EURO Journal on Computational Optimization 10 (2022) 100048

Contents lists available at ScienceDirect

EURO Journal on Computational Optimization

www.elsevier.com/locate/ejco

Accelerated variance-reduced methods for saddle-point problems

Ekaterina Borodich [a,*], Vladislav Tominin [a], Yaroslav Tominin [a], Dmitry Kovalev [d], Alexander Gasnikov [a,b,e], Pavel Dvurechensky [c]

[a] *Moscow Institute of Physics and Technology, Moscow, Russia*
[b] *HSE University, Moscow, Russia*
[c] *Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany*
[d] *King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*
[e] *Institute for Information Transmission Problems RAS, Moscow, Russia*

ARTICLE INFO

Keywords:
Saddle-point problem
Minimax optimization
Composite optimization
Stochastic variance-reduced algorithms
Accelerated algorithms

ABSTRACT

We consider composite minimax optimization problems where the goal is to find a saddle-point of a large sum of non-bilinear objective functions augmented by simple composite regularizers for the primal and dual variables. For such problems, under the average-smoothness assumption, we propose accelerated stochastic variance-reduced algorithms with optimal up to logarithmic factors complexity bounds. In particular, we consider strongly-convex-strongly-concave, convex-strongly-concave, and convex-concave objectives. To the best of our knowledge, these are the first nearly-optimal algorithms for this setting.

* Corresponding author.
E-mail addresses: borodich.ed@phystech.edu (E. Borodich), tominin.vd@phystech.edu (V. Tominin), tominin.yad@phystech.edu (Y. Tominin), dmitry.kovalev@kaust.edu.sa (D. Kovalev), gasnikov.av@mipt.ru (A. Gasnikov), pavel.dvurechensky@wias-berlin.de (P. Dvurechensky).

https://doi.org/10.1016/j.ejco.2022.100048

1. Introduction

Saddle-point optimization problems have many applications in different areas of modeling and optimization. The most classical example is, perhaps, two-player zero-sum games [25,26], including differential games [15]. More recent examples include imaging problems [4] and machine learning problems [36], where primal-dual saddle-point representations of large-scale optimization problems are constructed, and primal-dual first-order methods are used for their efficient solution. Many non-smooth optimization problems, such as ℓ_∞ or ℓ_1 regression admit a saddle-point representation, which allows one to propose methods [31,27] having faster convergence than the standard subgradient scheme.

Recently saddle-point problems started to attract more attention from the machine learning community motivated by applications to generative adversarial networks training [5,23], where the training process consists of a competition of a generator of non-real images and a discriminator which tries to distinguish between real and artificial images. Another application example is equilibrium problems in two-stage congested traffic flow models [8,11].

From the algorithmic viewpoint, the most studied setting deals with saddle-point problems having bilinear structure [31,27,3,37,42], where the cross term between the primal and dual variable is linear with respect to each variable. The extensions include bilinear problems with prox-friendly (i.e., admitting a proximal operator in closed form) composite terms [4,19]. A related line of research studies variational inequalities [27,19] since any convex-concave saddle-point problem can be reformulated as a variational inequality problem with a monotone operator. In this area, lower bounds for first-order methods are known [28] and many optimal methods exist [27,32,33,6,19,39,38]. Notably, these works do not rely on the bilinear structure and allow to solve convex-concave saddle-point problems with Lipschitz-continuous gradients, including differential games [7]. An alternative approach, which mostly inspired this paper, is based on a representation of a saddle-point problem $\min_x \max_y \tilde{F}(x, y)$ as either a primal minimization problem with an implicitly given objective $\hat{g}(x) = \max_y \tilde{F}(x, y)$ or a dual maximization problem with an implicitly given objective $\tilde{g}(y) = \min_x \tilde{F}(x, y)$. This approach was used in [31,30] for problems with bilinear structure and later extended in [13] for general saddle-point problems. Such connection to optimization turned out to be quite productive since it allows for the exploitation of accelerated optimization methods. In particular, recent advances in this direction are due to an observation [9,2,14,22] that primal and dual problems can have different condition numbers, which opens up a possibility to obtain theoretically faster algorithms, including accelerated near-optimal algorithms [22].

In this paper, we focus on strongly-convex-strongly-concave and convex-concave saddle-point problems with the finite-sum structure[1] and prox-friendly composite terms:

[1] Algorithms for general infinite-expectation saddle-point problems can be found in [43] and references therein.

$$\min_{x\in\mathbb{R}^{d_x}}\max_{y\in\mathbb{R}^{d_y}}\left\{f(x)+G(x,y)-h(y)\right\},\quad G(x,y):=\frac{1}{m}\sum_{i=1}^{m}G_i(x,y), \tag{1.1}$$

with the assumption that $G(x,y)$ is L-average-smooth (precise definition will be given below in (1.8)), $f(x), h(y)$ are prox-friendly and μ_x-, μ_y-(strongly)-convex respectively with $\mu_x, \mu_y \geq 0$. Our goal is to develop accelerated first-order stochastic variance-reduced algorithms that find an (ε,σ)-solution, where $\varepsilon > 0$ is the desired accuracy in terms of the duality gap and $\sigma \in (0,1)$ is a confidence level, i.e., the duality gap is guaranteed to be smaller than ε with probability at least $1-\sigma$.

Recently, the authors of [12] proved the lower complexity bounds for stochastic first-order algorithms for the problem

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\frac{1}{m}\sum_{i=1}^{m}G_i(x,y), \tag{1.2}$$

where $G(x,y)$ is assumed to be L-average-smooth and (μ_x,μ_y)-(strongly)-convex-(strongly)-concave, with $\mu_x \geq 0$ and $\mu_y \geq 0$, $\operatorname{diam}(\mathcal{X}) \leq D_x$, $\operatorname{diam}(\mathcal{Y}) \leq D_y$, where $\operatorname{diam}(\cdot)$ denotes the diameter of a set, i.e., maximum distance between its two points. Their lower bounds for finding an ε-solution in terms of the expectation of the duality gap are

$$\begin{cases}\Omega\left(\sqrt{m}\sqrt{\left(\sqrt{m}+\frac{L}{\mu_x}\right)\left(\sqrt{m}+\frac{L}{\mu_y}\right)}\ln\frac{1}{\varepsilon}\right), & \mu_x>0,\mu_y>0;\\ \Omega\left(m+D_x m^{\frac{3}{4}}\sqrt{\frac{L}{\varepsilon}}+\frac{\sqrt{m}LD_x}{\sqrt{\mu_y\varepsilon}}+m^{\frac{3}{4}}\sqrt{\frac{L}{\mu_y}}\ln\frac{1}{\varepsilon}\right), & \mu_x=0,\mu_y>0;\\ \Omega\left(m+\frac{\sqrt{m}LD_xD_y}{\varepsilon}+(D_x+D_y)m^{\frac{3}{4}}\sqrt{\frac{L}{\varepsilon}}\right), & \mu_x=0,\mu_y=0.\end{cases} \tag{1.3}$$

These results raised the question of whether these "accelerated" lower bounds can be achieved by some algorithms.

When an optimization problem has the finite-sum structure, also known as the empirical risk minimization problem, stochastic variance-reduced methods, see, e.g., [17,19], allow for faster convergence rates. For optimization problems, such methods are well developed and achieve lower complexity bounds [41]. On the contrary, the literature on variance-reduced methods for saddle-point problems is quite scarce. To our knowledge, first, such techniques were applied to saddle-point problems in [34], but the obtained bounds did not have separation between μ_x and μ_y, and there was no acceleration. Recently, these bounds were improved in [1], where an algorithm was proposed with non-accelerated complexity.

$$\widetilde{O}\left(m+\sqrt{m}\left(\frac{L}{\mu_x}+\frac{L}{\mu_y}\right)\right)$$

for (μ_x,μ_y)-strongly-convex-strongly-concave setting with $\mu_x,\mu_y > 0$. Moreover, for convex-concave setting their algorithm has complexity

$$O\left(m+\sqrt{m}\frac{L({D_x}^2+{D_y}^2)}{\varepsilon}\right).$$

Despite these bounds are optimal in the particular cases $\mu_x = \mu_y$ (strongly-convex-strongly-concave setting) or $D_x = D_y$ (convex-concave setting), the theoretical gap between the lower and upper bounds still remains when $\mu_x \neq \mu_y$ or $D_x \neq D_y$. This is especially important if, e.g., $\mu_x \ll \mu_y$ since then the lower bound becomes much smaller than the upper bound. Moreover, the case when $\mu_x \ll \mu_y$ highlights the benefits of separation between μ_x and μ_y. Indeed, under additional assumption that $L/\mu_x, L/\mu_y \gg \sqrt{m}$, the non-separated bound [34] is then $\widetilde{O}\left(\frac{L\sqrt{m}}{\min\{\mu_x,\mu_y\}}\right)$ which is much worse than the separated bound $\widetilde{O}\left(\frac{L\sqrt{m}}{\sqrt{\mu_x\mu_y}}\right)$. Thus, we focus on the case when the strong convexity parameters (or diameters of the sets) are different.

Our contribution. In this paper, we continue the line of research [2,10] by exploring additional properties of problem (1.1), namely, the finite-sum structure and the presence of prox-friendly composite terms. Since the problem class (1.1) contains the problem class (1.2), the lower bounds (1.3) are also valid for solving problem (1.1) by stochastic first-order methods. Our main contribution in this paper is, to a large extent, theoretical. We propose accelerated stochastic first-order variance-reduced algorithms that have nearly-optimal complexity, i.e., their complexity coincides with the bounds (1.3) up to logarithmic factors:

$$\begin{cases} \mathrm{O}\left(\left(m+m^{\frac{3}{4}}\sqrt{\frac{L}{\mu_x}}+m^{\frac{3}{4}}\sqrt{\frac{L}{\mu_y}}+\frac{L\sqrt{m}}{\sqrt{\mu_x\mu_y}}\right)\ln^3\frac{1}{\varepsilon}\right), & \mu_x>0,\mu_y>0; \\ \mathrm{O}\left(\left(m+m^{\frac{3}{4}}R_x\sqrt{\frac{L}{\varepsilon}}+m^{\frac{3}{4}}\sqrt{\frac{L}{\mu_y}}+\frac{R_xL\sqrt{m}}{\sqrt{\varepsilon\mu_y}}\right)\ln^3\frac{1}{\varepsilon}\right), & \mu_x=0,\mu_y>0; \\ \mathrm{O}\left(\left(m+(R_x+R_y)m^{\frac{3}{4}}\sqrt{\frac{L}{\varepsilon}}+\frac{R_xR_yL\sqrt{m}}{\varepsilon}\right)\ln^3\frac{1}{\varepsilon}\right), & \mu_x=0,\mu_y=0, \end{cases}$$

where in the absence of strong convexity and/or strong concavity, we assume that there exists a saddle point (x^*, y^*) for problem (1.1) satisfying $\|x^*\| \leq R_x$, $\|y^*\| \leq R_y$. Importantly, our algorithms guarantee the accuracy ε with high probability, rather than in expectation. To the best of our knowledge, these are the first nearly-optimal algorithms for this setting.[2] Our algorithms have multi-loop structure and provide a conceptual understanding that the lower complexity bounds (1.3) are achievable. Efficient implementation and/or loop-less algorithms achieving lower bounds are left for the future work.

Notation and definitions. We introduce some notation and necessary definitions used throughout the paper. We denote by $\|x\|$ and $\|y\|$ the standard Euclidean norms for $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$ respectively. This leads to the Euclidean norm on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ defined as $\|(x_1, y_1) - (x_2, y_2)\|^2 = \|x_1 - x_2\|^2 + \|y_1 - y_2\|^2$, $x_1, x_2, \in \mathbb{R}^{d_x}, y_1, y_2 \in \mathbb{R}^{d_y}$. $\mathcal{B}_2(\theta_0, R)$ denotes the Euclidean ball with center at θ_0 and radius R, i.e., $\left\{\theta \in \mathbb{R}^{d_\theta} : \|\theta - \theta_0\| \leq R\right\}$.

[2] The first version of our results appeared as a part of the preprint [40]. We believe that Algorithm 1 from [1] can be accelerated to achieve optimal complexity bounds.

We say that a function f is μ_f-strongly-convex if, for some $\mu_f > 0$ and for any its subgradient $\nabla f(x_1)$, it holds that

$$f(x_2) \geq f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle + \frac{\mu_f}{2}\|x_1 - x_2\|^2, \quad x_1, x_2 \in \mathrm{dom} f.$$

Note that when $\mu_f = 0$, we also say that f is convex. We say that a function f is L-smooth if its gradient is Lipschitz-continuous, i.e., $\|\nabla f(x_1) - \nabla f(x_2)\| \leq \|x_1 - x_2\|$, $x_1, x_2 \in \mathrm{dom} f$. We say that a function f is prox-friendly if it admits a tractable proximal operator [24]. This means that the evaluation of the point

$$prox_f^{\lambda}(\bar{x}) = \arg\min_{x \in \mathrm{dom} f} \left\{ \lambda f(x) + \frac{1}{2}\|x - \bar{x}\|^2 \right\} \tag{1.4}$$

for some fixed $\bar{x} \in \mathbb{R}^{d_x}$ and $\lambda > 0$ can be made either in closed form or numerically very efficiently up to machine precision.

For an optimization problem $\min_x f(x)$, we say that a random point $\hat{x}$ is an (ε, σ)-solution to this problem for some $\varepsilon > 0$ and $\sigma \in (0, 1)$, if

$$f(\hat{x}) - \min_x f(x) \leq \varepsilon \quad \text{with probability at least } 1 - \sigma. \tag{1.5}$$

We refer to ε as *accuracy* and to σ as *confidence level*.

We say that a function $G(x, y)$ is (strongly)-convex-(strongly)-concave if the function $G(\cdot, y)$ is (strongly)-convex for any fixed y and the function $G(x, \cdot)$ is (strongly)-concave for any fixed x. For a strongly-convex-strongly-concave saddle-point problem $\min_x \max_y G(x, y)$ a point $(\hat{x}, \hat{y})$ is called an (ε, σ)-solution for some $\varepsilon > 0$ and $\sigma \in (0, 1)$, if

$$\max_y G(\hat{x}, y) - \min_x G(x, \hat{y}) \leq \varepsilon \quad \text{with probability at least } 1 - \sigma. \tag{1.6}$$

Note that since the saddle-point problem is strongly-convex-strongly-concave, the quantity in the l.h.s. of (1.6) is correctly defined.

Notation $\widetilde{O}(\cdot)$ hides constant and polylogarithmic in ε^{-1} and σ^{-1} factors. More precisely, $\psi_1(\varepsilon, \sigma) = \widetilde{O}(\psi_2(\varepsilon, \sigma))$ if there exist constants $C > 0$, a, b such that, for all $\varepsilon > 0$, $\sigma \in (0, 1)$, $\psi_1(\varepsilon, \sigma) \leq C\psi_2(\varepsilon, \sigma) \ln^a \frac{1}{\varepsilon} \ln^b \frac{1}{\sigma}$. We use $O(\cdot)$-notation when $a = b = 0$. For a function $\xi(\varepsilon)$, where $\varepsilon \in \mathbb{R}_+$ we write $\xi(\varepsilon) = \mathbf{poly}(\varepsilon)$ if $\xi(\cdot) = \widetilde{O}(f(\varepsilon))$, where $f(\varepsilon)$ is a polynomial function of ε with non-negative, possibly fractional powers. For a function $\xi(\varepsilon, \sigma)$, where $\varepsilon, \sigma \in \mathbb{R}_+$ we write $\xi(\varepsilon, \sigma) = \mathbf{poly}(\varepsilon, \sigma)$ if $\xi(\cdot, \sigma)$ is a polynomial function of ε and $\xi(\varepsilon, \cdot)$ is a polynomial function of σ.

Problem formulation. The main problem formulation, we are interested in, is the composite strongly-convex-strongly-concave saddle-point problem:

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} \{f(x) + G(x, y) - h(y)\}, \quad G(x, y) := \frac{1}{m}\sum_{i=1}^{m} G_i(x, y). \tag{1.7}$$

We postpone the consideration of convex-strongly-concave and convex-concave problems until Section 5, where these cases will be considered by reduction to the strongly-convex-strongly-concave setting. For now, we make the following assumption.

Assumption 1.

1. $f(x)$ is μ_x-strongly-convex, $h(y)$ is μ_y-strongly-convex, where $\mu_x, \mu_y > 0$;
2. Each function $G_i(x, y)$, $i \in 1, \dots, m$ is continuously differentiable, convex-concave and $G(x, y)$ is L-average-smooth, i.e., for each $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

$$\frac{1}{m}\sum_{i=1}^{m} \|\nabla G_i(x_1, y_1) - \nabla G_i(x_2, y_2)\| \leq L\|(x_1, y_1) - (x_2, y_2)\|; \tag{1.8}$$

3. $f(x), h(y)$ are prox-friendly (smoothness is not required).

Remark 1. As an example of composite terms f and h we can consider the elastic net regularization [44]: $f(x) = \lambda_{1,x}\|x\|_1 + \lambda_{2,x}\|x\|^2$, $h(y) = \lambda_{1,y}\|y\|_1 + \lambda_{2,y}\|y\|^2$, where $\|\cdot\|_1$ denotes the ℓ_1-norm of a vector. We also can "move" strong convexity between the coupling term G and the composites f, h. For that, assume that $G(x, y)$ is (μ_x, μ_y)-strongly-convex-strongly-concave and f, h are just convex. Then, we can define $\hat{G}(x, y) = G(x, y) - \frac{\mu_x}{2}\|x\|^2 + \frac{\mu_y}{2}\|y\|^2$, $\hat{f}(x) = f(x) + \frac{\mu_x}{2}\|x\|^2$, $\hat{h}(y) = h(y) + \frac{\mu_y}{2}\|y\|^2$. It is easy to see that $\hat{f}(x)$ is μ_x-strongly-convex and $\hat{h}(y)$ is μ_y-strongly-convex, and, since f, h are prox-friendly [35], so are the new functions $\hat{f}(x), \hat{h}(y)$. Thus, in general, we can consider any convex prox-friendly composite terms, e.g., indicator functions of convex sets, regularizers, etc.

Paper organization. We start with describing two building blocks for our algorithm: the Catalyst framework [20] adapted and slightly generalized for our setting and variance-reduced algorithm SAGA proposed in [34] which we also adapt to our setting (Section 2). The former algorithm is an optimization algorithm, the latter is designed for saddle-point problems, and we use these algorithms in the system of inner-outer loops, each of which is designed to solve a special optimization subproblem up to a chosen accuracy. To be able to connect the output of an inner loop with the requirement of an outer loop, we prove several technical lemmas (Section 3). After that, we collect all the pieces together and describe the loops of our algorithm as well as present its complexity theorem for the strongly-convex-strongly-concave case (Section 4). Finally, we present the regularization idea and complexity theorem for convex-strongly-concave and convex-concave cases (Section 5).

2. Algorithmic building blocks

In this section, we describe two algorithms that are used as building blocks in our algorithm to find an (ε, σ)-solution to problem (1.7) under Assumption 1. Namely, we

describe the Catalyst Meta-Algorithm [20,21] and the SAGA algorithm [34]. For each algorithm, we describe the problem that is solved by this algorithm under certain assumptions, describe the algorithm itself, and formulate convergence rate and complexity theorems.

2.1. Catalyst meta-algorithm [20,21]

In this subsection, we focus on the optimization problem of the form

$$\min_{x\in\mathbb{R}^{d_x}} F(x) := \varphi(x) + \psi(x) \tag{2.1}$$

under the following assumption:

Assumption 2. $\varphi(x)$ is μ-strongly-convex with $\mu > 0$ and $\psi(x)$ is L_ψ-smooth and convex.

We denote by x^* the solution to this problem. Assume that problem (2.1) can be solved by a linearly convergent method $\mathcal{M}$. One way to accelerate this method is to apply the Catalyst algorithm [20,21] with the inner method $\mathcal{M}$, which is a special case of Accelerated Proximal Point Algorithm where the proximal step is computed inexactly by the inner method $\mathcal{M}$. The resulting algorithm is listed below as Algorithm 1. Each iteration k of this algorithm requires to find an approximate solution to a special optimization problem with accuracy ε_k. Below, in Theorem 1, we show how to choose the sequence $(\varepsilon_k)_{k\geq 0}$ in order to guarantee that Algorithm 1 outputs an ε-solution to problem (2.1) (cf. (1.5)). Further, to be able to use a randomized method as the inner method $\mathcal{M}$, we study the case when the auxiliary problem in each iteration k of Algorithm 1 is solved inexactly with the required accuracy, but only with some probability given by a confidence level σ_k. The total complexity of Algorithm 1 with a randomized inner method $\mathcal{M}$ together with sufficient conditions on $(\varepsilon_k)_{k\geq 0}$ and $(\sigma)_{k\geq 0}$ are given in Theorem 2, which is the main theorem of this subsection.

As said, we start with the result on a sufficient accuracy of the solution to the auxiliary problem in each iteration of Algorithm 1.

Theorem 1. *Let us define $q = \mu/(\mu + H)$ and consider Algorithm 1 satisfying*

$$\varepsilon_k = \frac{2}{9}(F(x_0) - F(x^*))(1-\rho)^k \quad \text{with} \quad \rho = 0.9\sqrt{q},\ k \geq 0. \tag{2.2}$$

Then, after at most $\mathcal{N} = \widetilde{O}\left(\sqrt{1+\frac{H}{\mu}}\right)$ iterations of Algorithm 1, we get $x_{\mathcal{N}}$ such that $F(x_{\mathcal{N}}) - F(x^*) \leq \varepsilon$.

Proof. Recall that $\rho = 0.9\sqrt{q} = 0.9\sqrt{\mu/(\mu+H)}$ and define $C = 8/(\sqrt{q}-\rho)^2$. Choosing

Algorithm 1 Catalyst [20,21].

Input: Starting point $x_0 \in \mathbb{R}^{d_x}$, algorithm parameters $H, \alpha_0 > 0$, strong convexity parameter μ, optimization method $\mathcal{M}$ and a sequence $(\varepsilon_k)_{k\geq 0}$.
Initialize $x_0^{md} = x_0$, $q = \frac{\mu}{\mu+H}$.
for $k = 0, 1, \ldots$ **do**
  Find an approximate solution of the following problem by $\mathcal{M}$

$$x_k \approx \arg\min_{x\in\mathbb{R}^{d_x}} \left\{ S_k(x) := F(x) + \frac{H}{2}\|x - x_{k-1}^{md}\|^2 \right\}$$

  such that $S_k(x_k) - S_k(x_k^*) \leq \varepsilon_k$, where $x_k^* = \arg\min_{x\in\mathbb{R}^{d_x}} S_k(x)$;
  Update $\alpha_k \in (0,1)$ from equation $\alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 + q\alpha_k$;
  Compute x_k^{md} using Nesterov's extrapolation step

$$x_k^{md} = x_k + \beta_k(x_k - x_{k-1}) \quad \text{with} \quad \beta_k = \frac{\alpha_{k-1}(1-\alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}.$$

end for
Output: x_k (final estimate).

$$\begin{aligned}\mathcal{N} &= \left\lceil \frac{1}{\rho} \ln \frac{C(1-\rho)(F(x_0) - F(x^*))}{\varepsilon} \right\rceil = \left\lceil \frac{1}{\rho} \ln \frac{8(1-\rho)(F(x_0) - F(x^*))}{(\sqrt{q}-\rho)^2\varepsilon} \right\rceil \\ &= \left\lceil \frac{\sqrt{\mu+H}}{0.9\sqrt{\mu}} \ln \frac{8(1-0.9\sqrt{\mu/(\mu+H)})(F(x_0) - F(x^*))(H+\mu)}{0.01\mu\varepsilon} \right\rceil \\ &= \widetilde{O}\left(\sqrt{1+\frac{H}{\mu}}\right), \end{aligned} \tag{2.3}$$

using the choice of ε_k according to (2.2), by Theorem 3.1 in [20], we obtain

$$\begin{aligned} F(x_{\mathcal{N}}) - F(x^*) &\leq C(1-\rho)^{\mathcal{N}+1}(F(x_0) - F(x^*)) \\ &\leq C(1-\rho)e^{-\rho\mathcal{N}}(F(x_0) - F(x^*)) \leq \varepsilon. \quad \square \end{aligned}$$

The following theorem is the main theorem of this subsection. It gives sufficient conditions for the accuracy of solution to the auxiliary problem in each iteration of Algorithm 1 and its overall complexity.

Theorem 2. *Let us define $q = \mu/(\mu+H)$ and let Assumption 2 hold. Assume also that at each iteration $k \geq 0$ of Algorithm 1 we find x_k such that $S_k(x_k) - S_k(x_k^*) \leq \varepsilon_k$ with probability at least $1-\sigma_k$, where*

$$0 < \varepsilon_k \leq \frac{2}{9}(F(x_0) - F(x^*))(1-\rho)^k = \boldsymbol{poly}(\varepsilon) \quad \textit{with} \quad \rho = 0.9\sqrt{q}, \tag{2.4}$$

$$0 < \sigma_k \leq \sigma_a = \frac{\sigma}{\frac{\sqrt{\mu+H}}{0.9\sqrt{\mu}} \ln \frac{8(1-0.9\sqrt{\mu/(\mu+H)})(F(x_0)-F(x^*))(H+\mu)}{0.01\mu\varepsilon}} = \boldsymbol{poly}(\varepsilon, \sigma). \tag{2.5}$$

Then, after

$$\mathcal{N} = \widetilde{O}\left(\sqrt{1+\frac{H}{\mu}}\right)$$

iterations, Algorithm 1 finds an (ε, σ)*-solution to problem* (2.1)*, i.e.,* (1.5) *holds.*

Proof. We start by showing that $\varepsilon_k = \mathbf{poly}(\varepsilon)$ for all $k = 0, ..., \mathcal{N}$. Indeed,

$$\begin{aligned}
\varepsilon_k &= \frac{2}{9}(F(x_0) - F(x^*))(1-\rho)^k \leq \frac{2}{9}(F(x_0) - F(x^*))e^{-\rho k} \\
&= \frac{2}{9}(F(x_0) - F(x^*))e^{-\rho\mathcal{N}\frac{k}{\mathcal{N}}} \\
&\overset{(2.3)}{\leq} \frac{2}{9}(F(x_0) - F(x^*))\left(\frac{\varepsilon}{C(1-\rho)(F(x_0) - F(x^*))}\right)^{\frac{k}{\mathcal{N}}} \\
&= \frac{2}{9}\frac{\varepsilon^{\frac{k}{\mathcal{N}}}(F(x_0) - F(x^*))^{\frac{\mathcal{N}-k}{\mathcal{N}}}}{C^{\frac{k}{\mathcal{N}}}(1-\rho)^{\frac{k}{\mathcal{N}}}}.
\end{aligned}$$

Since $k \leq \mathcal{N}$, we have that $\varepsilon_k = \mathbf{poly}(\varepsilon)$.

To show the theorem result, we notice that, by the theorem assumptions on ε_k, at each iteration of Algorithm 1 condition (2.2) is satisfied with probability at least $1 - \sigma_k$, where σ_k is given in (2.5). Thus, by the union bound, after $\mathcal{N}$ iterations, where $\mathcal{N}$ is given in (2.3), condition (2.2) is satisfied in all the $\mathcal{N}$ iterations with probability at least $\prod_{i=1}^{\mathcal{N}}(1-\sigma_i)$. Combining (2.3), (2.5), we obtain that

$$\prod_{i=1}^{\mathcal{N}}(1-\sigma_i) \geq (1-\sigma_a)^{\mathcal{N}} = \left(1 - \frac{\sigma}{\mathcal{N}}\right)^{\mathcal{N}} \geq 1 - \sigma.$$

Thus, with probability at least $1-\sigma$ the conditions of Theorem 1 are satisfied, which guarantees that $F(x_{\mathcal{N}}) - F(x^*) \leq \varepsilon$ with probability at least $1-\sigma$, where $\mathcal{N} = \widetilde{O}\left(\sqrt{1+\frac{H}{\mu}}\right)$ is given in (2.3). □

In contrast to the Catalyst algorithm in [20,21], we analyze Algorithm 1 in the setting when the auxiliary problem is step 4 is solved inexactly with some probability. In particular, we estimate how to choose this probability to guarantee that the point returned by the algorithm is an (ε, σ)-solution to (2.1) and show that the total complexity remains the same.

2.2. SAGA algorithm [34]

In this subsection, we focus on the optimization problem of the form

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \{K(x,y) + M(x,y)\} \tag{2.6}$$

under the following assumption.

Assumption 3.

1. M is (μ_x, μ_y)-strongly-convex-strongly-concave and has tractable proximal operator

$$prox_M^{\lambda}(x', y') = \arg\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left\{\lambda M(x,y) + \frac{\mu_x}{2}\|x - x'\|^2 - \frac{\mu_y}{2}\|y - y'\|^2\right\};$$

2. K is convex-concave;
3. The vector-valued function $B(x,y) = (\nabla_x K(x,y), -\nabla_y K(x,y)) \in \mathbb{R}^{d_x+d_y}$ may be split into a family of vector-valued functions as $B = \sum\limits_{i\in\mathcal{J}} B_i$, where $B_i(x,y) = (\nabla_x K_i(x,y), -\nabla_y K_i(x,y))$, and B is L-average-smooth.

Under Assumption 3.3, problem (2.6) has the finite-sum structure, motivating the use of stochastic variance-reduced methods. Instead of expensive calculation of the full operator B in each iteration, such methods pick at random one B_i and only rarely calculate the full operator B. In particular, problem (2.6) under Assumption 3 can be solved by the SAGA algorithm proposed in [34] and listed below as Algorithm 2. Next, we, first, show in Lemma 1 that a problem with the form (1.7) satisfies Assumption 3. After that, in the main theorem of this subsection (Theorem 3), we give the complexity of the SAGA algorithm.

Algorithm 2 SAGA: online stochastic variance reduction for saddle points [34].

Input: Function M, operators $(B_i)_{i=1}^m$, probabilities $(\pi_i)_{i=1}^m$, smoothness constants $\bar{L}(\pi)$ (see (2.11)) and L, starting point $z_0 = (x_0, y_0)$, number of iterations t, number of updates per iteration (mini-batch size) s.
Set $\lambda = \left(\max\left\{\frac{3|\mathcal{J}|}{2s} - 1, L^2 + \frac{3\bar{L}^2}{s}\right\}\right)^{-1}$;
Initialize $w^i = B_i(x_0, y_0)$ for all $i \in \mathcal{J}$ and $W = \sum\limits_{i\in\mathcal{J}} w^i$;
for $l = 1$ to t **do**
  Sample $i_1, \ldots, i_s \in \mathcal{J}$ from the probability vector $(\pi_i)_{i=1}^m$ with replacement;
  Compute $v_k = B_{i_k}(x_l, y_l)$ for $k \in \{1, \ldots, s\}$;
  (x_l, y_l)
  $= prox_M^{\lambda}\left\{(x_{l-1}, y_{l-1}) - \lambda \begin{pmatrix} \frac{1}{\mu_x} & 0 \\ 0 & \frac{1}{\mu_y} \end{pmatrix} \left(W + \frac{1}{s}\sum_{k=1}^s \left\{\frac{1}{\pi_{i_k}} v_k - \frac{1}{\pi_{i_k}} w^{i_k}\right\}\right)\right\}$;
  Replace $W = W - \sum_{k=1}^s \{w^{i_k} - v_k\}$ and $w^{i_k} = v_k$ for $k \in \{1, \ldots, s\}$.
end for
Output: Approximate solution $z_t = (x_t, y_t)$.

Lemma 1. *Consider the following special case of problem (2.6) with* $M(x,y) = \tilde{f}(x) - \tilde{h}(y), K(x,y) = \frac{1}{m}\sum_{i=1}^m G_i(x,y)$*:*

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left\{ \tilde{f}(x) + \frac{1}{m}\sum_{i=1}^m G_i(x,y) - \tilde{h}(y) \right\}, \tag{2.7}$$

where $\tilde{f}(x)$ *is prox-friendly* $\tilde{\mu}_x$*-strongly-convex,* $\tilde{h}(y)$ *is prox-friendly* $\tilde{\mu}_y$*-strongly-convex, each function* $G_i(x,y)$ *is continuously differentiable, convex-concave, and* $\frac{1}{m}\sum_{i=1}^m G_i(x,y)$ *is* L*-average-smooth. Under these assumptions, problem (2.7) satisfies Assumption 3.*

Proof. 1. Since $\tilde{f}(x)$ is $\tilde{\mu}_x$-strongly-convex, $-h(y)$ is $\tilde{\mu}_y$-strongly-concave, $M(x,y)$ is $(\tilde{\mu}_x, \tilde{\mu}_y)$-strongly-convex-strongly-concave. Moreover,

$$\begin{aligned}
&prox_M^{\lambda}(x',y') \\
&= \arg\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left\{ \lambda M(x,y) + \frac{\tilde{\mu}_x}{2}\|x-x'\|^2 - \frac{\tilde{\mu}_y}{2}\|y-y'\|^2 \right\} \\
&= \arg\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left\{ \lambda(\tilde{f}(x) - \tilde{h}(y)) + \frac{\tilde{\mu}_x}{2}\|x-x'\|^2 - \frac{\tilde{\mu}_y}{2}\|y-y'\|^2 \right\} \\
&= \left(\arg\min_{x\in\mathbb{R}^{d_x}} \left\{ \lambda\tilde{f}(x) + \frac{\tilde{\mu}_x}{2}\|x-x'\|^2 \right\}, \arg\max_{y\in\mathbb{R}^{d_y}} \left\{ -\lambda\tilde{h}(y) - \frac{\tilde{\mu}_y}{2}\|y-y'\|^2 \right\} \right) \\
&= \left(prox_{\tilde{f}}^{\lambda/\tilde{\mu}_x}(x'), prox_{\tilde{h}}^{\lambda/\tilde{\mu}_y}(y') \right).
\end{aligned}$$

Thus, since $\tilde{f}(x), \tilde{h}(y)$ are prox-friendly, $\left(prox_{\tilde{f}}^{\lambda/\tilde{\mu}_x}(x'), prox_{\tilde{h}}^{\lambda/\tilde{\mu}_y}(y') \right)$ is easy to compute, $prox_M^{\lambda}(x',y')$ is easy to compute as well, and Assumption 3.1 holds.

2. Assumption 3.2 holds since $K(x,y) = \frac{1}{m}\sum_{i=1}^m G_i(x,y)$ is convex-concave.

3. Defining $B_i(x,y) = \frac{1}{m}(\nabla_x G_i(x,y), -\nabla_y G_i(x,y))$, we see that

$$\begin{aligned}
B(x,y) &= (\nabla_x K(x,y), -\nabla_y K(x,y)) \\
&= \left(\nabla_x \frac{1}{m}\sum_{i=1}^m G_i(x,y), -\nabla_y \frac{1}{m}\sum_{i=1}^m G_i(x,y) \right) \\
&= \frac{1}{m}\sum_{i=1}^m (\nabla_x G_i(x,y), -\nabla_y G_i(x,y)) = \sum_{i=1}^m B_i(x,y).
\end{aligned}$$

Since $\frac{1}{m}\sum_{i=1}^m G_i(x,y)$ is L-average-smooth, we have, for all $(x_1,y_1), (x_2,y_2)$,

$$\begin{aligned}
\sum_{i=1}^m \|B_i(x_1,y_1) - B_i(x_2,y_2)\| &= \frac{1}{m}\sum_{i=1}^m \|\nabla G_i(x_1,y_1) - \nabla G_i(x_2,y_2)\| \\
&\le L\|(x_1,y_1) - (x_2,y_2)\|.
\end{aligned}$$

Thus, Assumption 3.3 holds. □

In the next theorem, we state the iteration complexity of Algorithm 2 to find an (ε, σ)-solution to problem (2.7). Note that this is in contrast to [34], where the result is stated in terms of expectation. Let R_0 be a number such that for the solution $z^* = (x^*, y^*)$ of (2.7) and starting point $z_0 = (x_0, y_0)$ of Algorithm 2 it holds that $\|z_0 - z^*\| \leq R_0$. Since $\tilde{f}$ and $\tilde{h}$ are convex, by Theorem 3.1.8 in [29], the function $\tilde{f}(x) + \tilde{h}(y)$ is Lipschitz-continuous with constant $M_{\tilde{f}+\tilde{h}}$ on the ball $\mathcal{B}_2(z_0, 2R_0)$. More precisely, for $z_1 = (x_1, y_1), z_2 = (x_2, y_2) \in \mathcal{B}_2(z_0, 2R_0)$,

$$\left|\tilde{f}(x_1) + \tilde{h}(y_1) - \tilde{f}(x_2) - \tilde{h}(y_2)\right| \leq M_{\tilde{f}+\tilde{h}}\|z_1 - z_2\|. \tag{2.8}$$

Theorem 3. *Let the assumptions of Lemma 1 hold and let $\varepsilon', \sigma' > 0$ satisfy*

$$\varepsilon' \leq \min\left\{\varepsilon, R_0^2, \frac{\varepsilon}{\left(2L + \frac{L^2}{\tilde{\mu}_y} + \frac{L^2}{\tilde{\mu}_x}\right)}, \frac{\varepsilon^2}{4M_{\tilde{f}+\tilde{h}}{}^2}\right\} = \boldsymbol{poly}(\varepsilon), \quad \sigma' \leq \sigma = \boldsymbol{poly}(\sigma), \tag{2.9}$$

where $\varepsilon > 0$, $\sigma \in (0, 1)$. Then, after

$$\mathcal{N} = 6\left(m + \frac{L^2}{(\min\{\tilde{\mu}_x, \tilde{\mu}_y\})^2}\right)\ln\frac{2R_0^2}{\varepsilon'\sigma'} = \widetilde{O}\left(m + \frac{L^2}{(\min\{\tilde{\mu}_x, \tilde{\mu}_y\})^2}\right) \tag{2.10}$$

iterations, Algorithm 2 (with $s = 1$, $\pi_i = \frac{1}{m}$ for $i = 1, ..., m$) finds an (ε, σ)-solution to problem (2.7), i.e., (1.6) holds.

Proof. The proof is organized in three steps and relies on the complexity Theorem 2 in [34]. The first step is to estimate several constants that are used in that complexity theorem. The second step is to apply that theorem and show that the output $z_{\mathcal{N}}$ of Algorithm 2 is sufficiently close to the solution z^*. The final step is to show that $\hat{z} = z_{\mathcal{N}}$ also satisfies (1.6).

Step 1. The goal is to define and estimate for problem (2.7) the constants L_{A_2} (in [34] this constant is denoted as L), $\bar{L}$, μ that are used in Appendix A, D.2 of [34]. We start by defining the operators $A_1(x, y), A_2(x, y)$ that are used in Appendix A of [34]:

$$A_1 = \left(\frac{1}{\tilde{\mu}_x}\partial\tilde{f}(x), \frac{1}{\tilde{\mu}_y}\partial\tilde{h}(y)\right), \text{ and } A_2(x, y) = \left(\frac{1}{\tilde{\mu}_x}\nabla_x G(x, y), -\frac{1}{\tilde{\mu}_y}\nabla_y G(x, y)\right).$$

Clearly,

$$A_2(x, y) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\tilde{\mu}_x}\nabla_x G_i(x, y), -\frac{1}{\tilde{\mu}_y}\nabla_y G_i(x, y)\right) =: \sum_{i=1}^{m} A_2^i(x, y).$$

The constants $\mu, L_{A_2}, \bar{L}$ are defined as follows.

(a) μ is the strong monotonicity constant of the operator A_1. Using that $\tilde{f}(x)$ is $\tilde{\mu}_x$-strongly-convex, $\tilde{h}(y)$ is $\tilde{\mu}_y$-strongly-convex, we have:

$$\begin{aligned}
&(A_1(z) - A_1(z'))^T(z - z') \\
&= \frac{1}{\tilde{\mu}_x}(\partial\tilde{f}(x) - \partial\tilde{f}(x'))(x - x') + \frac{1}{\tilde{\mu}_y}(\partial\tilde{h}(y) - \partial\tilde{h}(y'))(y - y') \\
&\geq \|x - x'\|^2 + \|y - y'\|^2 \geq \|z - z'\|^2.
\end{aligned}$$

Thus, $A_1(x, y)$ is μ-strongly-monotone with $\mu = 1$.

(b) L_{A_2} is the Lipschitz constant of $A_2(x, y)$ with respect to the Euclidean norm. Taking arbitrary $z = (x, y), z' = (x', y') \in \mathbb{R}^{d_x+d_y}$, we have

$$\begin{aligned}
&\|A_2(x, y) - A_2(x', y')\| \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\tilde{\mu}_x}\|\nabla_x G_i(x, y) - \nabla_x G_i(x', y')\| + \frac{1}{\tilde{\mu}_y}\|\nabla_y G_i(x, y) - \nabla_y G_i(x', y')\|\right) \\
&\leq L\left(\frac{1}{\tilde{\mu}_x} + \frac{1}{\tilde{\mu}_y}\right)\|z - z'\| \leq \frac{2L}{\min\{\tilde{\mu}_x, \tilde{\mu}_y\}}\|z - z'\|
\end{aligned}$$

by the L-average-smoothness of $G(x, y)$. Thus, $L_{A_2} \leq \frac{2L}{\min\{\tilde{\mu}_x, \tilde{\mu}_y\}}$.

(c) The constant $\bar{L}^2 = \bar{L}^2(\pi)$ is defined as follows:

$$\begin{aligned}
\bar{L}^2 &:= \sup_{z,z' \in \mathbb{R}^{d_x+d_y}} \frac{1}{\|z - z'\|^2}\sum_{i=1}^{m}\frac{1}{\pi_i}\left\|A_2^i(x, y) - A_2^i(x', y')\right\|^2 \qquad (2.11) \\
&= \sup_{z,z' \in \mathbb{R}^{d_x+d_y}} \frac{1}{\|z - z'\|^2}\sum_{i=1}^{m} m\left\|A_2^i(x, y) - A_2^i(x', y')\right\|^2 \\
&= \sup_{z,z' \in \mathbb{R}^{d_x+d_y}} \frac{1}{\|z - z'\|^2}\sum_{i=1}^{m}\frac{m}{m^2}\frac{1}{\tilde{\mu}_x^2}\|\nabla_x G_i(x, y) - \nabla_x G_i(x', y')\|^2 \\
&+ \frac{1}{\|z - z'\|^2}\sum_{i=1}^{m}\frac{m}{m^2}\frac{1}{\tilde{\mu}_y^2}\|\nabla_y G_i(x, y) - \nabla_y G_i(x', y')\|^2 \\
&\leq \sup_{z,z' \in \mathbb{R}^{d_x+d_y}} \frac{1}{\|z - z'\|^2}L^2\left(\frac{1}{\tilde{\mu}_x^2} + \frac{1}{\tilde{\mu}_y^2}\right)\|z - z'\|^2 \leq \frac{2L^2}{\min\{\tilde{\mu}_x, \tilde{\mu}_y\}^2}
\end{aligned}$$

by the L-average-smoothness of $G(x, y)$. Thus, $\bar{L}^2 \leq \frac{2L^2}{\min\{\tilde{\mu}_x, \tilde{\mu}_y\}^2}$. **Step 2.** By Lemma 1, problem (2.7) satisfies Assumption 3 and, by Theorem 2 in [34], after t iteration of the SAGA Algorithm 2 with $s = 1$, $\pi_i = \frac{1}{m}$ for $i = 1, ..., m$, we have

$$\mathbb{E}\|z_t - z^*\|^2 \leq 2\left(1 - \frac{1}{4}\left(\max\left\{\frac{3|\mathcal{J}|}{2}, 1 + \frac{L_{A_2}^2}{\mu^2} + \frac{3\bar{L}^2}{\mu^2}\right\}\right)^{-1}\right)^t \|z_0 - z^*\|^2,$$

where $|\mathcal{J}| = m$ and $\mu, L_{A_2}, \bar{L}$ are defined above. Defining

$$\eta = \left(\max\left\{\frac{3|\mathcal{J}|}{2}, 1 + \frac{L_{A_2}^2}{\mu^2} + \frac{3\bar{L}^2}{\mu^2}\right\}\right)^{-1} = \left(\max\left\{\frac{3m}{2}, \frac{5\bar{L}^2}{\min\{\tilde{\mu}_x, \tilde{\mu}_y\}^2}\right\}\right)^{-1}$$

and taking $t = \mathcal{N}$, where $\mathcal{N}$ is defined in (2.10), we get $(\hat{x}, \hat{y}) = \hat{z} = z_{\mathcal{N}}$ s.t.

$$\mathbb{E}\|\hat{z} - z^*\|^2 = \mathbb{E}\|z_{\mathcal{N}} - z^*\|^2 \leq 2\left(1 - \frac{\eta}{4}\right)^{\mathcal{N}} \|z_0 - z^*\|^2 \leq 2e^{-\frac{\eta}{4}\mathcal{N}} R_0^2 \leq \varepsilon'\sigma'.$$

Since $\|\hat{z} - z^*\|^2 \geq 0$, by the Markov's inequality, we have

$$\mathbb{P}(\|\hat{z} - z^*\|^2 \leq \varepsilon') \geq 1 - \frac{\mathbb{E}\|\hat{z} - z^*\|^2}{\varepsilon'} \geq 1 - \sigma',$$

and with probability at least $1 - \sigma'$, $(\hat{x}, \hat{y}) = \hat{z}$ satisfies

$$\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2 = \|\hat{z} - z^*\|^2 \leq \sigma'\varepsilon' \leq \varepsilon'. \tag{2.12}$$

Step 3. By (2.12), since, by (2.9), $\varepsilon' \leq R_0^2$, with probability at least $1 - \sigma'$

$$\|\hat{z} - z_0\|^2 \leq 2(\|\hat{z} - z^*\|^2 + \|z^* - z_0\|^2) \leq 4R_0^2.$$

Thus, with probability at least $1 - \sigma'$, $\hat{z}, z^* \in \mathcal{B}_2(z_0, 2R_0)$ and, by (2.8),

$$\left|\tilde{f}(\hat{x}) + \tilde{h}(\hat{y}) - \tilde{f}(x^*) - \tilde{h}(y^*)\right| \leq M_{\tilde{f}+\tilde{h}}\|\hat{z} - z^*\|. \tag{2.13}$$

Since L-average-smoothness of $G(x, y)$ implies its L-smoothness, the function $u(x) = \max_{y \in \mathbb{R}^{d_y}}\{G(x, y) - \tilde{h}(y)\}$ is $\left(L + \frac{L^2}{\tilde{\mu}_y}\right)$-smooth by Lemma 2, and the function $w(y) = -\min_{x \in \mathbb{R}^{d_x}}\{\tilde{f}(x) + G(x, y)\} = \max_{x \in \mathbb{R}^{d_x}}\{-\tilde{f}(x) - G(x, y)\}$ is $\left(L + \frac{L^2}{\tilde{\mu}_x}\right)$-smooth by Lemma 2. Thus, with probability at least $1 - \sigma' \overset{(2.9)}{\geq} 1 - \sigma$

$$\begin{aligned}
&\max_{y \in \mathbb{R}^{d_y}} \left\{\tilde{f}(\hat{x}) + G(\hat{x}, y) - \tilde{h}(y)\right\} - \min_{x \in \mathbb{R}^{d_x}} \left\{\tilde{f}(x) + G(x, \hat{y}) - \tilde{h}(\hat{y})\right\} \\
&= \max_{y \in \mathbb{R}^{d_y}} \left\{\tilde{f}(\hat{x}) + G(\hat{x}, y) - \tilde{h}(y)\right\} - \{\tilde{f}(x^*) + G(x^*, y^*) - \tilde{h}(y^*)\} \\
&+ \{\tilde{f}(x^*) + G(x^*, y^*) - \tilde{h}(y^*)\} - \min_{x \in \mathbb{R}^{d_x}} \left\{\tilde{f}(x) + G(x, \hat{y}) - \tilde{h}(\hat{y})\right\} \\
&= \tilde{f}(\hat{x}) - \tilde{f}(x^*) + u(\hat{x}) - u(x^*) + \tilde{h}(\hat{y}) - \tilde{h}(y^*) + w(\hat{y}) - w(y^*) \\
&\overset{(2.13)}{\leq} M_{\tilde{f}+\tilde{h}}\|\hat{z} - z^*\| + \frac{1}{2}\left(L + \frac{L^2}{\tilde{\mu}_y}\right)\|\hat{x} - x^*\|^2 + \frac{1}{2}\left(L + \frac{L^2}{\tilde{\mu}_x}\right)\|\hat{y} - y^*\|^2 \\
&\leq M_{\tilde{f}+\tilde{h}}\|\hat{z} - z^*\| + \frac{1}{2}\left(2L + \frac{L^2}{\tilde{\mu}_y} + \frac{L^2}{\tilde{\mu}_x}\right)\|\hat{z} - z^*\|^2
\end{aligned}$$

$$\overset{(2.12)}{\leq} M_{\tilde{f}+\tilde{h}}\sqrt{\varepsilon'} + \frac{1}{2}\left(2L + \frac{L^2}{\tilde{\mu}_y} + \frac{L^2}{\tilde{\mu}_x}\right)\varepsilon' \overset{(2.9)}{\leq} \varepsilon.$$

Thus, $(\hat{x}, \hat{y})$ is an (ε, σ)-solution to problem (2.7) (i.e., (1.6) holds). Moreover, $(\hat{x}, \hat{y})$ is found at most after $\mathcal{N}$ iterations, where $\mathcal{N}$ is defined in (2.10). □

3. Preliminaries

In this section, we provide several technical results necessary for the analysis of our main algorithm. First, in Lemma 2 we state the properties of an implicit function given as a solution to parametric maximization problem. After that, we give three lemmas related to reformulations of saddle-point problems in the form of nested optimization problems: outer minimization and inner maximization problems. Namely, the saddle-point problem 1) $\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \tilde{F}(x,y)$, where $\tilde{F}(x,y)$ has similar structure to the objective in (1.7), can be reformulated in two ways:

$$2) \min_{x\in\mathbb{R}^{d_x}} \left\{\hat{g}(x) = \max_{y\in\mathbb{R}^{d_y}} \tilde{F}(x,y)\right\}, \text{ and } 3) \min_{y\in\mathbb{R}^{d_x}} \left\{-\tilde{g}(y) = \max_{x\in\mathbb{R}^{d_x}} (-\tilde{F}(x,y))\right\},$$

and the main goal of the three technical lemmas is to relate to each other (ε, σ)-solutions to these reformulations. We show that an approximate solution to 2) is an approximate solution to 1) (Lemma 3); that an approximate solution to 3) is an approximate solution to 2) (Lemma 4); that an approximate solution to 1) is an approximate solution to 3) (Lemma 5).

Lemma 2. *Let us consider the function*

$$p(x) = \max_{y\in\mathbb{R}^{d_y}} \left\{\hat{S}(x,y) = F(x,y) - q(y)\right\}, \tag{3.1}$$

where $F(x,y)$ is convex in x, concave in y and is L_F-smooth as a function of (x,y), $q(y)$ is μ_q-strongly-convex. Then, $p(x)$ is L_p-smooth with $L_p = L_F + \frac{L_F^2}{\mu_q}$ and $y^*(\cdot)$ is $\left(\frac{L_F}{\mu_q}\right)$-Lipschitz continuous, where the point y^* is defined as

$$y^*(x) := \arg\max_{y\in\mathbb{R}^{d_y}} \hat{S}(x,y), \quad \forall x \in \mathbb{R}^{d_x}.$$

Proof. The function $\hat{S}(x,\cdot)$ is μ_q-strongly-concave, and $\hat{S}(\cdot,y)$ is differentiable. Therefore, by the Demyanov–Danskin's theorem, we have

$$\nabla p(x) = \nabla_x \hat{S}(x, y^*(x)) = \nabla_x F(x, y^*(x)), \quad \forall x \in \mathbb{R}^{d_x}. \tag{3.2}$$

To prove that $\nabla p(\cdot)$ is L_p–Lipschitz with $L_p = L_F + \frac{L_F^2}{\mu_q}$, we first prove the Lipschitz condition for $y^*(\cdot)$. Since $\hat{S}(x,\cdot)$ is μ_q-strongly-concave, we have, for arbitrary $x_1, x_2 \in \mathbb{R}^{d_x}$,

$$\|y^*(x_1) - y^*(x_2)\|^2 \leq \frac{2}{\mu_q}\left(\hat{S}(x_1, y^*(x_1)) - \hat{S}(x_1, y^*(x_2))\right),$$
$$\|y^*(x_2) - y^*(x_1)\|^2 \leq \frac{2}{\mu_q}\left(\hat{S}(x_2, y^*(x_2)) - \hat{S}(x_2, y^*(x_1))\right). \tag{3.3}$$

At the same time,

$$\begin{aligned}
&\hat{S}(x_1, y^*(x_1)) - \hat{S}(x_1, y^*(x_2)) + \hat{S}(x_2, y^*(x_2)) - \hat{S}(x_2, y^*(x_1)) \\
&= \hat{S}(x_1, y^*(x_1)) - \hat{S}(x_1, y^*(x_2)) - \hat{S}(x_2, y^*(x_1)) + \hat{S}(x_2, y^*(x_2)) \\
&\overset{(3.1)}{=} (F(x_1, y^*(x_1)) - F(x_1, y^*(x_2))) - (F(x_2, y^*(x_1)) - F(x_2, y^*(x_2))) \\
&= \int_0^1 \langle \nabla_x F(x_1 + t(x_2 - x_1), y^*(x_1)) - \nabla_x F(x_1 + t(x_2 - x_1), y^*(x_2)), x_2 - x_1 \rangle dt \\
&\leq \|\nabla_x F(x_1 + t(x_2 - x_1), y^*(x_1)) - \nabla_x F(x_1 + t(x_2 - x_1), y^*(x_1))\| \cdot \|x_2 - x_1\| \\
&\leq L_F \|y^*(x_1) - y^*(x_2)\| \cdot \|x_2 - x_1\|.
\end{aligned} \tag{3.4}$$

Thus, (3.3) and (3.4) imply the inequality

$$\|y^*(x_2) - y^*(x_1)\| \leq \frac{L_F}{\mu_q}\|x_2 - x_1\|, \tag{3.5}$$

i.e., the function $y^*(\cdot)$ satisfies Lipschitz condition with the constant $\frac{L_F}{\mu_q}$. Next, from (3.2), we obtain

$$\begin{aligned}
&\|\nabla p(x_1) - \nabla p(x_2)\| = \|\nabla_x F(x_1, y^*(x_1)) - \nabla_x F(x_2, y^*(x_2))\| = \\
&= \|\nabla_x F(x_1, y^*(x_1)) - \nabla_x F(x_1, y^*(x_2)) + \nabla_x F(x_1, y^*(x_2)) - \nabla_x F(x_2, y^*(x_2))\| \\
&\leq \|\nabla_x F(x_1, y^*(x_1)) - \nabla_x F(x_1, y^*(x_2))\| \\
&\quad + \|\nabla_x F(x_1, y^*(x_2)) - \nabla_x F(x_2, y^*(x_2))\| \\
&\leq L_F \|y^*(x_1) - y^*(x_2)\| + L_F \|x_2 - x_1\| \overset{(3.5)}{\leq} \left(L_F + \frac{L_F^2}{\mu_q}\right)\|x_2 - x_1\|.
\end{aligned}$$

Thus, $p(x)$ has Lipschitz gradient with the constant $L_p = L_F + \frac{L_F^2}{\mu_q}$. □

As mentioned above, we next consider different reformulations of saddle-point problems. First, we consider the reformulation of the strongly-convex-strongly-concave problem

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} \left\{\tilde{f}(x) + G(x, y) - \tilde{h}(y)\right\} \tag{3.6}$$

as an equivalent minimization problem:

$$\min_{x\in\mathbb{R}^{d_x}}\left\{\tilde{f}(x)+\max_{y\in\mathbb{R}^{d_y}}\{G(x,y)-\tilde{h}(y)\}\right\}. \tag{3.7}$$

Our goal is to show that if $\hat{x}$ is an (ε_x,σ_x)-solution to the outer problem in (3.7) and $\hat{y}$ is an (ε_y,σ_y)-solution to the inner problem $\max_{y\in\mathbb{R}^{d_y}}\{G(\hat{x},y)-\tilde{h}(y)\}$, then $(\hat{x},\hat{y})$ is an (ε,σ)-solution to problem (3.6), where the dependencies $\varepsilon_x(\varepsilon)$, $\varepsilon_y(\varepsilon)$, $\sigma_x(\sigma)$, $\sigma_y(\sigma)$ are polynomial. Let R_y be a number such that $R_y^2\geq\varepsilon>0$ and for the solution $z^*=(x^*,y^*)$ of (3.6) it holds that $\|y^*\|\leq R_y$. Since $\tilde{h}$ is convex, by Theorem 3.1.8 in [29], $\tilde{h}(y)$ is Lipschitz-continuous with constant $M_{\tilde{h}}$ on the ball $\mathcal{B}_2(0,2R_y)$.

Lemma 3. *Assume, that in (3.6) $\tilde{f}(x)$ is $\tilde{\mu}_x$-strongly-convex, $\tilde{h}(y)$ is $\tilde{\mu}_y$-strongly-convex, and Assumption 1.2 holds. Let an approximate solution $(\hat{x},\hat{y})$ for (3.7) satisfy*

1. *$\hat{x}$ is an (ε_x,σ_x)-solution to the outer problem in (3.7), i.e., (1.5) holds for the objective $\tilde{f}(x)+\max_{y\in\mathbb{R}^{d_y}}\{G(x,y)-\tilde{h}(y)\}$;*
2. *$\hat{y}$ is an (ε_y,σ_y)-solution to the inner problem $\max_{y\in\mathbb{R}^{d_y}}\{G(\hat{x},y)-\tilde{h}(y)\}$,*

where

$$\varepsilon_y\leq\min\left\{\frac{\tilde{\mu}_y\varepsilon}{8},\frac{\varepsilon^2\tilde{\mu}_y}{72M_{\tilde{h}}^2},\frac{\varepsilon\tilde{\mu}_y}{12\left(L+\frac{L^2}{\tilde{\mu}_x}\right)}\right\}=\boldsymbol{poly}(\varepsilon), \tag{3.8}$$

$$\varepsilon_x\leq\min\left\{\frac{\varepsilon_y\tilde{\mu}_x\tilde{\mu}_y}{L^2},\frac{\varepsilon}{3}\right\}=\boldsymbol{poly}(\varepsilon),\quad \max\{\sigma_x,\sigma_y\}\leq\frac{\sigma}{2}=\boldsymbol{poly}(\sigma) \tag{3.9}$$

for some $\varepsilon>0$, $\sigma\in(0,1)$. Then, $(\hat{x},\hat{y})$ is an (ε,σ)-solution to problem (3.6), i.e., (1.6) holds.

Proof. Under the lemma assumptions, problems (3.6) and (3.7) have the unique solution (x^*,y^*). We denote $\Psi(x)=\tilde{f}(x)+\max_{y\in\mathbb{R}^{d_y}}\{G(x,y)-\tilde{h}(y)\}$ and notice that $\Psi(x)$ is $\tilde{\mu}_x$-strongly-convex and has the minimum at x^*. Thus, by condition 1. of the lemma, with probability at least $1-\sigma_x$,

$$\|\hat{x}-x^*\|^2\leq\frac{2}{\tilde{\mu}_x}\left(\Psi(\hat{x})-\Psi(x^*)\right)\leq\frac{2\varepsilon_x}{\tilde{\mu}_x}.$$

We denote $y^*(x)=\arg\max_{y\in\mathbb{R}^{d_y}}\{G(x,y)-\tilde{h}(y)\}$ and notice that, since L-average-smoothness of $G(x,y)$ implies its L-smoothness, by Lemma 2, $y^*(x)$ is $\left(\frac{L}{\mu_y}\right)$-Lipschitz continuous. Since $G(\hat{x},y)-\tilde{h}(y)$ is $\tilde{\mu}_y$-strongly-concave, by condition 2. of the lemma, we obtain that with probability at least $1-\sigma_x-\sigma_y$,

$$\|\hat{y}-y^*\|^2\leq2\|\hat{y}-y^*(\hat{x})\|^2+2\|y^*(\hat{x})-y^*(x^*)\|^2\leq\frac{4\varepsilon_y}{\tilde{\mu}_y}+2\left(\frac{L}{\tilde{\mu}_y}\right)^2\|\hat{x}-x^*\|^2$$

$$\leq \frac{4\varepsilon_y}{\tilde{\mu}_y} + 4\left(\frac{L}{\tilde{\mu}_y}\right)^2 \frac{\varepsilon_x}{\tilde{\mu}_x} \overset{(3.9)}{\leq} \frac{8\varepsilon_y}{\tilde{\mu}_y} \overset{(3.8)}{\leq} \varepsilon. \tag{3.10}$$

Further, with probability at least $1-\sigma_x-\sigma_y$, since $R_y^2 \geq \varepsilon > 0$, we have $\|\hat{y}\|^2 \leq 2(\|\hat{y}-y^*\|^2+\|y^*\|^2) \leq 4R_y^2$ and, hence, $y^*, \hat{y} \in \mathcal{B}_2(0, 2R_y)$. Thus, since $\tilde{h}(y)$ is Lipschitz-continuous with constant $M_{\tilde{h}}$ on $\mathcal{B}_2(0,2R_y)$ (see the justification before the statement of this lemma), we have with probability at least $1-\sigma_x-\sigma_y$ that $\left|\tilde{h}(\hat{y})-\tilde{h}(y^*)\right| \leq M_{\tilde{h}}\|\hat{y}-y^*\|$. Further, since L-average-smoothness of $G(x,y)$ implies its L-smoothness, by Lemma 2, $w(y) = -\min_{x\in\mathbb{R}^{d_x}}\{\tilde{f}(x)+G(x,y)\}$ is $\left(L+\frac{L^2}{\tilde{\mu}_x}\right)$-smooth. Thus, for $\Phi(y) = \min_{x\in\mathbb{R}^{d_x}}\{\tilde{f}(x)+G(x,y)\} - \tilde{h}(y)$, with probability at least $1-\sigma_x-\sigma_y$ we have:

$$\begin{aligned}
0 &\leq \Phi(y^*) - \Phi(\hat{y}) = \tilde{h}(\hat{y}) - \tilde{h}(y^*) + w(\hat{y}) - w(y^*) \\
&\leq M_{\tilde{h}}\|\hat{y}-y^*\| + \frac{1}{2}\left(L+\frac{L^2}{\tilde{\mu}_x}\right)\|\hat{y}-y^*\|^2 \\
&\overset{(3.10)}{\leq} M_{\tilde{h}}\sqrt{\frac{8\varepsilon_y}{\tilde{\mu}_y}} + \frac{1}{2}\left(L+\frac{L^2}{\tilde{\mu}_x}\right)\left(\frac{8\varepsilon_y}{\tilde{\mu}_y}\right) \\
&\overset{(3.8)}{\leq} M_{\tilde{h}}\sqrt{\frac{8\varepsilon^2\tilde{\mu}_y}{\tilde{\mu}_y 72 M_{\tilde{h}}^2}} + \frac{1}{2}\left(L+\frac{L^2}{\tilde{\mu}_x}\right)\left(\frac{8\varepsilon\tilde{\mu}_y}{\tilde{\mu}_y 12\left(L+\frac{L^2}{\tilde{\mu}_x}\right)}\right) = \frac{2\varepsilon}{3}.
\end{aligned} \tag{3.11}$$

Finally, with probability at least $1-\sigma_x-\sigma_y \overset{(3.9)}{\geq} 1-\sigma$

$$\begin{aligned}
&\max_{y\in\mathbb{R}^{d_y}}\left\{\tilde{f}(\hat{x})+G(\hat{x},y)-\tilde{h}(y)\right\} - \min_{x\in\mathbb{R}^{d_x}}\left\{\tilde{f}(x)+G(x,\hat{y})-\tilde{h}(\hat{y})\right\} \\
&= \max_{y\in\mathbb{R}^{d_y}}\left\{\tilde{f}(\hat{x})+G(\hat{x},y)-\tilde{h}(y)\right\} - \left\{\tilde{f}(x^*)+G(x^*,y^*)-\tilde{h}(y^*)\right\} \\
&+ \left\{\tilde{f}(x^*)+G(x^*,y^*)-\tilde{h}(y^*)\right\} - \min_{x\in\mathbb{R}^{d_x}}\left\{\tilde{f}(x)+G(x,\hat{y})-\tilde{h}(\hat{y})\right\} \\
&= \Psi(\hat{x}) - \Psi(x^*) + \Phi(y^*) - \Phi(\hat{y}) \overset{(3.11)}{\leq} \varepsilon_x + \frac{2\varepsilon}{3} \overset{(3.9)}{\leq} \varepsilon. \quad \square
\end{aligned}$$

Second, we reformulate problem (3.7) as an equivalent problem

$$-\min_{y\in\mathbb{R}^{d_y}}\left\{\tilde{h}(y) + \max_{x\in\mathbb{R}^{d_x}}\left\{-G(x,y)-\tilde{f}(x)\right\}\right\}. \tag{3.12}$$

Our goal is to show that if $\hat{y}$ is an (ε_y,σ_y)-solution to the outer problem in (3.12) and $\hat{x}$ is an (ε_x,σ_x)-solution to the inner problem $\max_{x\in\mathbb{R}^{d_x}}\{-G(x,\hat{y})-\tilde{f}(x)\}$, then $\hat{x}$ is an $(\varepsilon'_x,\sigma'_x)$-solution to the outer problem in (3.7) and $\hat{y}$ is an $\left(\varepsilon'_y,\sigma'_y\right)$-solution to the inner problem $\max_{y\in\mathbb{R}^{d_y}}\{G(\hat{x},y)-\tilde{h}(y)\}$ in (3.7), where the dependencies $\varepsilon_x\left(\varepsilon_x\right)$, $\varepsilon_y\left(\varepsilon'_x,\varepsilon'_y\right)$, $\sigma_x\left(\sigma'_x\right)$, $\sigma_y\left(\sigma'_x,\sigma'_y\right)$ are polynomial. Let R_x, R_y be numbers such that $R_x^2 \geq \varepsilon'_x > 0$, $R_y^2 \geq \varepsilon'_y > 0$ and for the solution $z^* = (x^*,y^*)$ of (3.12) it holds that $\|x^*\| \leq R_x$ and

$\|y^*\| \leq R_y$. Since $\tilde{f}$, $\tilde{h}$ are convex, by Theorem 3.1.8 in [29], $\tilde{f}(x)$ is Lipschitz-continuous with constant $M_{\tilde{f}}$ on the ball $\mathcal{B}_2(0, 2R_x)$ and $\tilde{h}(y)$ is Lipschitz-continuous with constant $M_{\tilde{h}}$ on the ball $\mathcal{B}_2(0, 2R_y)$.

Lemma 4. *Assume, that in (3.12) $\tilde{f}(x)$ is $\tilde{\mu}_x$-strongly-convex, $\tilde{h}(y)$ is $\tilde{\mu}_y$-strongly-convex, and Assumption 1.2 holds. Let an approximate solution $(\hat{x}, \hat{y})$ for (3.12) satisfy*

1. *$\hat{y}$ is an $(\varepsilon_y, \sigma_y)$-solution to the outer problem in (3.12), i.e., (1.5) holds for the objective $\tilde{h}(y) + \max_{x \in \mathbb{R}^{d_x}} \{-G(x, y) - \tilde{f}(x)\}$;*
2. *$\hat{x}$ is an $(\varepsilon_x, \sigma_x)$-solution to the inner problem $\max_{x \in \mathbb{R}^{d_x}} \{-\tilde{f}(x) - G(x, \hat{y})\}$,*

where

$$\varepsilon_x \leq \min\left\{\frac{\tilde{\mu}_x \varepsilon'_x}{8}, \frac{\varepsilon'^2_x \tilde{\mu}_x}{32 M^2_{\tilde{f}}}, \frac{\varepsilon'_x \tilde{\mu}_x}{8\left(L + \frac{L^2}{\tilde{\mu}_y}\right)}\right\} = \boldsymbol{poly}(\varepsilon'_x), \tag{3.13}$$

$$\varepsilon_y \leq \min\left\{\frac{\tilde{\mu}_y \varepsilon'_y}{2}, \frac{\varepsilon_x \tilde{\mu}_x \tilde{\mu}_y}{L^2}, \frac{\varepsilon'^2_y \tilde{\mu}_y}{8 M^2_{\tilde{h}}}, \frac{\varepsilon'_y \tilde{\mu}_y}{2L}\right\} = \boldsymbol{poly}(\varepsilon'_x, \varepsilon'_y), \tag{3.14}$$

$$\sigma_x \leq \frac{\sigma'_x}{2} = \boldsymbol{poly}(\sigma'_x), \quad \sigma_y \leq \min\left\{\frac{\sigma'_x}{2}, \sigma'_y\right\} = \boldsymbol{poly}(\sigma'_x, \sigma'_y) \tag{3.15}$$

for some $\varepsilon > 0$, $\sigma \in (0, 1)$. Then, $\hat{x}$ is an $(\varepsilon'_x, \sigma'_x)$-solution to the outer problem in (3.7) and $\hat{y}$ is an $(\varepsilon'_y, \sigma'_y)$-solution to the inner problem in (3.7) $\max_{y \in \mathbb{R}^{d_y}} \{G(\hat{x}, y) - \tilde{h}(y)\}$.

Proof. We first prove that $\hat{x}$ is an $(\varepsilon'_x, \sigma'_x)$-solution to the outer problem in (3.7). Under the lemma assumptions, problems (3.7) and (3.12) have the unique solution (x^*, y^*). We denote $\Phi(y) = \tilde{h}(y) + \max_{x \in \mathbb{R}^{d_x}} \{-G(x, y) - \tilde{f}(x)\}$ and notice that $\Phi(y)$ is $\tilde{\mu}_y$-strongly-convex and has the minimum at y^*. Thus, by condition 1. of the lemma, with probability at least $1 - \sigma_y$,

$$\|\hat{y} - y^*\|^2 \leq \frac{2}{\tilde{\mu}_y}\left(\Phi(\hat{y}) - \Phi(y^*)\right) \leq \frac{2\varepsilon_y}{\tilde{\mu}_y}. \tag{3.16}$$

We denote $x^*(y) = \arg\max_{x \in \mathbb{R}^{d_x}} \{-G(x, y) - \tilde{f}(x)\}$ and notice that, since L-average-smoothness of $G(x, y)$ implies its L-smoothness, by Lemma 2, $x^*(y)$ is $\left(\frac{L}{\mu_x}\right)$-Lipschitz-continuous. Since $-G(x, \hat{y}) - \tilde{f}(x)$ is $\tilde{\mu}_x$-strongly-concave, using condition 2. of the lemma, we obtain that with probability at least $1 - \sigma_x - \sigma_y$,

$$\begin{aligned}\|\hat{x} - x^*\|^2 &\leq 2\|\hat{x} - x^*(\hat{y})\|^2 + 2\|x^*(\hat{y}) - x^*(y^*)\|^2 \leq \frac{4\varepsilon_x}{\tilde{\mu}_x} + 2\left(\frac{L}{\tilde{\mu}_x}\right)^2 \|\hat{y} - y^*\|^2 \\ &\leq \frac{4\varepsilon_x}{\tilde{\mu}_x} + 4\left(\frac{L}{\tilde{\mu}_x}\right)^2 \frac{\varepsilon_y}{\tilde{\mu}_y} \overset{(3.14)}{\leq} \frac{8\varepsilon_x}{\tilde{\mu}_x} \overset{(3.13)}{\leq} \varepsilon'_x.\end{aligned} \tag{3.17}$$

Further, with probability at least $1-\sigma_x-\sigma_y$, since $R_x^2 \geq \varepsilon_x' > 0$, we have $\|\hat{x}\|^2 \leq 2(\|\hat{x}-x^*\|^2+\|x^*\|^2) \leq 4R_x^2$ and, hence, $x^*, \hat{x} \in \mathcal{B}_2(0, 2R_x)$. Thus, since $\tilde{f}(x)$ is Lipschitz-continuous with constant $M_{\tilde{f}}$ on $\mathcal{B}_2(0, 2R_x)$ (see the justification before the statement of this lemma), we have with probability at least $1-\sigma_x-\sigma_y$ that $\left|\tilde{f}(\hat{x})-\tilde{f}(x^*)\right| \leq M_{\tilde{f}}\|\hat{x}-x^*\|$. Further, since L-average-smoothness of $G(x,y)$ implies its L-smoothness, by Lemma 2, $g(x)=\max_{y\in\mathbb{R}^{d_y}}\{G(x,y)-\tilde{h}(y)\}$ is $\left(L+\frac{L^2}{\tilde{\mu}_y}\right)$-smooth. Thus, with probability at least $1-\sigma_x-\sigma_y \overset{(3.15)}{\geq} 1-\sigma_x'$:

$$
\begin{aligned}
&\tilde{f}(\hat{x})+\max_{y\in\mathbb{R}^{d_y}}\{G(\hat{x},y)-\tilde{h}(y)\}-\min_{x\in\mathbb{R}^{d_x}}\{\tilde{f}(x)+\max_{y\in\mathbb{R}^{d_y}}\{G(x,y)-\tilde{h}(y)\}\}\\
&=\tilde{f}(\hat{x})-\tilde{f}(x^*)+g(\hat{x})-g(x^*)\\
&\leq M_{\tilde{f}}\|\hat{x}-x^*\|+\frac{1}{2}\left(L+\frac{L^2}{\tilde{\mu}_y}\right)\|\hat{x}-x^*\|^2 \overset{(3.17)}{\leq} M_{\tilde{f}}\sqrt{\frac{8\varepsilon_x}{\tilde{\mu}_x}}+\frac{1}{2}\left(L+\frac{L^2}{\tilde{\mu}_y}\right)\left(\frac{8\varepsilon_x}{\tilde{\mu}_x}\right)\\
&\overset{(3.13)}{\leq} M_{\tilde{f}}\sqrt{\frac{8\varepsilon_x'^2\tilde{\mu}_x}{\tilde{\mu}_x 32 M_{\tilde{f}}^2}}+\frac{1}{2}\left(L+\frac{L^2}{\tilde{\mu}_y}\right)\left(\frac{8\varepsilon_x'\tilde{\mu}_x}{\tilde{\mu}_x 8\left(L+\frac{L^2}{\tilde{\mu}_y}\right)}\right)=\varepsilon_x',
\end{aligned}
$$

justifying that $\hat{x}$ is an $(\varepsilon_x', \sigma_x')$-solution to the outer problem in (3.7).

Next, we prove that $\hat{y}$ is an $(\varepsilon_y', \sigma_y')$-solution to the inner problem in (3.7), i.e., $\max_{y\in\mathbb{R}^{d_y}}\{G(\hat{x},y)-\tilde{h}(y)\}$. With probability at least $1-\sigma_y$, from (3.16) and (3.14), we have that $\|\hat{y}-y^*\|^2 \leq \frac{2\varepsilon_y}{\tilde{\mu}_y} \leq \varepsilon_y' \leq R_y^2$, which implies $\|\hat{y}\|^2 \leq 2(\|\hat{y}-y^*\|^2+\|y^*\|^2) \leq 4R_y^2$ and, hence, $y^*, \hat{y} \in \mathcal{B}_2(0, 2R_y)$. Thus, since $\tilde{h}(y)$ is Lipschitz-continuous with constant $M_{\tilde{h}}$ on $\mathcal{B}_2(0, 2R_y)$ (see the justification before the statement of this lemma), we have with probability at least $1-\sigma_y$ that $\left|\tilde{h}(\hat{y})-\tilde{h}(y^*)\right| \leq M_{\tilde{h}}\|\hat{y}-y^*\|$. Further, L-average-smoothness of $G(x,y)$ implies its L-smoothness.

Thus, with probability at least $1-\sigma_y \overset{(3.15)}{\geq} 1-\sigma_y'$:

$$
\begin{aligned}
\left(G(\hat{x},y^*)-\tilde{h}(y^*)\right)-\left(G(\hat{x},\hat{y})-\tilde{h}(\hat{y})\right)&=\tilde{h}(\hat{y})-h(y^*)+G(\hat{x},y^*)-G(\hat{x},\hat{y})\\
&\leq M_{\tilde{h}}\|\hat{y}-y^*\|+\frac{L}{2}\|\hat{y}-y^*\|^2 \overset{(3.16)}{\leq} M_{\tilde{h}}\sqrt{\frac{2\varepsilon_y}{\tilde{\mu}_y}}+\frac{L}{2}\frac{2\varepsilon_y}{\tilde{\mu}_y}\\
&\overset{(3.14)}{\leq} M_{\tilde{h}}\sqrt{\frac{2\varepsilon_y'^2\tilde{\mu}_y}{\tilde{\mu}_y 8 M_{\tilde{h}}^2}}+\frac{L}{2}\frac{\varepsilon_y'\tilde{\mu}_y}{L\tilde{\mu}_y}=\varepsilon_y'.
\end{aligned}
$$

Hence, $\hat{y}$ is an $(\varepsilon_y', \sigma_y')$-solution to the inner problem in (3.7), i.e., $\max_{y\in\mathbb{R}^{d_y}}\{G(\hat{x},y)-\tilde{h}(y)\}$. □

Finally, we reformulate problem (3.12) as an equivalent strongly-convex-strongly-concave saddle-point problem (3.6). Our goal is to show that if $(\hat{x}, \hat{y})$ is an (ε, σ)-solution to problem (3.6), then $\hat{y}$ is an $(\varepsilon_y, \sigma_y)$-solution to the outer problem in (3.12) and $\hat{x}$ is

an $(\varepsilon_x, \sigma_x)$-solution to the inner problem $\max_{x\in\mathbb{R}^{d_x}}\{-G(x,\hat{y}) - \tilde{f}(x)\}$ in (3.12), where the dependencies $\varepsilon(\varepsilon_x, \varepsilon_y)$, $\sigma(\sigma_x, \sigma_y)$ are polynomial. Let R_x be a number such that $R_x^2 \geq \varepsilon_x > 0$ and for the solution $z^* = (x^*, y^*)$ of (3.6) it holds that $\|x^*\| \leq R_x$. Since $\tilde{f}$ is convex, by Theorem 3.1.8 in [29], $\tilde{f}$ is Lipschitz-continuous with constant $M_{\tilde{f}}$ on the ball $\mathcal{B}_2(0, 2R_x)$.

Lemma 5. *Assume, that in (3.6) $\tilde{f}(x)$ is $\tilde{\mu}_x$-strongly-convex, $\tilde{h}(y)$ is $\tilde{\mu}_y$-strongly-convex, and Assumption 1.2 holds. Let $(\hat{x}, \hat{y})$ be an (ε, σ)-solution to the saddle-point problem (3.6), i.e., (1.6) holds, where*

$$\varepsilon \leq \min\left\{\varepsilon_y, \frac{\varepsilon_x\tilde{\mu}_x}{2}, \frac{\varepsilon_x\tilde{\mu}_x}{2L}, \frac{\varepsilon_x^2\tilde{\mu}_x}{8M_{\tilde{f}}^2}\right\} = \boldsymbol{poly}(\varepsilon_x, \varepsilon_y), \sigma \leq \min\{\sigma_x, \sigma_y\} = \boldsymbol{poly}(\sigma_x, \sigma_y) \tag{3.18}$$

for some $\varepsilon_x, \varepsilon_y > 0$, $\sigma_x, \sigma_y \in (0,1)$. Then,

1. *$\hat{y}$ is an $(\varepsilon_y, \sigma_y)$-solution to the outer problem in (3.12), i.e., (1.5) holds for the objective $\tilde{h}(y) + \max_{x\in\mathbb{R}^{d_x}}\{-G(x,y) - \tilde{f}(x)\}$;*
2. *$\hat{x}$ is an $(\varepsilon_x, \sigma_x)$-solution to the inner problem $\max_{x\in\mathbb{R}^{d_x}}\{-G(x,\hat{y}) - \tilde{f}(x)\}$.*

Proof. Since $(\hat{x}, \hat{y})$ is an (ε, σ)-solution to the saddle-point problem (3.6), i.e., (1.6) holds, we have with probability at least $1 - \sigma$

$$\begin{aligned}
&\tilde{h}(\hat{y}) + \max_{x\in\mathbb{R}^{d_x}}\{-G(x,\hat{y}) - \tilde{f}(x)\} - \min_{y\in\mathbb{R}^{d_y}}\{\tilde{h}(y) + \max_{x\in\mathbb{R}^{d_x}}\{-G(x,y) - \tilde{f}(x)\}\} \\
&\quad\leq -\min_{x\in\mathbb{R}^{d_x}}\{\tilde{f}(x) + G(x,\hat{y}) - \tilde{h}(\hat{y})\} + \{\tilde{f}(x^*) + G(x^*,y^*) - \tilde{h}(y^*)\} \\
&\qquad + \max_{y\in\mathbb{R}^{d_y}}\{\tilde{f}(\hat{x}) + G(\hat{x},y) - \tilde{h}(y)\} - \{\tilde{f}(x^*) + G(x^*,y^*) - \tilde{h}(y^*)\} \\
&\quad= \max_{y\in\mathbb{R}^{d_y}}\{\tilde{f}(\hat{x}) + G(\hat{x},y) - \tilde{h}(y)\} - \min_{x\in\mathbb{R}^{d_x}}\{\tilde{f}(x) + G(x,\hat{y}) - \tilde{h}(\hat{y})\} \leq \varepsilon.
\end{aligned} \tag{3.19}$$

Since $\varepsilon \leq \varepsilon_y, \sigma \leq \sigma_y$, this immediately gives that $\hat{y}$ is an $(\varepsilon_y, \sigma_y)$-solution to the outer problem in (3.12).

Repeating similar steps as in (3.19), we obtain that $\hat{x}$ is an (ε, σ)-solution to the problem $\min_{x\in\mathbb{R}^{d_x}}\{\Psi(x) = \tilde{f}(x) + \max_{y\in\mathbb{R}^{d_y}}\{G(x,y) - \tilde{h}(y)\}\}$. $\Psi(x)$ is $\tilde{\mu}_x$-strongly-convex and has the minimum at x^*, whence, with probability at least $1 - \sigma$

$$\|\hat{x} - x^*\|^2 \leq \frac{2}{\tilde{\mu}_x}(\Psi(\hat{x}) - \Psi(x^*)) \leq \frac{2\varepsilon}{\tilde{\mu}_x} \overset{(3.18)}{\leq} \varepsilon_x. \tag{3.20}$$

Further, with probability at least $1 - \sigma$, since $R_x^2 \geq \varepsilon_x > 0$, we have $\|\hat{x}\|^2 \leq 2(\|\hat{x} - x^*\|^2 + \|x^*\|^2) \leq 4R_x^2$ and, hence, $x^*, \hat{x} \in \mathcal{B}_2(0, 2R_x)$. Thus, since $\tilde{f}(x)$ is Lipschitz-continuous with constant $M_{\tilde{f}}$ on $\mathcal{B}_2(0, 2R_x)$ (see the justification before the statement of this lemma), we have with probability at least $1 - \sigma$ that $\left|\tilde{f}(\hat{x}) - \tilde{f}(x^*)\right| \leq M_{\tilde{f}}\|\hat{x} - x^*\|$.

Thus, using that $G(x,y)$ is L-smooth by its L-average-smoothness, we get with probability at least $1-\sigma \overset{(3.18)}{\geq} 1-\sigma_x$

$$\begin{aligned}\left\{-G(x^*,\hat{y})-\tilde{f}(x^*)\right\}-\left\{-G(\hat{x},\hat{y})-\tilde{f}(\hat{x})\right\} &= \tilde{f}(\hat{x})-f(x^*)+G(\hat{x},\hat{y})-G(x^*,\hat{y})\\ &\leq M_{\tilde{f}}\|\hat{x}-x^*\|+\frac{L}{2}\|\hat{x}-x^*\|^2 \overset{(3.20)}{\leq} M_{\tilde{f}}\sqrt{\frac{2\varepsilon}{\tilde{\mu}_x}}+\frac{L}{2}\frac{2\varepsilon}{\tilde{\mu}_x}\\ &\overset{(3.18)}{\leq} M_{\tilde{f}}\sqrt{\frac{2\varepsilon_x^2\tilde{\mu}_x}{8\tilde{\mu}_x M_{\tilde{f}}^2}}+\frac{L}{2}\frac{\varepsilon_x\tilde{\mu}_x}{L\tilde{\mu}_x}=\varepsilon_x,\end{aligned}$$

which implies that $\hat{x}$ is an (ε_x,σ_x)-solution to the inner problem in (3.12), i.e., $\max_{x\in\mathbb{R}^{d_x}}\{-G(x,\hat{y})-\tilde{f}(x)\}$. □

4. Accelerated variance-reduced method for saddle-point problems

In this section, we describe in detail the structure of our three-loop algorithm, listed below as Algorithm 3, and prove the main complexity Theorem 4. The main idea is to accelerate SAGA algorithm (Algorithm 2) using extrapolation steps in x and in y as in the Catalyst algorithm (Algorithm 1). Therefore, in the first two loops, we apply Algorithm 1 and, in the third loop, we apply Algorithm 2. Each loop takes as input some outer problem and its required accuracy ε and confidence level σ. This outer problem is reformulated in the loop and the reformulation is solved either by Algorithm 1 or by Algorithm 2 with the accuracy ε' and confidence level σ' guaranteeing, by the results of Section 3, that for the outer problem we have an (ε,σ)-solution. Moreover, the algorithm used in each loop requires to solve some auxiliary problem with some accuracy and confidence level, which are then used as input for the next loop. After the loops' description, we summarize the complexity of each loop to obtain the total complexity of the algorithm in Theorem 4.

Loop 1 This loop starts at line 3 and ends at line 12. The goal of this loop is to accelerate SAGA Algorithm 2 in x using the Catalyst Algorithm 1. To that end, we reformulate the saddle-point problem (1.7) as the minimization problem

$$\min_{x\in\mathbb{R}^{d_x}}\left\{f(x)+\max_{y\in\mathbb{R}^{d_y}}\{G(x,y)-h(y)\}\right\}. \tag{4.1}$$

Let $\left(\varepsilon_x^{(1)},\sigma_x^{(1)}\right)=(\mathbf{poly}(\varepsilon),\mathbf{poly}(\sigma))$, $\left(\varepsilon_y^{(1)},\sigma_y^{(1)}\right)=(\mathbf{poly}(\varepsilon),\mathbf{poly}(\sigma))$ satisfy (3.8) and (3.9) with f and h playing the role of $\tilde{f}$ and $\tilde{h}$ respectively, meaning that $\tilde{\mu}_x=\mu_x$, $\tilde{\mu}_y=\mu_y$. If we find $\hat{x}$ that is an $\left(\varepsilon_x^{(1)},\sigma_x^{(1)}\right)$-solution to the outer problem (4.1), and $\hat{y}$ that is an $\left(\varepsilon_y^{(1)},\sigma_y^{(1)}\right)$-solution to the inner problem

$$\max_{y\in\mathbb{R}^{d_y}}\{G(\hat{x},y)-h(y)\}, \tag{4.2}$$

then, by Lemma 3, the pair $(\hat{x},\hat{y})$ is an (ε,σ)-solution to problem (1.7).

Algorithm 3 Accelerated variance-reduced algorithm for SPPs.

Input: Starting point $x_0\in\mathbb{R}^{d_x}$, algorithmic parameters $H_1>0$, $H_2>0$, $\alpha_{x_0}>0$, $\alpha_{y_0}>0$, strong convexity and strong concavity parameters μ_x, μ_y, and sequences of accuracies and confidence levels $\left(\varepsilon^{(3)}_{k,n},\sigma^{(3)}_{k,n}\right)_{k,n\geq 0}$.

Initialize $x_0^{md}=x_0$, $q_x=\frac{\mu_x}{\mu_x+H_1}$;

for $k=0,1,\ldots$ **do**

  $y_{k_0}^{md}=y_{k_0}$, $q_y=\frac{\mu_y}{\mu_y+H_2}$;

  for $n=0,1,\ldots$ **do**

    By the SAGA Algorithm 2 find an $(\varepsilon^{(3)}_{k,n},\sigma^{(3)}_{k,n})$-solution

$$(x_k,y_{k,n})\approx\min_{x\in\mathbb{R}^{d_x}}\max_{y\in\mathbb{R}^{d_y}}S_{k,n}(x,y),\quad\text{where}$$

$$S_{k,n}(x,y):=f(x)+\tfrac{H_1}{2}\|x-x_{k-1}^{md}\|^2+G(x,y)-h(y)-\tfrac{H_2}{2}\|y-y_{k,n-1}^{md}\|^2;$$

    Find $\alpha_{y_{k,n}}\in(0,1)$ from equation $\alpha^2_{y_{k,n}}=(1-\alpha_{y_{k,n}})\alpha^2_{y_{k,n-1}}+q_y\alpha_{y_{k,n}}$;

    Compute $y_{k,n}^{md}$ using Nesterov's extrapolation step

$$y_{k,n}^{md}=y_{k,n}+\beta_{y_{k,n}}(y_{k,n}-y_{k,n-1})\quad\text{with}\quad\beta_{y_{k,n}}=\frac{\alpha_{y_{k,n-1}}(1-\alpha_{y_{k,n-1}})}{\alpha^2_{y_{k,n-1}}+\alpha_{y_{k,n}}}.$$

  end for

  Find $\alpha_{x_k}\in(0,1)$ from equation $\alpha^2_{x_k}=(1-\alpha_{x_k})\alpha^2_{x_{k-1}}+q_x\alpha_{x_k}$;

  Compute x_k^{md} using Nesterov's extrapolation step

$$x_k^{md}=x_k+\beta_{x_k}(x_k-x_{k-1})\quad\text{with}\quad\beta_{x_k}=\frac{\alpha_{x_{k-1}}(1-\alpha_{x_{k-1}})}{\alpha^2_{x_{k-1}}+\alpha_{x_k}}.$$

end for

Output: x_k, $y_{k,n}$ (final estimate).

To find such pair $(\hat{x},\hat{y})$ with target parameters $\varepsilon_x^{(1)},\sigma_x^{(1)},\varepsilon_y^{(1)},\sigma_y^{(1)}$, we solve problem (4.1) using Algorithm 1 with parameter H_1 to be chosen later. The outer objective in (4.1) satisfies Assumption 2 with $\varphi(x)=f(x)$ which is μ_x-strongly-convex and $\psi(x)=\max_{y\in\mathbb{R}^{d_y}}\{G(x,y)-h(y)\}$ which is $\left(L+\frac{L^2}{\mu_y}\right)$-smooth. The latter holds by Lemma 2 since L-average smoothness of $G(x,y)$ implies that $G(x,y)$ is L-smooth as a function of (x,y).

In each iteration k of Algorithm 1 applied to (4.1), the problem

$$x_k=\arg\min_{x\in\mathbb{R}^{d_x}}\left\{f(x)+\max_{y\in\mathbb{R}^{d_y}}\{G(x,y)-h(y)\}+\frac{H_1}{2}\|x-x_{k-1}^{md}\|^2\right\} \tag{4.3}$$

needs to be solved inexactly. Assume that, for each $k\geq 0$, we can find an $\left(\varepsilon^{(1)}_{x_k},\sigma^{(1)}_{x_k}\right)$-solution to this problem, where $\varepsilon^{(1)}_{x_k}$ and $\sigma^{(1)}_{x_k}$ satisfy respectively (2.4) and (2.5) with

$\varepsilon_x^{(1)}$, $\sigma_x^{(1)}$, μ_x, H_1 playing the role of ε, σ, μ, H respectively. The latter also implies that $\varepsilon_{x_k}^{(1)} = \mathbf{poly}\left(\varepsilon_x^{(1)}\right) = \mathbf{poly}\left(\varepsilon\right)$ and $\sigma_{x_k}^{(1)} = \mathbf{poly}\left(\varepsilon_x^{(1)}, \sigma_x^{(1)}\right) = \mathbf{poly}\left(\varepsilon, \sigma\right)$. Applying Theorem 2, and using that $\varepsilon_x^{(1)} = \mathbf{poly}(\varepsilon)$, $\sigma_x^{(1)} = \mathbf{poly}(\sigma)$, we obtain that in $\mathcal{N}_1 = \widetilde{O}\left(\sqrt{1 + \frac{H_1}{\mu_x}}\right)$ iterations Algorithm 1 finds an $(\varepsilon_x^{(1)}, \sigma_x^{(1)})$-solution $\hat{x}$ to problem (4.1) and an $(\varepsilon_y^{(1)}, \sigma_y^{(1)})$-solution $\hat{y}$ to problem (4.2), which, by the choice of $\varepsilon_x^{(1)}, \sigma_x^{(1)}, \varepsilon_y^{(1)}, \sigma_y^{(1)}$, are also an (ε, σ)-solution to problem (1.7).

It remains to show how to find $\left(\varepsilon_{x_k}^{(1)}, \sigma_{x_k}^{(1)}\right)$-solution to problem (4.3) in each iteration of Algorithm 1. This is organized in Loop 2 below. Importantly, $\left(\varepsilon_{x_k}^{(1)}, \sigma_{x_k}^{(1)}\right) = (\mathbf{poly}(\varepsilon), \mathbf{poly}(\varepsilon, \sigma))$.

Loop 2 This loop starts at line 5 and ends at line 9. The goal of this loop is to accelerate SAGA Algorithm 2 in y using the Catalyst Algorithm 1 while finding an $\left(\varepsilon_{x_k}^{(1)}, \sigma_{x_k}^{(1)}\right)$-solution to problem (4.3). To that end, we reformulate the minimization problem in x (4.3) as the minimization problem in y:

$$\min_{y \in \mathbb{R}^{d_y}} \left\{ h(y) + \max_{x \in \mathbb{R}^{d_x}} \left\{ -\hat{f}_k(x) - G(x, y) \right\} \right\}, \tag{4.4}$$

where $\hat{f}_k(x) = f(x) + \frac{H_1}{2}\|x - x_{k-1}^{md}\|^2$ is $(\mu_x + H_1)$-strongly-convex and prox-friendly [35] since f is prox-friendly.

Set $\left(\varepsilon_{y_k}^{(1)}, \sigma_{y_k}^{(1)}\right) = \left(\varepsilon_{x_k}^{(1)}, \sigma_{x_k}^{(1)}\right)$ and let $\left(\varepsilon_{x_k}^{(2)}, \sigma_{x_k}^{(2)}\right) = \left(\mathbf{poly}\left(\varepsilon_{x_k}^{(1)}\right), \mathbf{poly}\left(\sigma_{x_k}^{(1)}\right)\right) = (\mathbf{poly}(\varepsilon), \mathbf{poly}(\varepsilon, \sigma))$, $\left(\varepsilon_{y_k}^{(2)}, \sigma_{y_k}^{(2)}\right) = \left(\mathbf{poly}\left(\varepsilon_{y_k}^{(1)}, \varepsilon_{x_k}^{(1)}\right), \mathbf{poly}\left(\sigma_{x_k}^{(1)}, \sigma_{y_k}^{(1)}\right)\right) = \left(\mathbf{poly}(\varepsilon), \mathbf{poly}(\varepsilon, \sigma)\right)$ satisfy (3.13), (3.14), and (3.15) with $\varepsilon_{x_k}^{(1)}$, $\sigma_{x_k}^{(1)}$, $\varepsilon_{y_k}^{(1)}$, $\sigma_{y_k}^{(1)}$ playing the role of ε_x', σ_x', ε_y', σ_y' respectively, and $\hat{f}_k$ and h playing the role of $\tilde{f}$ and $\tilde{h}$ respectively, meaning that $\tilde{\mu}_x = \mu_x + H_1$, $\tilde{\mu}_y = \mu_y$. If we find $\hat{y}_k$ that is an $\left(\varepsilon_{y_k}^{(2)}, \sigma_{y_k}^{(2)}\right)$-solution to the outer problem (4.4) and $\hat{x}_k$ that is an $\left(\varepsilon_{x_k}^{(2)}, \sigma_{x_k}^{(2)}\right)$-solution to the inner problem

$$\max_{x \in \mathbb{R}^{d_x}} \left\{ -\hat{f}_k(x) - G(x, \hat{y}_k) \right\}, \tag{4.5}$$

then, by Lemma 4, $\hat{x}_k$ is an $\left(\varepsilon_{x_k}^{(1)}, \sigma_{x_k}^{(1)}\right)$-solution to the outer problem in (4.3) and $\hat{y}_k$ is an $\left(\varepsilon_{y_k}^{(1)}, \sigma_{y_k}^{(1)}\right)$-solution to the inner problem in (4.3) and the goal of Loop 2 is achieved.

To find such pair $(\hat{x}_k, \hat{y}_k)$ with target parameters $\varepsilon_{x_k}^{(2)}, \sigma_{x_k}^{(2)}, \varepsilon_{y_k}^{(2)}, \sigma_{y_k}^{(2)}$, we solve problem (4.4) using Algorithm 1 with parameter H_2 to be chosen later. The outer objective in (4.4) satisfies Assumption 2 with μ_y-strongly-convex function $h(y)$ playing the role of $\varphi(\cdot)$ and $\left(L + \frac{L^2}{\mu_x + H_1}\right)$-smooth function $\max_{x \in \mathbb{R}^{d_x}} \left\{ -\hat{f}_k(x) - G(x, y) \right\}$ playing the role of $\psi(\cdot)$. The smoothness holds by Lemma 2 since L-average smoothness of $G(x, y)$ implies that $G(x, y)$ is L-smooth as a function of (x, y) and since $\hat{f}_k(x)$ is $(\mu_x + H_1)$-strongly-convex.

In each iteration n of Algorithm 1 applied to (4.4), the problem

$$y_{k,n} = \arg\min_{y\in\mathbb{R}^{d_y}} \left\{ h(y) + \max_{x\in\mathbb{R}^{d_x}} \left\{ -G(x,y) - \hat{f}_k(x) \right\} + \frac{H_2}{2}\|y - y_{k,n-1}^{md}\|^2 \right\} \quad (4.6)$$

needs to be solved inexactly. Assume that, for each $n \geq 0$, we can find an $\left(\varepsilon_{y_{k,n}}^{(2)}, \sigma_{y_{k,n}}^{(2)}\right)$-solution to this problem, where $\varepsilon_{y_{k,n}}^{(2)}$ and $\sigma_{y_{k,n}}^{(2)}$ satisfy respectively (2.4) and (2.5) with $\varepsilon_{y_k}^{(2)}$, $\sigma_{y_k}^{(2)}$, μ_y, H_2 playing the role of ε, σ, μ, H respectively. The latter also implies that $\varepsilon_{y_{k,n}}^{(2)} = \mathbf{poly}\left(\varepsilon_{y_k}^{(2)}\right) = \mathbf{poly}\left(\varepsilon\right)$ and $\sigma_{y_{k,n}}^{(2)} = \mathbf{poly}\left(\varepsilon_{y_k}^{(2)}, \sigma_{y_k}^{(2)}\right) = \mathbf{poly}\left(\varepsilon, \sigma\right)$. Applying Theorem 2, and using that $\varepsilon_{y_k}^{(2)} = \mathbf{poly}(\varepsilon)$, $\sigma_{y_k}^{(2)} = \mathbf{poly}(\varepsilon, \sigma)$, we obtain that in $\mathcal{N}_2 = \widetilde{O}\left(\sqrt{1 + \frac{H_2}{\mu_y}}\right)$ iterations Algorithm 1 finds an $(\varepsilon_{y_k}^{(2)}, \sigma_{y_k}^{(2)})$-solution $\hat{y}_k$ to problem (4.4) and an $(\varepsilon_{x_k}^{(2)}, \sigma_{x_k}^{(2)})$-solution $\hat{x}_k$ to problem (4.5), and, by the choice of $\varepsilon_{x_k}^{(2)}, \sigma_{x_k}^{(2)}, \varepsilon_{y_k}^{(2)}, \sigma_{y_k}^{(2)}$, $\hat{x}_k$ is also an $\left(\varepsilon_{x_k}^{(1)}, \sigma_{x_k}^{(1)}\right)$-solution to problem (4.3).

It remains to show how to find $\left(\varepsilon_{y_{k,n}}^{(2)}, \sigma_{y_{k,n}}^{(2)}\right)$-solution to problem (4.6) in each iteration of Algorithm 1. This is organized in Loop 3 below. Importantly, $\left(\varepsilon_{y_{k,n}}^{(2)}, \sigma_{y_{k,n}}^{(2)}\right) = (\mathbf{poly}(\varepsilon), \mathbf{poly}(\varepsilon, \sigma))$.

Loop 3 This loop is made inside step 6, where Algorithm 2 is applied. The goal of this loop is to find an $\left(\varepsilon_{y_{k,n}}^{(2)}, \sigma_{y_{k,n}}^{(2)}\right)$-solution to problem (4.6).

To that end, we reformulate the minimization problem in y (4.6) as the saddle-point problem

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left\{ \hat{f}_k(x) + G(x,y) - \hat{h}_{k,n}(y) \right\}, \quad (4.7)$$

where $\hat{h}_{k,n}(y) = h(y) + \frac{H_2}{2}\|y - y_{k,n-1}^{md}\|^2$ is $(\mu_y + H_2)$-strongly-convex and prox-friendly [35] since h is prox-friendly.

Set $\left(\varepsilon_{x_{k,n}}^{(2)}, \sigma_{x_{k,n}}^{(2)}\right) = \left(\varepsilon_{y_{k,n}}^{(2)}, \sigma_{y_{k,n}}^{(2)}\right)$ and let $\varepsilon_{k,n}^{(3)} = \mathbf{poly}\left(\varepsilon_{x_{k,n}}^{(2)}, \varepsilon_{y_{k,n}}^{(2)}\right) = \mathbf{poly}(\varepsilon)$, $\sigma_{k,n}^{(3)} = \mathbf{poly}\left(\sigma_{x_{k,n}}^{(2)}, \sigma_{y_{k,n}}^{(2)}\right) = \mathbf{poly}(\varepsilon, \sigma)$ satisfy (3.18) with $\varepsilon_{x_{k,n}}^{(2)}$, $\sigma_{x_{k,n}}^{(2)}$, $\varepsilon_{y_{k,n}}^{(2)}$, $\sigma_{y_{k,n}}^{(2)}$ playing the role of ε_x, σ_x, ε_y, σ_y respectively, and $\hat{f}_k$ and $\hat{h}_{k,n}$ playing the role of $\tilde{f}$ and $\tilde{h}$ respectively, meaning that $\tilde{\mu}_x = \mu_x + H_1$, $\tilde{\mu}_y = \mu_y + H_2$. If we find a pair $(\hat{x}_{k,n}, \hat{y}_{k,n})$ that is an $(\varepsilon_{k,n}^{(3)}, \sigma_{k,n}^{(3)})$-solution to the saddle-point problem (4.7), then, by Lemma 5, $\hat{y}_{k,n}$ is an $\left(\varepsilon_{y_{k,n}}^{(2)}, \sigma_{y_{k,n}}^{(2)}\right)$-solution to the outer problem in (4.6) and $\hat{x}_{k,n}$ is an $\left(\varepsilon_{x}^{(2)}, \sigma_{x}^{(2)}\right)$-solution to the inner problem in (4.6).

To find such pair $(\hat{x}_{k,n}, \hat{y}_{k,n})$ with target parameters $\varepsilon_{k,n}^{(3)}, \sigma_{k,n}^{(3)}$, we solve problem (4.7) using Algorithm 2 with

$$M(x,y) = \hat{f}_k(x) - \hat{h}_{k,n}(y), \qquad K(x,y) = G(x,y).$$

This problem satisfies the assumptions of Lemma 1 with $\tilde{f}(x) = \hat{f}_k(x)$, $\tilde{h}(y) = \hat{h}_{k,n}(y)$, $\tilde{\mu}_x = \mu_x + H_1$, $\tilde{\mu}_y = \mu_y + H_2$ since $\hat{f}_k(x)$, $\hat{h}_{k,n}(y)$ are strongly-convex and prox-friendly, and since $G(x, y)$ is L-average smooth. Therefore, by Lemma 1, Assumption 3 holds, which allows us to apply Algorithm 2 to solve problem (4.7) using the following parameters: $\pi_i = \frac{1}{m}$, $i = 1, ..., m$, $\bar{L} = \frac{2L}{\min\{\mu_x+H_1, \mu_y+H_2\}}$, $s = 1$. We set $(\varepsilon_{k,n}^{(3)})'$ and $(\sigma_{k,n}^{(3)})'$ to satisfy (2.9) with $\varepsilon_{k,n}^{(3)}$, $\sigma_{k,n}^{(3)}$ playing the role of ε, σ respectively, and $\tilde{\mu}_x = \mu_x + H_1$, $\tilde{\mu}_y = \mu_y + H_2$, which implies also that $(\varepsilon_{k,n}^{(3)})' = \mathbf{poly}\left(\varepsilon_{k,n}^{(3)}\right) = \mathbf{poly}\left(\varepsilon\right)$ and $(\sigma_{k,n}^{(3)})' = \mathbf{poly}\left(\sigma_{k,n}^{(3)}\right) = \mathbf{poly}\left(\varepsilon, \sigma\right)$. Applying Theorem 3, and using that $(\varepsilon_{k,n}^{(3)})' = \mathbf{poly}(\varepsilon)$, $(\sigma_{k,n}^{(3)})' = \mathbf{poly}(\varepsilon, \sigma)$, we obtain that in $\mathcal{N}_3 = \widetilde{O}\left(m + \frac{L^2}{\min\{\mu_x+H_1, \mu_y+H_2\}}\right)$ iterations Algorithm 2 finds an $(\varepsilon_{k,n}^{(3)}, \sigma_{k,n}^{(3)})$-solution $(\hat{x}_{k,n}, \hat{y}_{k,n})$ to problem (4.7), and, by the choice of $\varepsilon_{k,n}^{(3)}, \sigma_{k,n}^{(3)}$, $\hat{y}_{k,n}$ is also an $\left(\varepsilon_{y_{k,n}}^{(2)}, \sigma_{y_{k,n}}^{(2)}\right)$-solution to problem (4.6).

The total complexity of these three loops is summarized in the next result, which is the main result of the paper.

Theorem 4. *Let Assumption 1 hold. Then, Algorithm 3 after*

$$O\left(\left(m + m^{\frac{3}{4}}\sqrt{\frac{L}{\mu_x}} + m^{\frac{3}{4}}\sqrt{\frac{L}{\mu_y}} + \frac{L\sqrt{m}}{\sqrt{\mu_x\mu_y}}\right)\ln^3\frac{1}{\varepsilon\sigma}\right) \tag{4.8}$$

evaluations of stochastic gradients $\nabla_x G_i(x, y)$, $\nabla_y G_i(x, y)$ and proximal operators of $f(x)$ and $h(y)$ (see (1.4)) finds an (ε, σ)-solution to problem (1.7).

Proof. We evaluate stochastic gradients $\nabla_x G_i(x, y)$, $\nabla_y G_i(x, y)$ and proximal operators for the functions $f(x)$, $h(y)$ (when we evaluate the proximal operators for $\hat{f}_k(x)$, $\hat{h}_{k,n}(y)$) only in Loop 3, where Algorithm 2 is used. Multiplying the number of iterations $\mathcal{N}_1$, $\mathcal{N}_2$, $\mathcal{N}_3$ in each loop, we obtain that the total number of these basic operations is

$$\mathcal{N} = \widetilde{O}\left(\sqrt{1 + \frac{H_1}{\mu_x}}\right) \cdot \widetilde{O}\left(\sqrt{1 + \frac{H_2}{\mu_y}}\right) \cdot \widetilde{O}\left(m + \frac{L^2}{\min(H_1 + \mu_x, H_2 + \mu_y)^2}\right).$$

Choosing $H_1 = \max\left\{\mu_x, \frac{L}{\sqrt{m}}\right\}, H_2 = \max\left\{\mu_y, \frac{L}{\sqrt{m}}\right\}$, we obtain

$$\begin{aligned}
\mathcal{N} &\leq \widetilde{O}\left(1 + \sqrt{\frac{L}{\mu_x\sqrt{m}}}\right) \cdot \widetilde{O}\left(1 + \sqrt{\frac{L}{\mu_y\sqrt{m}}}\right) \cdot \widetilde{O}\left(m + L^2\max\left(\frac{1}{H_1}, \frac{1}{H_2}\right)^2\right) \\
&\leq \widetilde{O}\left(1 + \sqrt{\frac{L}{\mu_x\sqrt{m}}}\right) \cdot \widetilde{O}\left(1 + \sqrt{\frac{L}{\mu_y\sqrt{m}}}\right) \cdot \widetilde{O}\left(m + L^2\frac{m}{L^2}\right) \\
&= \widetilde{O}\left(1 + \sqrt{\frac{L}{\mu_x\sqrt{m}}}\right) \cdot \widetilde{O}\left(1 + \sqrt{\frac{L}{\mu_y\sqrt{m}}}\right) \cdot \widetilde{O}\left(m\right)
\end{aligned}$$

$$= \widetilde{O}\left(m + m^{\frac{3}{4}}\sqrt{\frac{L}{\mu_x}} + m^{\frac{3}{4}}\sqrt{\frac{L}{\mu_y}} + \frac{L\sqrt{m}}{\sqrt{\mu_x\mu_y}}\right),$$

where in the second inequality we used that $\frac{1}{H_1}, \frac{1}{H_2} \leq \frac{\sqrt{m}}{L}$. Since each loop gives one logarithmic term, we have the third power of the logarithm in (4.8). □

5. Convex-strongly-concave and convex-concave cases

In this section, we consider problem (1.1) in the convex-strongly-concave and convex-concave settings under the following assumption.

Assumption 4.

1. $f(x)$ is convex ($\mu_x = 0$) and $h(y)$ is μ_y-strongly-convex with $\mu_y \geq 0$;
2. There exists a solution (x^*, y^*) to problem (1.1) such that $\|x^*\| \leq R_x$ and, if $\mu_y = 0$, $\|y^*\| \leq R_y$;
3. Assumption 1.2 and 1.3 hold.

The main idea is to use a reduction technique based on regularization. When lacking strong convexity/concavity w.r.t. one of the variables, we add a quadratic regularization for this variable, which reduces the problem to problem (1.7) under Assumption 1 and we can apply Algorithm 3. The following result shows that a solution to such-regularized problems gives is a solution to the original problem (1.1) when the regularization parameter is sufficiently small.

Lemma 6. *Under Assumption 4:*

if $\mu_x = 0$, $\mu_y > 0$ and $(\hat{x}, \hat{y})$ is an $\left(\frac{2\varepsilon}{3}, \sigma\right)$-solution to problem:

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left\{ f(x) + \frac{\varepsilon}{12R_x^2}\|x\|^2 + G(x,y) - h(y) \right\};$$

or if $\mu_x = 0$, $\mu_y = 0$ and $(\hat{x}, \hat{y})$ is an $\left(\frac{\varepsilon}{2}, \sigma\right)$-solution to the problem:

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left\{ f(x) + \frac{\varepsilon}{16R_x^2}\|x\|^2 + G(x,y) - h(y) - \frac{\varepsilon}{16R_y^2}\|y\|^2 \right\},$$

then, $(\hat{x}, \hat{y})$ is an (ε, σ)-solution to problem (1.1).

Proof. If $\mu_x = 0, \mu_y > 0$, then, with probability at least $1 - \sigma$

$$\frac{2\varepsilon}{3} \geq \max_{y\in\mathbb{R}^{d_y}} \left\{ f(\hat{x}) + \frac{\varepsilon}{12R_x^2}\|\hat{x}\|^2 + G(\hat{x}, y) - h(y) \right\}$$

$$
\begin{aligned}
&- \min_{x\in\mathbb{R}^{d_x}} \left\{ f(x) + \frac{\varepsilon}{12R_x^2}\|x\|^2 + G(x,\hat{y}) - h(\hat{y}) \right\} \\
&\geq \max_{\|y\|\leq 2R_y} \left\{ f(\hat{x}) + \frac{\varepsilon}{12R_x^2}\|\hat{x}\|^2 + G(\hat{x},y) - h(y) \right\} \\
&- \min_{\|x\|\leq 2R_x} \left\{ f(x) + \frac{\varepsilon}{12R_x^2}\|x\|^2 + G(x,\hat{y}) - h(\hat{y}) \right\} \\
&\geq \max_{\|y\|\leq 2R_y} \{ f(\hat{x}) + G(\hat{x},y) - h(y) \} + \frac{\varepsilon}{12R_x^2}\|\hat{x}\|^2 \\
&- \min_{\|x\|\leq 2R_x} \{ f(x) + G(x,\hat{y}) - h(\hat{y}) \} - \frac{\varepsilon}{12R_x^2}4R_x^2 \\
&\geq \max_{\|y\|\leq 2R_y} \{ f(\hat{x}) + G(\hat{x},y) - h(y) \} \\
&- \min_{\|x\|\leq 2R_x} \{ f(x) + G(x,\hat{y}) - h(\hat{y}) \} - \frac{\varepsilon}{3},
\end{aligned}
$$

and $(\hat{x},\hat{y})$ is an (ε,σ)-solution to the problem

$$
\min_{\|x\|\leq 2R_x} \max_{\|y\|\leq 2R_y} \{ f(x) + G(x,y) - h(y) \}.
$$

This problem is equivalent to problem (1.1) under assumption $\|x^*\| \leq R_x$.

If $\mu_x = 0, \mu_y = 0$, then with probability at least $1-\sigma$

$$
\begin{aligned}
\frac{\varepsilon}{2} &\geq \max_{y\in\mathbb{R}^{d_y}} \left\{ f(\hat{x}) + \frac{\varepsilon}{16R_x^2}\|\hat{x}\|^2 + G(\hat{x},y) - h(y) - \frac{\varepsilon}{16R_y^2}\|y\|^2 \right\} \\
&- \min_{x\in\mathbb{R}^{d_x}} \left\{ f(x) + \frac{\varepsilon}{16R_x^2}\|x\|^2 + G(x,\hat{y}) - h(\hat{y}) - \frac{\varepsilon}{16R_y^2}\|\hat{y}\|^2 \right\} \\
&\geq \max_{\|y\|\leq 2R_y} \left\{ f(\hat{x}) + \frac{\varepsilon}{16R_x^2}\|\hat{x}\|^2 + G(\hat{x},y) - h(y) - \frac{\varepsilon}{16R_y^2}\|y\|^2 \right\} \\
&- \min_{\|x\|\leq 2R_x} \left\{ f(x) + \frac{\varepsilon}{16R_x^2}\|x\|^2 + G(x,\hat{y}) - h(\hat{y}) - \frac{\varepsilon}{16R_y^2}\|\hat{y}\|^2 \right\} \\
&\geq \max_{\|y\|\leq 2R_y} \{ f(\hat{x}) + G(\hat{x},y) - h(y) \} + \frac{\varepsilon}{16R_x^2}\|\hat{x}\|^2 - \frac{\varepsilon}{16R_y^2}4R_y^2 \\
&- \min_{\|x\|\leq 2R_x} \{ f(x) + G(x,\hat{y}) - h(\hat{y}) \} - \frac{\varepsilon}{16R_x^2}4R_x^2 + \frac{\varepsilon}{16R_y^2}\|\hat{y}\|^2 \\
&\geq \max_{\|y\|\leq 2R_y} \{ f(\hat{x}) + G(\hat{x},y) - h(y) \} - \frac{\varepsilon}{4} \\
&- \min_{\|x\|\leq 2R_x} \{ f(x) + G(x,\hat{y}) - h(\hat{y}) \} - \frac{\varepsilon}{4},
\end{aligned}
$$

and $(\hat{x},\hat{y})$ is an (ε,σ)-solution to the problem

$$\min_{\|x\|\leq 2R_x} \max_{\|y\|\leq 2R_y} \{f(x) + G(x,y) - h(y)\}.$$

This problem is equivalent to problem (1.1) under assumption $\|x^*\| \leq R_x$, $\|y^*\| \leq R_y$. □

We are now in a position to apply Algorithm 3 to problems stated in the previous Lemma and obtain the corresponding complexity bounds for problem (1.1) as a corollary of Theorem 4.

Corollary 1. *Let Assumption 4 hold. Then, Algorithm 3 after (in convex-strongly-concave case)*

$$O\left(\left(m + m^{\frac{3}{4}} R_x \sqrt{\frac{L}{\varepsilon}} + m^{\frac{3}{4}} \sqrt{\frac{L}{\mu_y}} + \frac{R_x L \sqrt{m}}{\sqrt{\varepsilon \mu_y}}\right) \ln^3 \frac{1}{\varepsilon\sigma}\right),$$

or (in convex-concave case)

$$O\left(\left(m + (R_x + R_y) m^{\frac{3}{4}} \sqrt{\frac{L}{\varepsilon}} + \frac{R_x R_y L \sqrt{m}}{\varepsilon}\right) \ln^3 \frac{1}{\varepsilon\sigma}\right)$$

evaluations of stochastic gradients $\nabla_x G_i(x,y)$, $\nabla_y G_i(x,y)$ and proximal operators of $f(x)$ and $h(y)$ finds an (ε,σ)-solution to problem (1.1).

Proof. We regularize problem (1.1) as proposed in Lemma 6, obtain strongly-convex-strongly-concave problem with constants either $\mu_x = \frac{\varepsilon}{12R_x^2}$, μ_y or $\mu_x = \frac{\varepsilon}{16R_x^2}$, $\mu_y = \frac{\varepsilon}{16R_y^2}$, and solve it by Algorithm 3. The result follows by substituting these values of the parameters to the complexity bound in Theorem 4. □

6. Conclusions and future work

In this paper, we propose accelerated variance-reduced methods for saddle-point problems with finite-sum structure and composite terms. For these methods, we propose complexity estimates that up to constant and logarithmic factors coincide with the lower bounds [12] for this class of problems. Our algorithms are based on several nested loops and to make them more practical it is desired to propose their loop-less counterparts. The recent progress in this field [18] gives a hope of the possibility to construct a direct optimal method without auxiliary loops and logarithmic factors in the complexity bounds. Future research includes also oracle complexity separation [16,40] techniques for the setting when $f(x)$ and $h(y)$ are not prox-friendly and their gradients need to be evaluated.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

References

[1] A. Alacaoglu, Y. Malitsky, Stochastic variance reduction for variational inequality methods, in: Proceedings of Thirty Fifth Conference on Learning Theory, vol. 178, 2022, pp. 778–816.
[2] M. Alkousa, A. Gasnikov, D. Dvinskikh, D. Kovalev, F. Stonyakin, Accelerated methods for saddle-point problem, Comput. Math. Math. Phys. 60 (11) (2020) 1787–1809.
[3] Y. Carmon, Y. Jin, A. Sidford, K. Tian, Variance reduction for matrix games, in: Advances in Neural Information Processing Systems, 2019, pp. 11381–11392.
[4] A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, J. Math. Imaging Vis. 40 (1) (2011) 120–145.
[5] X. Chen, J. Wang, H. Ge, Training generative adversarial networks via primal-dual subgradient methods: a lagrangian perspective on gan, in: Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, 2018.
[6] Y. Chen, G. Lan, Y. Ouyang, Accelerated schemes for a class of variational inequalities, Math. Program. 165 (1) (2017) 113–149.
[7] P. Dvurechensky, Y. Nesterov, V. Spokoiny, Primal-dual methods for solving infinite-dimensional games, J. Optim. Theory Appl. 166 (1) (2015) 23–51.
[8] A. Gasnikov, Searching equillibriums in large transport networks, arXiv:1607.03142, 2016.
[9] A. Gasnikov, P. Dvurechensky, Y. Nesterov, Stochastic gradient methods with inexact oracle, Proc. Moscow Inst. Phys. Technol. 8 (1) (2016) 41–91.
[10] A. Gasnikov, D. Dvinskikh, P. Dvurechensky, D. Kamzolov, V. Matyukhin, D. Pasechnyuk, N. Tupitsa, A. Chernov, Accelerated meta-algorithm for convex optimization problems, Comput. Math. Math. Phys. 61 (1) (2021) 17–28.
[11] A.V. Gasnikov, Reduction of searching competetive equillibrium to the minimax problem in application to different network problems, Mat. Model. 27 (12) (2015) 121–136.
[12] Y. Han, G. Xie, Z. Zhang, Lower complexity bounds of finite-sum optimization problems: the results and construction, arXiv:2103.08280, 2021.
[13] L.T.K. Hien, R. Zhao, W.B. Haskell, An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems, arXiv:1711.03669, 2020.
[14] A. Ibrahim, W. Azizian, G. Gidel, I. Mitliagkas, Linear lower bounds and conditioning of differentiable games, in: International Conference on Machine Learning, PMLR, 2020, pp. 4583–4593.
[15] R. Isaacs, Differential Games: a Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization, Courier Corporation, 1999.
[16] A. Ivanova, P. Dvurechensky, E. Vorontsova, D. Pasechnyuk, A. Gasnikov, D. Dvinskikh, A. Tyurin, Oracle complexity separation in convex optimization, J. Optim. Theory Appl. 193 (1) (2022) 462–490.
[17] R. Johnson, T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc., 2013.

[18] D. Kovalev, A. Gasnikov, P. Richtárik, Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling, arXiv:2112.15199, 2021.
[19] G. Lan, First-Order and Stochastic Optimization Methods for Machine Learning, Springer, 2020.
[20] H. Lin, J. Mairal, Z. Harchaoui, A universal catalyst for first-order optimization, in: Proceedings of 29^{th} International Conference Neural Information Processing Systems, NIPS, 2015.
[21] H. Lin, J. Mairal, Z. Harchaoui, Catalyst acceleration for first-order convex optimization: from theory to practice, J. Mach. Learn. Res. (2018).
[22] T. Lin, C. Jin, M.I. Jordan, Near-optimal algorithms for minimax optimization, in: J. Abernethy, S. Agarwal (Eds.), Proceedings of Thirty Third Conference on Learning Theory, PMLR, Proc. Mach. Learn. Res. 125 (2020) 2738–2779.
[23] S. Lu, R. Singh, X. Chen, Y. Chen, M. Hong, Understand the dynamics of gans via primal-dual optimization, in: Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, 2018.
[24] J.J. Moreau, Proximité et dualité dans un espace hilbertien, Bull. Soc. Math. Fr. 93 (1965) 273–299.
[25] O. Morgenstern, J. Von Neumann, Theory of Games and Economic Behavior, Princeton University Press, 1953.
[26] J.F. Nash Jr, The bargaining problem, Econometrica (1950) 155–162.
[27] A. Nemirovski, Prox-method with rate of convergence $o(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, SIAM J. Optim. 15 (1) (2004) 229–251.
[28] A. Nemirovsky, D. Yudin, Problem Complexity and Method Efficiency in Optimization, J. Wiley & Sons, New York, 1983.
[29] Y. Nesterov, Introductory Lectures on Convex Optimization: A Basic Course, Springer, 2004.
[30] Y. Nesterov, Excessive gap technique in nonsmooth convex minimization, SIAM J. Optim. 16 (1) (2005) 235–249.
[31] Y. Nesterov, Smooth minimization of non-smooth functions, Math. Program. 103 (1) (2005) 127–152.
[32] Y. Nesterov, Dual extrapolation and its applications to solving variational inequalities and related problems, Math. Program. 109 (2–3) (2007) 319–344.
[33] Y. Nesterov, L. Scrimali, Solving strongly monotone variational and quasi-variational inequalities, https://doi.org/10.3934/dcds.2011.31.1383, 2011.
[34] B. Palaniappan, F. Bach, Stochastic variance reduction methods for saddle-point problems, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 29, Curran Associates, Inc., 2016.
[35] N. Parikh, S. Boyd, Proximal algorithms, Found. Trends Optim. 1 (3) (2014) 127–239, https://doi.org/10.1561/2400000003.
[36] S. Shalev-Shwartz, T. Zhang, Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization, in: E.P. Xing, T. Jebara (Eds.), Proceedings of the 31st International Conference on Machine Learning, PMLR, Bejing, China, in: Proceedings of Machine Learning Research, vol. 32, 2014, pp. 64–72.
[37] C. Song, S.J. Wright, J. Diakonikolas, Variance reduction via primal-dual accelerated dual averaging for nonsmooth convex finite-sums, in: Proceedings of the 38th International Conference on Machine Learning, vol. 139, 2021, pp. 9824–9834.
[38] F. Stonyakin, A. Tyurin, A. Gasnikov, P. Dvurechensky, A. Agafonov, D. Dvinskikh, M. Alkousa, D. Pasechnyuk, S. Artamonov, V. Piskunova, Inexact model: a framework for optimization and variational inequalities, Optim. Methods Softw. 36 (6) (2021) 1155–1201.
[39] F. Stonyakin, A. Gasnikov, P. Dvurechensky, A. Titov, M. Alkousa, Generalized mirror prox algorithm for monotone variational inequalities: universality and inexact oracle, J. Optim. Theory Appl. 194 (3) (2022) 988–1013.
[40] V. Tominin, Y. Tominin, E. Borodich, D. Kovalev, A. Gasnikov, P. Dvurechensky, On accelerated methods for saddle-point problems with composite structure, arXiv:2103.09344, 2021.
[41] B.E. Woodworth, N. Srebro, Tight complexity bounds for optimizing composite objectives, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 29, Curran Associates, Inc., 2016.
[42] G. Xie, Y. Han, Z. Zhang, Dippa: an improved method for bilinear saddle point problems, arXiv:2103.08270, 2021.
[43] R. Zhao, Accelerated stochastic algorithms for convex-concave saddle-point problems, Math. Oper. Res. 47 (2) (2022) 1443–1473.

[44] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc., Ser. B 67 (2) (2005) 301–320.