adjoint operator $A^T : E_2^* \to E_1^*$ in the following way $\langle u, Ax\rangle = \langle A^T u, x\rangle, \quad \forall u \in E_2^*, \quad x \in E_1$. We say that a function $f : E \to \mathbb{R}$ has a L-Lipschitz-continuous gradient w.r.t. norm $\|\cdot\|_{E,*}$ if it is differentiable and its gradient satisfies Lipschitz condition $\|\nabla f(x) - \nabla f(y)\|_{E,*} \le L\|x - y\|_E, \quad \forall x, y \in E$.

Our next goal is to provide an algorithm for a primal-dual pair of problems

$$(P) \min_{x\in Q\subseteq E} \{f(x) : Ax = b\}, \ (D) \quad \min_{\lambda\in\Lambda} \left\{ \langle \lambda, b\rangle + \max_{x\in Q} \left(-f(x) - \langle A^T\lambda, x\rangle\right) \right\}.$$

where Q is a simple closed convex set, $A : E \to H$ is given linear operator, $b \in H$ is given, $\Lambda = H^*$. We define

$$\varphi(\lambda) := \langle \lambda, b\rangle + \max_{x\in Q} \left(-f(x) - \langle A^T\lambda, x\rangle\right) = \langle \lambda, b\rangle + f^*(-A^T\lambda) \tag{92}$$

and assume it to be smooth with L-Lipschitz-continuous gradient. Here f^* is the Fenchel-Legendre dual for f. We also assume that $f^*(-A^T\lambda) =$

$\mathbb{E}_\xi F^*(-A^T\lambda,\xi)$, where ξ is random vector. Also, we define $F(x,\xi)$ to be the Fenchel-Legendre conjugate function to F^*, i.e. it satisfies $F^*(-A^T\lambda,\xi) = \max_{x\in Q}\{\langle -A^T\lambda, x\rangle - F(x,\xi)\}$ and $x(\lambda,\xi)$ to be the solution of this maximization problem. Under these assumptions, the dual problem (D) can be accessed by a stochastic oracle

$$(\Phi(x,\xi), \nabla\Phi(\lambda,\xi)) = (F^*(-A^T\lambda,\xi), \nabla F^*(-A^T\lambda,\xi))$$

satisfying $\mathbb{E}_\xi\Phi(\lambda,\xi) = \varphi(\lambda)$, $\mathbb{E}_\xi\nabla\Phi(\lambda,\xi) = \nabla\varphi(\lambda)$, which we use in our algorithm. Finally, we assume that dual problem (D) has a solution λ^* and there exists some $R > 0$ such that $\|\lambda^*\|_2 \le R < +\infty$.

We additionally assume that the variance of the stochastic approximation $\nabla\Phi(\lambda,\xi)$ for the gradient of φ can be controlled and made small as we desire, e.g. by mini-batching. Also, since $\nabla\Phi(\lambda,\xi) = b - A\nabla F^*(-A^T\lambda,\xi) = b - Ax(\lambda,\xi)$, on each iteration, to find $\nabla\Phi(\lambda,\xi)$ we find the vector $x(\lambda,\xi)$ and use it for the primal iterates.

Algorithm 14 Accelerated Primal-Dual Stochastic Gradient Method (APDSGM)

Require: Number of iterations N.

1: $C_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
2: **for** $k = 0, \dots, N-1$ **do**
3: Find α_{k+1} as the largest root of the equation $C_{k+1} := C_k + \alpha_{k+1} = 2L\alpha_{k+1}^2$. $\tau_{k+1} = \alpha_{k+1}/C_{k+1}$.
4: $\lambda_{k+1} = \tau_{k+1}\zeta_k + (1-\tau_{k+1})\eta_k$
5: $\zeta_{k+1} = \zeta_k - \alpha_{k+1}\nabla\Phi(\lambda_{k+1},\xi_{k+1})$.
6: $\eta_{k+1} = \tau_{k+1}\zeta_{k+1} + (1-\tau_{k+1})\eta_k$.
7: $\hat{x}_{k+1} = \tau_{k+1}x(\lambda_{k+1},\xi_{k+1}) + (1-\tau_{k+1})\hat{x}_k$.
8: **end for**

Ensure: The points $\hat{x}_{k+1}$, η_{k+1}.

Theorem 3.5. *Let φ have L-Lipschitz-continuous gradient w.r.t. 2-norm and $\|\lambda^*\|_2 \le R$, where λ^* is a solution of dual problem (D). Given desired accuracy ε, assume that, at each iteration of Algorithm 14, the stochastic gradient $\nabla\Phi(\lambda_k,\xi_k)$ is chosen in such a way that $\mathbb{E}_\xi\|\nabla\Phi(\lambda_k,\xi_k) - \nabla\varphi(\lambda_k)\|_2^2 \le \frac{\varepsilon L\alpha_k}{C_k}$. Then, for any $\varepsilon > 0$ and $N \ge 0$, and expectation $\mathbb{E}$ w.r.t. all the randomness $\xi_1,\dots,\xi_N$, the outputs η_N and $\hat{x}_N$ generated by the Algorithm 14 satisfy*

$$f(\mathbb{E}\hat{x}_N) - f^* \le \frac{32LR^2}{N^2} + \frac{\varepsilon}{2} \quad \textit{and} \quad \|A\mathbb{E}\hat{x}_N - b\|_2 \le \frac{32LR}{N^2} + \frac{\varepsilon}{2R}, \tag{93}$$

Next, we apply the general algorithm to solve the primal-dual pair of problems (90)-(91) and approximate the regularized Wasserstein barycenter which is a solution to (90).

Lemma 3.2. *The gradient of the objective function $\mathcal{W}_\gamma^*(\boldsymbol{\lambda})$ in the dual problem (91) is $\lambda_{\max}(W)/\gamma$-Lipschitz-continuous w.r.t. 2-norm. If its stochastic approximation is defined as*

$$[\widetilde{\nabla}\mathcal{W}_\gamma^*(\boldsymbol{\lambda})]_i = \sum_{j=1}^m \sqrt{W}_{ij}\widetilde{\nabla}\mathcal{W}_{\gamma,\mu_j}^*(\bar{\lambda}_j),\ i=1,...,m,\ \textit{with}$$

$$\widetilde{\nabla}\mathcal{W}_{\gamma,\mu_j}^*(\bar{\lambda}_j) = \frac{1}{M}\sum_{r=1}^M p_j(\bar{\lambda}_j),\ \textit{and}\ [p_j(\bar{\lambda}_j)]_l = \frac{\exp(([\bar{\lambda}_j]_l - c_l(Y_r^j))/\gamma)}{\sum_{\ell=1}^n \exp(([\bar{\lambda}_j]_\ell - c_\ell(Y_r^j))/\gamma)}. \tag{94}$$

where M is the batch size, $\bar{\lambda}_j := [\sqrt{W}\boldsymbol{\lambda}]_j$, $j=1,...,m$, $Y_1^j,...,Y_r^j$ is a sample from the measure μ_j, $j=1,...,m$. Then $\mathbb{E}_{Y_r^j\sim\mu_j, j=1,...,m,r=1,...,M}\widetilde{\nabla}\mathcal{W}_\gamma^*(\boldsymbol{\lambda}) = \nabla\mathcal{W}_\gamma^*(\boldsymbol{\lambda})$ and

$$\mathbb{E}_{Y_r^j\sim\mu_j, j=1,...,m,r=1,...,M}\|\widetilde{\nabla}\mathcal{W}_\gamma^*(\boldsymbol{\lambda}) - \nabla\mathcal{W}_\gamma^*(\boldsymbol{\lambda})\|_2^2 \le \frac{\lambda_{\max}(W)}{M},\ \boldsymbol{\lambda}\in\mathbb{R}^{mn}. \tag{95}$$

Based on this lemma, we see that if, on each iteration of Algorithm 14, the mini-batch size M_k satisfies $M_k \ge \frac{\lambda_{max}(W)C_k}{L\alpha_k\varepsilon}$, the assumptions of Theorem 3.5 hold.

For the particular problem (91) the step 5 of Algorithm 14 can be written block-wise $[\boldsymbol{\zeta}_{k+1}]_i = [\boldsymbol{\zeta}_k]_i - \alpha_{k+1}\sum_{j=1}^m \sqrt{W}_{ij}\widetilde{\nabla}\mathcal{W}_{\gamma,\mu_j}^*([\sqrt{W}\boldsymbol{\lambda}_{k+1}]_j)$, $i=1,...,m$. We change the variables and denote $\bar{\boldsymbol{\lambda}} = \sqrt{W}\boldsymbol{\lambda}$, $\bar{\boldsymbol{\eta}} = \sqrt{W}\boldsymbol{\eta}$, $\bar{\boldsymbol{\zeta}} = \sqrt{W}\boldsymbol{\zeta}$. Then the step 5 of Algorithm 14 becomes $[\bar{\boldsymbol{\zeta}}_{k+1}]_i = [\bar{\boldsymbol{\zeta}}_k]_i - \alpha_{k+1}\sum_{j=1}^m W_{ij}\widetilde{\nabla}\mathcal{W}_{\gamma,\mu_j}^*([\bar{\boldsymbol{\lambda}}_{k+1}]_j)$, $i=1,...,m$.

Algorithm 15 Distributed computation of Wasserstein barycenter

Require: Each agent $i \in V$ is assigned its measure μ_i.

All agents set $[\bar{\boldsymbol{\eta}}_0]_i = [\bar{\boldsymbol{\zeta}}_0]_i = [\bar{\boldsymbol{\lambda}}_0]_i = \mathbf{0} \in \mathbb{R}^n$,
$C_0 = \alpha_0 = 0$ and N

For each agent $i \in V$:

for $k = 0, \dots, N-1$ **do**

  Find α_{k+1} as the largest root of the equation
  $C_{k+1} := C_k + \alpha_{k+1} = 2L\alpha_{k+1}^2$.
  $\tau_{k+1} = \alpha_{k+1}/C_{k+1}$.

  Set $M_{k+1} = \max\{1,\ \lambda_{\max}(W)C_{k+1}/(L\alpha_{k+1}\varepsilon)\}$

  $[\bar{\boldsymbol{\lambda}}_{k+1}]_i = \tau_{k+1}[\bar{\boldsymbol{\zeta}}_k]_i + (1-\tau_{k+1})[\bar{\boldsymbol{\eta}}_k]_i$

  Generate M_{k+1} samples $\{Y_r^i\}_{r=1}^{M_{k+1}}$ from the measure μ_i and set $\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_i}([\bar{\boldsymbol{\lambda}}_{k+1}]_i)$ as in (94).

  Share $\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_i}([\bar{\boldsymbol{\lambda}}_{k+1}]_i)$ with $\{j \mid (i,j) \in E\}$

  $[\bar{\boldsymbol{\zeta}}_{k+1}]_i = [\bar{\boldsymbol{\zeta}}_k]_i - \alpha_{k+1}\sum_{j=1}^m W_{ij}\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_j}([\bar{\boldsymbol{\lambda}}_{k+1}]_j)$

  $[\bar{\boldsymbol{\eta}}_{k+1}]_i = \tau_{k+1}[\bar{\boldsymbol{\zeta}}_{k+1}]_i + (1-\tau_{k+1})[\bar{\boldsymbol{\eta}}_{k+1}]_i$

  $[\hat{\mathtt{p}}_{k+1}]_i = \tau_{k+1}p_i([\bar{\boldsymbol{\lambda}}_{k+1}]_i) + (1-\tau_{k+1})[\hat{\mathtt{p}}_{k+1}]_i$, where $p_i(\cdot)$ is defined in (94).

end for

Ensure: $\hat{\mathtt{p}}_N$.

Theorem 3.6. *Under the above assumptions, Algorithm 15 after $N = \sqrt{16\lambda_{max}(W)R^2/(\varepsilon\gamma)}$ iterations returns an approximation $\hat{\mathtt{p}}_N$ for the barycenter, which satisfies*

$$\sum_{i=1}^m \mathcal{W}_{\gamma,\mu_i}(\mathbb{E}[\hat{\mathtt{p}}_N]_i) - \sum_{i=1}^m \mathcal{W}_{\gamma,\mu_i}([\mathtt{p}^*]_i) \le \varepsilon, \quad \|\sqrt{W}\mathbb{E}\hat{\mathtt{p}}_N\|_2 \le \varepsilon/R. \tag{96}$$

Moreover, the total complexity is $O\left(n\max\lambda_{\max}(W)R^2/\varepsilon^2, \sqrt{\lambda_{\max}(W)R^2/(\varepsilon\gamma)}\right)$ arithmetic operations.

3.4 Primal-dual accelerated gradient method with small-dimensional relaxation oracle

The results of this subsection are published in [40, 41].

We consider the following minimization problem

$$(P_1) \qquad \min_{x\in Q\subseteq E}\{f(x) : \mathbf{A}x = b\},$$

where E is a finite-dimensional real vector space, Q is a simple closed convex set, $\mathbf{A}$ is given linear operator from E to some finite-dimensional real vector space H, $b \in H$ is given. The Lagrange dual problem to Problem (P_1) is

$$(D_1) \qquad \max_{\lambda\in\Lambda}\left\{-\langle\lambda, b\rangle + \min_{x\in Q}\left(f(x) + \langle\mathbf{A}^T\lambda, x\rangle\right)\right\}.$$

Here we denote $\Lambda = H^*$. It is convenient to rewrite Problem (D_1) in the equivalent form of a minimization problem

$$(P_2) \quad \min_{\lambda \in \Lambda} \left\{ \langle \lambda, b \rangle + \max_{x \in Q} \left(-f(x) - \langle \mathbf{A}^T \lambda, x \rangle \right) \right\}.$$

We denote

$$\varphi(\lambda) = \langle \lambda, b \rangle + \max_{x \in Q} \left(-f(x) - \langle \mathbf{A}^T \lambda, x \rangle \right). \tag{97}$$

Since f is convex, $\varphi(\lambda)$ is a convex function and, by Danskin's theorem, its subgradient is equal to (see e.g. [36])

$$\nabla \varphi(\lambda) = b - \mathbf{A} x(\lambda) \tag{98}$$

where $x(\lambda)$ is some solution of the convex problem

$$\max_{x \in Q} \left(-f(x) - \langle \mathbf{A}^T \lambda, x \rangle \right). \tag{99}$$

In what follows, we make the following assumptions about the dual problem (D_1)

- Subgradient of the objective function $\varphi(\lambda)$ satisfies Hölder condition with constant M_ν, i.e., for all $\lambda, \mu \in \Lambda$ and some $\nu \in [0,1]$

$$\|\nabla \varphi(\lambda) - \nabla \varphi(\mu)\|_* \leqslant M_\nu \|\lambda - \mu\|^\nu. \tag{100}$$

- The dual problem (D_1) has a solution λ^* and there exist some $R > 0$ such that

$$\|\lambda^*\|_2 \leqslant R < +\infty. \tag{101}$$

We choose Euclidean proximal setup in the dual space, which means that we introduce Euclidean norm $\|\cdot\|_2$ in the space of vectors λ and choose the prox-function $d(\lambda) = \frac{1}{2}\|\lambda\|_2^2$. Then, we have for the Bregman distance $V[\zeta](\lambda) = \frac{1}{2}\|\lambda - \zeta\|_2^2$. Our primal-dual algorithm for Problem (P_1) is listed below as Algorithm 16.

Theorem 3.7. *Let the objective φ in the problem (P_2) have Hölder-continuous subgradient and the solution of this problem be bounded, i.e. $\|\lambda^*\|_2 \leqslant R$. Then, for the sequence $\hat{x}^{k+1}, \eta^{k+1}$, $k \geqslant 0$, generated by Algorithm 16,*

$$\|\mathbf{A}\hat{x}^k - b\|_2 \leqslant \frac{2R}{A_k} + \frac{\varepsilon}{2R}, \quad |\varphi(\eta^k) + f(\hat{x}^k)| \leqslant \frac{2R^2}{A_k} + \frac{\varepsilon}{2}, \tag{102}$$

where $A_k \geqslant \left[\frac{1+\nu}{1-\nu}\right]^{\frac{1-\nu}{1+\nu}} \frac{k^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}$.

Algorithm 16 PDUGDsDR

Require: starting point $\lambda_0 = 0$, accuracy $\tilde{\varepsilon}_f, \tilde{\varepsilon}_{eq} > 0$.

1: Set $k = 0$, $A_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.

2: **repeat**

3: $\beta_k = \mathrm{argmin}_{\beta\in[0,1]} \varphi\left(\zeta^k + \beta(\eta^k - \zeta^k)\right)$; $\lambda^k = \zeta^k + \beta_k(\eta^k - \zeta^k)$

4: $h_{k+1} = \mathrm{argmin}_{h\geqslant 0} \varphi\left(\lambda^k - h\nabla\varphi(\lambda^k)\right)$; $\eta^{k+1} = \lambda^k - h_{k+1}\nabla\varphi(\lambda^k)$ // Choose $\nabla\varphi(\lambda^k) : \langle\nabla\varphi(\lambda^k), \zeta^k - \lambda^k\rangle \geqslant 0$

5: Choose a_{k+1} from $\varphi(\eta^{k+1}) = \varphi(\lambda^k) - \frac{a_{k+1}^2}{2A_{k+1}}\|\nabla\varphi(\lambda^k)\|_2^2 + \frac{\varepsilon a_{k+1}}{2A_{k+1}}$ // $A_{k+1} = A_k + a_{k+1}$

6: $\zeta^{k+1} = \zeta^k - a_{k+1}\nabla\varphi(\lambda^k)$

7: Set
$$\hat{x}^{k+1} = \frac{1}{A_{k+1}}\sum_{i=0}^{k} a_{i+1}x(\lambda^i) = \frac{a_{k+1}x(\lambda^k) + A_k\hat{x}^k}{A_{k+1}}.$$

8: Set $k = k + 1$.

9: **until** $|f(\hat{x}^{k+1}) + \varphi(\eta^{k+1})| \leqslant \tilde{\varepsilon}_f$, $\|\mathbf{A}\hat{x}^{k+1} - b\|_2 \leqslant \tilde{\varepsilon}_{eq}$.

Ensure: The points $\hat{x}^{k+1}$, η^{k+1}.

Let us make a remark on complexity. As it can be seen from Theorem 3.7, whenever $A_k \geqslant 2R^2/\varepsilon$, the error in the objective value and equality constraints is smaller than ε. At the same time, using the lower bound for A_k, we obtain that the number of iterations to achieve this accuracy is $O\left(\left(\frac{M_\nu^{\frac{2}{1+\nu}}R^2}{\varepsilon^{\frac{2}{1+\nu}}}\right)^{\frac{1+\nu}{1+3\nu}}\right)$. Since the algorithm does not use the value of ν, we can take infimum in $\nu \in [0,1]$ of this complexity. This means that the method is uniformly optimal for the class of problems with Hölder-continuous gradient.

4 Conclusion

This thesis is based on published papers [21, 22, 24, 27, 30, 33, 34, 38, 40, 41].

In papers [21, 22, 24, 27] we developed optimization methods with (stochastic) inexact first-order oracle, inexact zero-order oracle, inexact directional derivative oracle. We also considered a particular application to learning a parametric model for web-page ranking.

Papers [30, 33, 34, 38, 40, 41] devoted to primal-dual methods for convex problems with linear constraints. In particular, we consider infinite-dimensional problems and propose dimension-independent convergence rates for this problem. We also consider (stochastic) convex problems with linear constraints and propose accelerated gradient methods with optimal convergence rates. We apply these methods for approximating optimal transport distance and barycenters.

Let us list the main results that are obtained in this thesis and submitted for defense.

1. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle.
2. Gradient method with inexact oracle for deterministic non-convex optimization and gradient-free method with inexact oracle for deterministic convex optimization.
3. A concept of inexact oracle for the methods which use directional derivatives, accelerated and non-accelerated inexact directional derivative method for strongly convex smooth stochastic optimization.
4. Primal-dual methods for solving infinite-dimensional games in convex-concave and strongly convex-concave setting.
5. Non-adaptive and adaptive accelerated primal-dual gradient method for strongly convex minimization problems with linear equality and inequality constraints.
6. New complexity estimates for the optimal transport distance problem.
7. Stochastic primal-dual accelerated gradient method for problems with linear constraints and its application to the problem of approximation of Wasserstein barycenter.
8. A universal primal-dual accelerated gradient method with line-search.

Acknowledgements

The dissertation was supported by Russian Science Foundation (project 18-71-10108), by RFBR project number 18-31-20005 mol-a-ved, by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) No.075-00337-20-03, project No. 0714-2020-0005, and by RFBR project number 18-29-03071-mk.

5 References

[1] N. Karmarkar. "A new polynomial-time algorithm for linear programming". In: *Combinatorica* 4.4 (1984), pp. 373–395. ISSN: 1439-6912. DOI: 10.1007/BF02579150.

[2] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming.* SIAM, 1994.

[3] Augustin Cauchy. "Méthode générale pour la résolution des systémes d'équations simultanées". In: *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 55 (1847), pp. 536–538.

[4] Boris Polyak. "Gradient methods for the minimisation of functionals". In: *USSR Computational Mathematics and Mathematical Physics* 3.4 (1963), pp. 864–878. ISSN: 0041-5553. DOI: http://dx.doi.org/10.1016/0041-5553(63)90382-3.

[5] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *Ann. Math. Statist.* 22.3 (Sept. 1951), pp. 400–407. DOI: 10.1214/aoms/1177729586.

[6] A.S. Nemirovsky and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* J. Wiley & Sons, New York, 1983.

[7] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate $O(1/k^2)$". In: *Soviet Mathematics Doklady* 27.2 (1983), pp. 372–376.

[8] Amir Beck and Marc Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202. DOI: 10.1137/080716542.

[9] Yurii Nesterov. "Gradient methods for minimizing composite functions". In: *Mathematical Programming* 140.1 (2013). First appeared in 2007 as CORE discussion paper 2007/76, pp. 125–161.

[10] Guanghui Lan. "An optimal method for stochastic composite optimization". In: *Mathematical Programming* 133.1 (2012). Firs appeared in June 2008, pp. 365–397. ISSN: 1436-4646.

[11] Rie Johnson and Tong Zhang. "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., 2013, pp. 315–323. URL: http://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predictive-variance-reduction.pdf.

[12] Qihang Lin, Zhaosong Lu, and Lin Xiao. "An Accelerated Proximal Coordinate Gradient Method". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. First appeared in arXiv:1407.1296. Curran Associates, Inc., 2014, pp. 3059–3067. URL: http://papers.nips.cc/paper/5356-an-accelerated-proximal-coordinate-gradient-method.pdf.

[13] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. "A Universal Catalyst for First-order Optimization". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 3384–3392.

[14] Guanghui Lan and Yi Zhou. "An optimal randomized incremental gradient method". In: *Mathematical Programming* (2017). ISSN: 1436-4646. DOI: 10.1007/s10107-017-1173-0.

[15] Shai Shalev-Shwartz and Tong Zhang. "Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research. First appeared in arXiv:1309.2375. Bejing, China: PMLR, 2014, pp. 64–72. URL: http://proceedings.mlr.press/v32/shalev-shwartz14.html.

[16] Yurii Nesterov. "Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems". In: *SIAM Journal on Optimization* 22.2 (2012). First appeared in 2010 as CORE discussion paper 2010/2, pp. 341–362. DOI: 10.1137/100802001.

[17] Yurii Nesterov and Vladimir Spokoiny. "Random Gradient-Free Minimization of Convex Functions". In: *Found. Comput. Math.* 17.2 (Apr. 2017). First appeared in 2011 as CORE discussion paper 2011/16, pp. 527–566. ISSN: 1615-3375. DOI: 10.1007/s10208-015-9296-2.

[18] Alexandre d'Aspremont. "Smooth Optimization with Approximate Gradient". In: *SIAM J. on Optimization* 19.3 (Oct. 2008), pp. 1171–1183. ISSN: 1052-6234. DOI: 10.1137/060676386.

[19] Olivier Devolder, François Glineur, and Yurii Nesterov. "First-order methods of smooth convex optimization with inexact oracle". In: *Mathematical Programming* 146.1 (2014), pp. 37–75. ISSN: 1436-4646. DOI: 10.1007/s10107-013-0677-5.

[20] Amir Beck and Marc Teboulle. "A fast dual proximal gradient algorithm for convex minimization and applications". In: *Operations Research Letters* 42.1 (2014), pp. 1–6.

[21] Pavel Dvurechensky and Alexander Gasnikov. "Stochastic Intermediate Gradient Method for Convex Problems with Stochastic Inexact Oracle". In: *Journal of Optimization Theory and Applications* 171.1 (2016), pp. 121–145.

[22] A. V. Gasnikov and P. E. Dvurechensky. "Stochastic intermediate gradient method for convex optimization problems". In: *Doklady Mathematics* 93.2 (2016), pp. 148–151.

[23] Olivier Devolder. "Stochastic first order methods in smooth convex optimization". In: *CORE Discussion Paper 2011/70* (2011).

[24] Lev Bogolubsky, Pavel Dvurechensky, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrei M Raigorodskii, Aleksey Tikhonov, and Maksim Zhukovskii. "Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. arXiv:1603.00717. Curran Associates, Inc., 2016, pp. 4914–4922.

[25] Yurii Nesterov and Arkadi Nemirovski. "Finding the stationary states of Markov chains by iterative methods". In: *Applied Mathematics and Computation* 255 (2015). Special issue devoted to the international conference "Numerical computations: Theory and Algorithms" June 17–23, 2013, Falerna, Italy, pp. 58–65. ISSN: 0096-3003. DOI: `https://doi.org/10.1016/j.amc.2014.04.053`.

[26] *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. PhD thesis. CORE UCL, 2013.

[27] Pavel Dvurechensky, Eduard Gorbunov, and Alexander Gasnikov. "An Accelerated Directional Derivative Method for Smooth Stochastic Convex Optimization". In: *European Journal of Operational Research* (2020). ISSN: 0377-2217. DOI: `https://doi.org/10.1016/j.ejor.2020.08.027`.

[28] *Ben-Tal A., Nemirovski A.* Lectures on Modern Convex Optimization. Philadelphia: SIAM, 2015. URL: `http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf`.

[29] Anatoli Juditsky and Yuri Nesterov. "Deterministic and Stochastic Primal-Dual Subgradient Algorithms for Uniformly Convex Minimization". In: *Stochastic Systems* 4.1 (2014), pp. 44–80. DOI: `10.1287/10-SSY010`.

[30] Pavel Dvurechensky, Yurii Nesterov, and Vladimir Spokoiny. "Primal-Dual Methods for Solving Infinite-Dimensional Games". In: *Journal of Optimization Theory and Applications* 166.1 (2015), pp. 23–51.

[31] Yurii Nesterov. "Primal-dual subgradient methods for convex problems". In: *Mathematical Programming* 120.1 (2009). First appeared in 2005 as CORE discussion paper 2005/67, pp. 221–259. ISSN: 1436-4646. DOI: `10.1007/s10107-007-0149-x`.

[32] Yurii Nesterov. "Dual extrapolation and its applications to solving variational inequalities and related problems". In: *Mathematical Programming* 109.2-3 (2007). First appeared in 2003 as CORE discussion paper 2003/68, pp. 319–344.

[33] Alexey Chernov, Pavel Dvurechensky, and Alexander Gasnikov. "Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints". In: *Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings*. Ed. by Yury Kochetov, Michael Khachay, Vladimir Beresnev, Evgeni Nurminski, and Panos Pardalos. Springer International Publishing, 2016, pp. 391–403.

[34] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. "Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. arXiv:1802.04367. 2018, pp. 1367–1376.

[35] Leonid Kantorovich. "On the translocation of masses". In: *Doklady Acad. Sci. USSR (N.S.)* 37 (1942), pp. 199–201.

[36] Yurii Nesterov. "Smooth minimization of non-smooth functions". In: *Mathematical Programming* 103.1 (2005), pp. 127–152.

[37] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. "Near-linear time approxFimation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. arXiv:1705.09634. Curran Associates, Inc., 2017, pp. 1961–1971.

[38] Pavel Dvurechensky, Darina Dvinskikh, Alexander Gasnikov, César A. Uribe, and Angelia Nedić. "Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. NeurIPS 2018. arXiv:1806.03915. Curran Associates, Inc., 2018, pp. 10783–10793.

[39] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., 2013, pp. 2292–2300.

[40] S. V. Guminov, Yu. E. Nesterov, P. E. Dvurechensky, and A. V. Gasnikov. "Accelerated Primal-Dual Gradient Descent with Linesearch for Convex, Nonconvex, and Nonsmooth Optimization Problems". In: *Doklady Mathematics* 99.2 (2019), pp. 125–128.

[41] Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. "Primal-dual accelerated gradient methods with small-dimensional relaxation oracle". In: *Optimization Methods and Software* (2020), pp. 1–28. DOI: 10.1080/10556788.2020.1731747.