Solving smooth min-min and min-max problems by mixed oracle algorithms⋆

Egor Gladin[1,2][0000−0002−9086−996X], Abdurakhmon Sadiev[1], Alexander Gasnikov[1,3,5][0000−0002−7386−039X], Pavel Dvurechensky[4,3,5][0000−0003−1201−2343], Aleksandr Beznosikov[1,3][0000−0002−3217−3614], and Mohammad Alkousa[1][0000−0001−5470−0182]

[1] Moscow Institute of Physics and Technology, Russia
[2] Skolkovo Institute of Science and Technology, Russia
[3] HSE University, Russia
[4] Weierstrass Institute for Applied Analysis and Stochastics, Germany
[5] Institute for Information Transmission Problems RAS, Russia

Abstract. In this paper we consider two types of problems which have some similarity in their structure, namely, min-min problems and min-max saddle-point problems. Our approach is based on considering the outer minimization problem as a minimization problem with inexact oracle. This inexact oracle is calculated via inexact solution of the inner problem, which is either a minimization or a maximization problem. Our main assumptions are that the problem is smooth and the available oracle is mixed: it is only possible to evaluate the gradient w.r.t. the outer block of variables which corresponds to the outer minimization problem, whereas for the inner problem only zeroth-order oracle is available. To solve the inner problem we use accelerated gradient-free method with zeroth-order oracle. To solve the outer problem we use either inexact variant of the Vaydya's cutting-plane method or a variant of accelerated gradient method. As a result we propose a framework which leads to non-asymptotic complexity bounds for both min-min and min-max problems. Moreover, we estimate separately the number of first- and zeroth-order oracle calls which are sufficient to reach any desired accuracy.

Keywords: First-order methods · Zeroth-order methods · Cutting-plane methods · Saddle-point problems.

1 Introduction

In this paper, we consider smooth optimization problems in which the decision variable is decomposed into two blocks with minimization w.r.t. one block, which we call the outer block, and two types of operations w.r.t. the second block,

which we call the inner block: minimization or maximization. In other words, we consider smooth min-min problems and min-max problems. The main difference of our setting with existing in the literature is that we assume that it is possible to evaluate the gradient w.r.t. the outer block of the variables, i.e. first-order oracle, and only function values, i.e. zeroth-order oracle, when we deal with the inner block of variables. Thus, we operate with mixed type of oracle: first-order in one block of variables and zeroth-order in the second block of variables.

Our motivation, firstly, comes from min-max saddle-point problems, which have recently became of an increased interest in machine learning community in application to training Generative Adversarial Networks [11], and other adversarial models [17], as well as to robust reinforcement learning [23]. The standard process is to simultaneously train neural network, find adversarial examples and learn the network to distinguish the true examples from the artificially generated. In the training process the gradient is available through the backpropagation, whereas for the generating adversarial examples the network is available as a black box and only zeroth-order oracle is available. Another close application area is Adversarial Attacks [12,29] on neural networks, in particular the Black-Box Adversarial Attacks [18]. Here the goal is for a trained network to find a perturbation of the data in such a way that the network outputs wrong prediction. Then the training is repeated to make the network robust to such attacks. Since the attacking model does not have access to the architecture of the main network, but only to the input and output of the network, the only available oracle for the attacker is the zeroth-order oracle for the loss function. The motivation for min-min problems comes from simulation optimization [8,27], where some parts of the optimized system can be given as a black box with unavailable or computationally expensive gradients, and other parts of the objective are differentiable.

Separately zeroth-order [5] and first-order [21] are very well developed areas of modern numerical optimization. There are also plenty of works on first-order [14,20,19,3] and zeroth-order methods [32,16,2,26] for saddle-point problems. Our main idea and contribution in this paper is to consider mixed oracles, which seems to be an underdeveloped area of optimization and saddle-point problems. In [26] the authors consider methods with mixed oracle, but, unlike this work, only in the context of saddle-point problems and without acceleration techniques.

Notably, in this paper we develop a generic approach which is suitable for both types of problems: min-min and min-max, and is based on the same idea for both problems: we consider the minimization problem w.r.t. the outer group of variables as a minimization problem with inexact oracle. This inexact oracle is evaluated via inexact solution of the inner problem, which is either a minimization or a maximization problem. We carefully estimate with what accuracy one needs to solve the inner problem to be able to solve the outer problem with the desired accuracy. Moreover, we have to account for the random nature of the solution to the inner problem since we use randomized gradient-free methods with zeroth-order oracle to solve the inner problem. In our approach we consider two settings for the outer problem. If the dimension of the outer problem is small,

we use Vaydya's cutting-plane method [31,30], for which we extend the analysis to the case of approximate subgradients. The drawback of this method is that it scales quite badly with the dimension. Thus, if the dimension of the outer problem is large, we exploit accelerated gradient method, for which we develop an analysis in the case when an inexact oracle is available only with some probability, which may be of independent interest. Our approach based on inner-outer loops allows also to separate complexities, i.e. the number of calls to each of the oracles: first-order oracle for the outer block of variables and zeroth-order oracle for the inner block of variables.

The rest of the paper is organized as follows. First we consider min-min problems in two settings: small and large dimension of the outer problem. In the first case we develop an inexact variant of the Vaydya's method use it in the outer loop in a combination with accelerated random gradient-free method in the inner loop. In the second case, we apply accelerated gradient method in the outer loop combined with the same method in the inner loop. After that we consider saddle-point min-max problems again in two settings. When the dimension of the outer problem is small, we use the same scheme with inexact Vaydya's method and accelerated random gradient-free method. The situation is more complicated when the dimension of the outer problem is large. In this case we use a three-loop structure with the Catalyst acceleration scheme [15] combined with accelerated gradient method with inexact oracle and accelerated random gradient-free method.

Table 1. Main results

		Oracle Complexity	
		1-st order	**0**-th order
Min Min	Small Scale	$\widetilde{O}\left(n_x\right)$	$\widetilde{O}\left(n_x n_y \sqrt{\frac{L_{yy}}{\mu_y}}\right)$
Min Max	Small Scale	$\widetilde{O}\left(n_x\right)$	$\widetilde{O}\left(n_x n_y \sqrt{\frac{L_{yy}}{\mu_y}}\right)$
	Large Scale	$\widetilde{O}\left(\sqrt{\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x \mu_y}}\right)$	$\widetilde{O}\left(n_y \sqrt{\frac{L_{xx}L_{yy}}{\mu_x \mu_y} + \frac{2L_{xy}^2}{\mu_x \mu_y}}\right)$

2   Solving Min-Min Problems

Consider the problem

$$\min_{x \in \mathcal{X}} \min_{y \in \mathbb{R}^{n_y}} f(x,y), \tag{1}$$

where $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ is a closed convex set, $f(x,y)$ is a convex function equipped with a mixed oracle, i.e. we have access to a first-order oracle for the outer

problem (minimization w.r.t. x) and a zeroth-order oracle for the inner problem (minimization w.r.t. y). In the sections below we will describe the two approaches to solving such problems together with additional assumptions they require.

The general idea of the approaches is as follows. Let us introduce the function

$$g(x) = \min_{y \in \mathbb{R}^{n_y}} f(x, y) \tag{2}$$

and rewrite the initial problem (1) as

$$\min_{x \in \mathcal{X}} g(x). \tag{3}$$

Using an iterative method for the outer problem (3) requires solving the inner problem (2) numerically on each iteration. An error of the solution of the inner problem results in an inexact oracle for the outer problem.

2.1   Small dimension of the outer problem

The approach described in the present subsection requires the following assumptions about the problem (1):

1. $\mathcal{X} \subset \mathbb{R}^{n_x}$ is a compact convex set with nonempty interior;
2. n_x is relatively small (up to a hundred);
3. $f(x, y)$ is a continuous convex function which is also μ_y-strongly convex in y;
4. for all $x \in \mathcal{X}$ the function $f(x, \cdot)$ is L_{yy}-smooth, i.e.

$$\|\nabla_y f(x, y) - \nabla_y f(x, y')\|_2 \le L_{yy} \|y - y'\|_2 \quad \forall y, y' \in \mathbb{R}^{n_y}.$$

5. for any $x \in \mathcal{X}$ the minimization problem (2) has solution $y(x)$.

The algorithms used in the proposed approach and related convergence theorems are given in the subsequent paragraphs. Our proposed approach goes as follows:

Approach 1. *The outer problem* (3) *is solved via Vaidya's cutting plane method [30,31]. The inner problem* (2) *is solved via Accelerated Randomized Directional Derivative method for strongly convex functions (ARDDsc) [7], see Algorithm 2.*[6]

The complexity of approach 1 is given in the following theorem:

Theorem 1. *Approach 1 arrives at ε-solution of the problem* (3) *after*[7] $\widetilde{O}(n_x)$ *calls to the first-order oracle and* $\widetilde{O}\left(n_x n_y \sqrt{\frac{L_{yy}}{\mu_y}}\right)$ *calls to the zeroth-order oracle.*

---

[6] Here and below instead of ARDDsc we can use Accelerated coordinate descent methods [22,10] with replacing partial derivatives by finite-differences. In this case we lost opportunity to play on the choice of the norm (that could save $\sqrt{n_y}$-factor in gradient-free oracle complexity estimate [7]), but, we gain a possibility to replace the wort case L_{yy} to the average one (that could be n_y-times smaller [22]). At the end this could also save $\sqrt{n_y}$-factor in gradient-free oracle complexity estimate [22].

[7] $\widetilde{O}(\cdot) = O(\cdot)$ up to a small power of logarithmic factor

Remark 1. As far as the arithmetic complexity of the iteration is concerned, Vaidya's cutting plane method involves inversions of $n_x \times n_x$ matrices, hence the assumption that n_x is relatively small.

The complexity bounds from theorem 1 are derived in the paragraph *Analysis of the approach* which follows the description of algorithms.

Vaidya's cutting plane method Vaidya proposed a cutting plane method [30,31] for solving problems of the form

$$\min_{x \in \mathcal{X}} g(x), \tag{4}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is a compact convex set with non-empty interior, and $g : \mathcal{X} \to \mathbb{R}$ is a continuous convex function.

We will now introduce the notation and describe the algorithm. Let $P = \{x \in \mathbb{R}^n : Ax \geq b\}$ be the bounded full-dimensional polytope, where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. The logarithmic barrier for P is defined as

$$L(x) := -\sum_{i=1}^m \ln\left(a_i^\top x - b_i\right),$$

where $a_i^\top$ is the ith row of A. The Hessian of $L(x)$ is given by

$$H(x) = \sum_{i=1}^m \frac{a_i a_i^\top}{\left(a_i^\top x - b_i\right)^2} \tag{5}$$

and is positive definite for all x in the interior of P. The *volumetric barrier* for P is defined as

$$F(x) = \frac{1}{2} \ln\left(\det H(x)\right),$$

where $\det H(x)$ denotes the determinant of $H(x)$. The point $\boldsymbol{\omega}$ that minimizes $F(x)$ over P will be called the *volumetric center* of P. Let $\sigma_i(x)$ be defined as

$$\sigma_i(x) = \frac{a_i^\top \left(H(x)\right)^{-1} a_i}{\left(a_i^\top x - b_i\right)^2}, \quad 1 \leq i \leq m. \tag{6}$$

Now, let $\mathcal{R}$ be a radius of some Euclidean ball $\mathcal{B}_\mathcal{R}$ that contains $\mathcal{X}$. W.l.o.g we will assume that $\mathcal{B}_\mathcal{R}$ is centered at the origin. The parameters of the method $\eta > 0$ and $\gamma > 0$ are small constants such that $\eta \leqslant 10^{-4}$, and $\gamma \leqslant 10^{-3}\eta$. The algorithm starts out with the simplex

$$P_0 = \left\{x \in \mathbb{R}^n : x_j \geqslant -\mathcal{R}, j = \overline{1, n}, \sum_{j=1}^n x_j \leqslant n\mathcal{R}\right\} \supseteq \mathcal{B}_\mathcal{R} \supseteq \mathcal{X}. \tag{7}$$

and produces a sequence of pairs $(A_k, b_k) \in \mathbb{R}^{m_k \times n} \times \mathbb{R}^{m_k}$, such that the corresponding polytope $P_k = \{x \in \mathbb{R}^n : A_k x \geq b_k\}$ always contains a solution of the

problem (4). At the beginning of each iteration k we have an approximation z_k to the volumetric center of P_k (for more details on computing the approximation see [31,30]). In particular, on the 0-th iteration we can compute the volumetric center explicitly:

Proposition 1. *The volumetric center for P_0 is $\boldsymbol{\omega} = \omega\mathbf{1}_n$, where $\omega := \frac{n-1}{n+1}\mathcal{R}$ and $\mathbf{1}_n$ denotes the vector $(1,\ldots,1)^\top \in \mathbb{R}^n$.*

For $k \geq 0$, the next polytope (A_{k+1}, b_{k+1}) is defined by either adding or removing a constraint to the current polytope, depending on the values $\{\sigma_i(z_k)\}_{i=1}^m$ associated to P_k:

1. If for some $i \in \{1,\ldots,m\}$ one has $\sigma_i(z_k) = \min\limits_{1\leq j\leq m} \sigma_j(z_k) < \gamma$, then (A_{k+1}, b_{k+1}) is defined by removing the ith row from (A_k, b_k).

2. Otherwise, i.e. if $\min\limits_{1\leq j\leq m} \sigma_j(z_k) \geq \gamma$, the oracle is called with the current point z_k as input. If $z_k \in \mathcal{X}$, it returns a vector c_k, such that $-c_k \in \partial g(z_k)$, i.e. $-c_k$ is a subgradient of g at z_k. Otherwise, it returns a vector c_k such that $c_k^\top x \geq c_k^\top z_k \ \forall x \in \mathcal{X}$. We choose $\beta_k \in \mathbb{R}$ such that $c_k^\top z_k \geq \beta_k$ and

$$\frac{c_k^\top \left(H(z_k)\right)^{-1} c_k}{\left(c_k^\top z_k - \beta_k\right)^2} = \frac{1}{2}\sqrt{\eta\gamma}.$$

Then we define (A_{k+1}, b_{k+1}) by adding the row given by $\left(c_k^\top, \beta_k\right)$ to (A_k, b_k).

After N iterations, the method returns a point $x^N := \arg\min\limits_{1\leq k\leq N} g(z_k)$.

Now, let us introduce the concept of inexact subgradient.

Definition 1. *The vector $c \in \mathbb{R}^n$ is called a δ-subgradient of a convex function g at $z \in \operatorname{dom} f$ (we denote $c \in \partial_\delta g(z)$), if*

$$g(x) \geq g(z) + c^\top(x - z) - \delta \quad \forall x \in \operatorname{dom} f.$$

In fact, a δ-subgradient can be used in Vaidya's method instead of the exact subgradient. In this case, we will call the algorithm Vaidya's method with δ-subgradient. We will now present a theorem that justifies this claim.

Theorem 2. *Let $\mathcal{B}_\rho$ and $\mathcal{B}_\mathcal{R}$ be some Euclidean balls of radii ρ and $\mathcal{R}$, respectively, such that $\mathcal{B}_\rho \subseteq \mathcal{X} \subseteq \mathcal{B}_\mathcal{R}$, and let a number $B > 0$ be such that $|g(x) - g(x')| \leq B \ \forall x, x' \in \mathcal{X}$. After $N \geq \frac{2n}{\gamma}\ln\left(\frac{n^{1.5}\mathcal{R}}{\gamma\rho}\right) + \frac{1}{\gamma}\ln\pi$ iterations Vaidya's method with δ-subgradient for the problem (4) returns a point x^N such that*

$$g(x^N) - g(x_*) \leq \frac{Bn^{1.5}\mathcal{R}}{\gamma\rho}\exp\left(\frac{\ln\pi - \gamma N}{2n}\right) + \delta, \tag{8}$$

where $\gamma > 0$ is the parameter of the algorithm and x_* is a solution of the problem (4).

Accelerated Randomized Directional Derivative method We refer to the work [7]. For convenience, we present algorithms from this paper, taking into account that the problem will be a classical optimization problem:

$$\min_{x \in \mathbb{R}^{n_x}} f(x). \tag{9}$$

Gradient approximation:

$$\text{grad}_f(x, \tau, \mathbf{e}) = \frac{n_x}{\tau} \left(f(x + \tau \mathbf{e}) - f(x) \right) \mathbf{e}, \tag{10}$$

where $\mathbf{e} \in \mathcal{RS}_2^{n_x}(1)$, i.e. be a random vector uniformly distributed on the surface of the unit Euclidean sphere in $\mathbb{R}^{n_x}$.

Definition 2. *Function $d(x) : \mathbb{R}^{n_x} \to \mathbb{R}$ is called prox-function if $d(x)$ is 1-strongly convex w.r.t. $\| \cdot \|_p$-norm and differentiable on $\mathbb{R}^{n_x}$ function.*

It is worth noting that in the case of $p = 2$, the prox-function $d(x)$ looks like this

$$d(x) = \frac{1}{2} \|x\|_2^2$$

Definition 3. *Let $d(x) : \mathbb{R}^{n_x} \to \mathbb{R}$ is prox-function. For any two points $x, x' \in \mathbb{R}^{n_x}$ we define Bregman divergence $V_x(x')$ associated with $d(x)$ as follows:*

$$V_x(x') = d(x') - d(x) - \langle \nabla d(x), x' - x \rangle.$$

It is worth noting that in the case of $p = 2$, the Bregman divergence $V_x(x')$ looks like this

$$V_x(x') = \frac{1}{2} \|x' - x\|_2^2$$

Let x^* be fixed point and x be random vector such that $\mathbb{E}_x \|x - x^*\|_p^2 \leq R_p^2$, then

$$\mathbb{E}_x d \left(\frac{x - x^*}{R_p} \right) \leq \frac{\Omega_p}{2}, \tag{11}$$

where $\mathbb{E}_x$ denotes the expectation with respect to random vector x and Ω_p is defined as follows. We note that for $p = 2$ $\Omega_p = 1$

Algorithm 1 Accelerated Randomized Directional Derivative (ARDD) method [7]

Input: x_0 - starting point, N - number of iterations, L - smoothness parameter, τ.

$y_0 := x_0, \ w_0 = x_0$

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

    Sample $\mathbf{e}_{k+1} \in \mathcal{RS}_2^{n_x}(1)$.

    Set
$$t_k := \frac{2}{k+2}, \ x_{k+1} := t_k w_k + (1 - t_k) y_k$$

    Calculate $\mathrm{grad}_f(x_{k+1}, \tau, \mathbf{e}_{k+1})$ using (10)

    Compute
$$y_{k+1} := x_{k+1} - \frac{1}{2L} \mathrm{grad}_f(x_{k+1}, \tau, \mathbf{e}_{k+1})$$

    Set
$$\alpha_k := \frac{k+1}{96n^2 L}$$

    Compute
$$w_{k+1} := \underset{z \in \mathbb{R}^{n_x}}{\mathrm{argmin}} \left\{ \alpha_{k+1} \langle \mathrm{grad}_f(x_{k+1}, \tau_k, \mathbf{e}_{k+1}), w - w_k \rangle + V_{z_k}(z) \right\}$$

end for

Output: y_N or $\bar{y}_N$.

Algorithm 2 Accelerated Randomized Directional Derivative method for strongly convex functions (ARDDsc) [7]

Input: x_0 - starting point s.t. $\|x_0 - x^*\|_p^2$, N -number of iterations, μ_p -strong convexity parameter.

Set
$$N_0 = \left\lceil \sqrt{\frac{8 a L_2 \Omega_p}{\mu_p}} \right\rceil$$

where $a = 384 n^2 \rho_n$, $\rho_n = \min\left\{q - 1, 16\ln n - 8\right\} n^{\frac{2}{q}-1}$.

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

    Set $R_k^2 = R_p^2 2^{-k}$

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

    Run ARDD 1 with starting point u_k and prox-function $d_k(x)$ 2 for N_0 steps.

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

end for

Output: u_N

Theorem 3 (see [7]). *Let $p \in [1, 2]$ and $q \in [2, +\infty]$ be defined such that $\frac{1}{p} + \frac{1}{q} = 1$. Let function f in problem (9) be μ_p-strongly convex w.r.t. $\|.\|_p$ and L_2-smooth w.r.t. $\|.\|_2$ and ARDDsc method 2 be applied to solve this problem.*

Then

$$\mathbb{E}f(u_N) - \min_{x \in \mathbb{R}^{n_x}} f(x) \leq \frac{\mu_p R_p^2}{2} 2^{-N}$$

Moreover, the oracle complexity to achieve ε-accuracy of the solution is

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

Analysis of the approach 1 Fix a point $x' \in \mathcal{X}$. The following theorem gives the recipe to obtaining the δ-subgradient $c \in \partial_\delta g(x')$ for the outer problem (3):

Theorem 4. *Let $\tilde{\delta} > 0$ and $\tilde{y} \in \mathbb{R}^{n_y}$ satisfy $f(x', \tilde{y}) - g(x') \leq \tilde{\delta}$, then $\partial_x f(x', \tilde{y}) \in \partial_\delta g(x')$ with*

$$\delta = 2\left(\sqrt{\tilde{\delta}} + 2\sqrt{D}\right)\sqrt{\frac{L_{yy}\tilde{\delta}}{\mu_y}}, \tag{12}$$

where $D := \max_{x \in \mathcal{X}} (f(x, \mathbf{0}) - g(x)) < +\infty$.

The theorem 4 is based on the two following lemmas:

Lemma 1. *Let $h : \mathbb{R}^{n_y} \to \mathbb{R}$ be an L-smooth convex function, and let the point $\tilde{y} \in \mathbb{R}^{n_y}$ satisfy $h(\tilde{y}) - h(y_*) \leq \tilde{\delta}$ for some delta > 0, where $y_* \in \underset{y \in \mathbb{R}^{n_y}}{\mathrm{Argmin}} \, h(y)$.*

Then

$$\langle \nabla h(\tilde{y}), \tilde{y} - y \rangle \leq \|\tilde{y} - y\|_2 \sqrt{2L\tilde{\delta}} \quad \forall y \in \mathbb{R}^{n_y}.$$

Lemma 2 (see [9], p.12). *Let $\tilde{y} \in \mathbb{R}^{n_y}$ satisfy*

$$\langle \nabla_y f(x, \tilde{y}), \tilde{y} - y(x') \rangle \leq \delta \quad \forall x' \in \mathcal{X}, \tag{13}$$

then $\partial_x f(x, \tilde{y}) \in \partial_\delta g(x)$.

According to theorem 4, we need to solve the inner problem (2) with sufficient accuracy to obtain the δ-subgradient. In fact, we can simplify the formula (12) to

$$\delta = 6\sqrt{\frac{L_{yy}D\tilde{\delta}}{\mu_y}}. \tag{14}$$

Indeed, since $\tilde{y}$ is an $\tilde{\delta}$-solution of the inner problem (2), we can assume that $\tilde{\delta} \leq D \equiv \max_{x \in \mathcal{X}}(f(x, \mathbf{0}) - g(x))$. If this inequality doesn't hold, we can always take $\mathbf{0}$ as an approximate solution of (2).

Now, to derive the complexity of approach 1, we will use the theorem 2 and put

$$\frac{Bn^{1.5}\mathcal{R}}{\gamma\rho} \exp\left(\frac{\ln \pi - \gamma N}{2n}\right) = \frac{\varepsilon}{2} \text{ and } \delta = \frac{\varepsilon}{2},$$

i.e. Vaidya's method will perform

$$N_x = O\left(n_x \ln\left(\frac{n_x^{1.5} B\mathcal{R}}{\varepsilon\rho}\right)\right),$$

steps (first-order oracle calls), and at each of them ARDDsc will perform

$$N_y = \widetilde{O}\left(n_y\sqrt{\frac{L_{yy}}{\mu_y}}\right)$$

iterations (see theorem 3). Thus, the number of zeroth-order oracle calls is

$$N_x \cdot N_y = \widetilde{O}\left(n_x n_y\sqrt{\frac{L_{yy}}{\mu_y}}\right),$$

which finishes the analysis of approach 1.

3 Solving Min-Max Saddle-Point Problems

Consider the problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathbb{R}^{n_y}} f(x, y), \tag{15}$$

where $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ is a closed convex set, $f(x, y)$ is a convex-concave function (i.e. convex in x and concave in y) equipped with a mixed oracle, i.e. we have access to a first-order oracle for the outer problem (minimization w.r.t. x) and a zeroth-order oracle for the inner problem (maximization w.r.t. y). In the subsections below we describe the two approaches for solving such problems together with additional assumptions they require.

The general idea of the approaches is as follows. Let us introduce the function

$$g(x) = \max_{y\in\mathbb{R}^{n_y}} f(x, y) \tag{16}$$

and rewrite the initial problem (15) as

$$\min_{x\in\mathcal{X}} g(x). \tag{17}$$

Using an iterative method for the outer problem (17) requires solving the inner problem (16) numerically in each iteration. An error of the solution of the inner problem results in an inexact oracle for the outer problem.

3.1 Small dimension of the outer problem

The approach described in the present section requires the following assumptions about the problem (15):

1. $\mathcal{X} \subset \mathbb{R}^{n_x}$ is a compact convex set with nonempty interior;

2. n_x is relatively small (up to a hundred);
3. $f(x, y)$ is a continuous function which is convex in x and μ_y-strongly concave in y;
4. for all $x \in \mathcal{X}$ the function $f(x, \cdot)$ is L_{yy}-smooth, i.e.

$$\|\nabla_y f(x, y) - \nabla_y f(x, y')\|_2 \le L_{yy} \|y - y'\|_2 \quad \forall y, y' \in \mathbb{R}^{n_y}.$$

5. for any $x \in \mathcal{X}$ the maximization problem (16) has solution $y(x)$.

The algorithms used in the approach and related convergence theorems are given in the previous section. The approach goes as follows:

Approach 2. *The outer problem* (17) *is solved via Vaidya's cutting plane method. The inner problem* (2) *is solved via ARDDsc, see Algorithm 2.*

Complexity of the approach is given in the following theorem:

Theorem 5. *Approach 2 arrives at ε-solution of the problem* (17) *after $\widetilde{O}(n_x)$ calls to the first-order oracle and $\widetilde{O}\left(n_x n_y \sqrt{\frac{L_{yy}}{\mu_y}}\right)$ calls to the zeroth-order oracle.*

Remark 2. As far as the arithmetic complexity of the iteration is concerned, Vaidya's cutting plane method involves inversions of $n_x \times n_x$ matrices, hence the assumption that n_x is relatively small.

Complexity bounds from theorem 5 are derived in the following paragraph.

Analysis of the approach 2 Fix a point $x' \in \mathcal{X}$. The following lemma gives the recipe to obtaining the δ-subgradient $c \in \partial_\delta g(x')$ for the outer problem (17):

Lemma 3 (see [24]). *Let $\tilde{y} \in \mathbb{R}^{n_y}$ satisfy $g(x') - f(x', \tilde{y}) \le \delta$, then $\partial_x f(x', \tilde{y}) \in \partial_\delta g(x')$.*

According to lemma 3, we need to solve the inner problem (16) with accuracy δ to obtain the δ-subgradient.

Now, to derive the complexity of approach 2, we will use the theorem 2 and put

$$\frac{Bn^{1.5}\mathcal{R}}{\gamma\rho} \exp\left(\frac{\ln \pi - \gamma N}{2n}\right) = \frac{\varepsilon}{2} \text{ and } \delta = \frac{\varepsilon}{2},$$

i.e. Vaidya's method will perform

$$N_x = O\left(n_x \ln\left(\frac{n_x^{1.5} B \mathcal{R}}{\varepsilon\rho}\right)\right),$$

steps (first-order oracle calls), and at each of them ARDDsc will perform

$$N_y = \widetilde{O}\left(n_y\sqrt{\frac{L_{yy}}{\mu_y}}\right)$$

iterations (see theorem 3). Thus, the number of zeroth-order oracle calls is

$$N_x \cdot N_y = \widetilde{O}\left(n_x n_y \sqrt{\frac{L_{yy}}{\mu_y}}\right),$$

which finishes the analysis of approach 2.

3.2   Large dimension of the outer problem

For a detailed study of the convergence of the methods, we introduce some assumptions about the objective function $f(x, y)$.

Assumption 1. $f(x, y)$ is convex-concave. It means that $f(\cdot, y)$ is convex for all y and $f(x, \cdot)$ is concave for all x.

Assumption 1(s). $f(x, y)$ is strongly-convex-strongly-concave. It means that $f(\cdot, y)$ is μ_x-strongly convex for all y and $f(x, \cdot)$ is μ_y-strongly concave for all x w.r.t. $\| \cdot \|_2$, i.e. for all $x_1, x_2 \in \mathcal{X}$ and for all $y_1, y_2 \in \mathbb{R}^{n_y}$ we have

$$f(x_1, y_2) \geq f(x_2, y_2) + \langle \nabla_x f(x_2, y_2), x_1 - x_2 \rangle + \frac{\mu_x}{2}\|x_1 - x_2\|_2^2,$$

$$-f(x_2, y_1) \geq -f(x_2, y_2) - \langle \nabla_y f(x_2, y_2), y_1 - y_2 \rangle + \frac{\mu_y}{2}\|y_1 - y_2\|_2^2. \quad (18)$$

Assumption 2. $f(x, y)$ is (L_{xx}, L_{xy}, L_{yy})-smooth w.r.t $\| \cdot \|_2$, i.e. for all $x, x' \in \mathcal{X}, \; y, y' \in \mathbb{R}^{n_y}$

$$\begin{aligned}
\|\nabla_x f(x, y) - \nabla_x f(x', y)\|_2 &\leq L_{xx} \|x - x'\|_2 ; \\
\|\nabla_x f(x, y) - \nabla_x f(x, y')\|_2 &\leq L_{xy} \|y - y'\|_2 \\
\|\nabla_y f(x, y) - \nabla_y f(x', y)\|_2 &\leq L_{xy} \|x - x'\|_2 . \\
\|\nabla_y f(x, y) - \nabla_y f(x, y')\|_2 &\leq L_{yy} \|y - y'\|_2 .
\end{aligned} \quad (19)$$

As mentioned above, we have access to a first-order oracle $\nabla_x f(x, y)$ for the outer problem (minimization problem with variables x) and a zeroth-order oracle $f(x, y)$ for the inner problem (maximization problem with variables y). Since we do not have access to the values of the gradient $\nabla_y f(x, y)$, it is logical to approximate it using finite differences using the value of the function $f(x, y)$ at two close points as follows

$$\mathrm{grad}_f(x, y, \tau, \mathbf{e}) = -\frac{n}{\tau}\left(f(x, y + \tau \mathbf{e}) - f(x, y)\right)\mathbf{e}, \quad (20)$$

where $\mathbf{e} \in \mathcal{RS}_2^{n_y}(1)$, i.e. is a random vector uniformly distributed on the surface of the unit Euclidean sphere in $\mathbb{R}^{n_y}$. So we get a mixed oracle

$$G(x, y, \tau, \mathbf{e}) = \begin{pmatrix} \nabla_x f(x, y) \\ \mathrm{grad}_f(x, y, \tau, \mathbf{e}) \end{pmatrix}. \quad (21)$$

Using the mixed oracle (21), we provide our approach for solving the initial saddle-point problem (15). First, we can use the following trick with the help of Sion's theorem:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathbb{R}^{n_y}} f(x, y) = \max_{y \in \mathbb{R}^{n_y}} \min_{x \in \mathcal{X}} f(x, y) = \max_{y \in \mathbb{R}^{n_y}} h(y), \text{ where } h(y) = \min_{x \in \mathcal{X}} f(x, y).$$

For the new problem, we apply the Catalyst algorithm [15] to the outer maximization problem:

$$\max_{y \in \mathbb{R}^{n_y}} \left\{ h(y) = \min_{x \in \mathcal{X}} f(x,y) \right\}. \tag{22}$$

Algorithm 3 Catalyst [15]

1: **Input:** starting point x_0, parameters H_1 and α_0, accuracy of solution to subproblem $\tilde{\varepsilon}$, optimization method $\mathcal{M}$.

2: Initialize
$$q = \frac{\mu_y}{(\mu_y + H_1)}$$

3: **while** the desired stopping criterion is not satisfied **do**

4:     Find an approximate solution of the following problem using $\mathcal{M}$:

$$y_k \approx \operatorname*{argmax}_{y \in \mathbb{R}^{n_y}} \left\{ \varphi_k(y) = h(y) - \frac{H_1}{2} \|y - z_{k-1}\|_2^2 \right\}, \tag{23}$$

$$\text{such that } \varphi_k^* - \varphi_k(y_k) \leq \tilde{\varepsilon}$$

5:     Compute $\alpha_k \in (0,1)$ from equation

$$\alpha_k^2 = (1 - \alpha_i)\alpha_{k-1}^2 + q\alpha_k$$

6:     Compute

$$z_k = y_k + \beta_k (y_k - y_{k-1}), \text{ where } \beta_k = \frac{\alpha_{k-1}(1 - \alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}$$

7: **end while**

8: **Output:** y_{final}.

Now the question arises how to solve the auxiliary problem (23). This subproblem is equivalent to solving the following problem:

$$\max_{y \in \mathbb{R}^{n_y}} \left\{ h(y) - \frac{H_1}{2} \|y - z_{k-1}\|_2^2 \right\} = \max_{y \in \mathbb{R}^{n_y}} \left\{ \min_{x \in \mathcal{X}} \{f(x,y)\} - \frac{H_1}{2} \|y - z_{k-1}\|_2^2 \right\},$$

for which we again use the Sion's theorem and rewrite equivalently the problem as:

$$\max_{y \in \mathbb{R}^{n_y}} \min_{x \in \mathcal{X}} \left\{ f(x,y) - \frac{H_1}{2} \|y - z_{k-1}\|_2^2 \right\} = \min_{x \in \mathcal{X}} \max_{y \in \mathbb{R}^{n_y}} \left\{ f(x,y) - \frac{H_1}{2} \|y - z_{k-1}\|_2^2 \right\}.$$

For convenience, we denote

$$\psi(x,y) = f(x,y) - \frac{H_1}{2} \|y - z_k\|_2^2. \tag{24}$$

Thus, to solve the auxiliary problem (23), we first solve the following saddle-point problem

$$\min_{x \in \mathcal{X}} \max_{y \in \mathbb{R}^{n_y}} \psi(x, y) = \min_{x \in \mathcal{X}} \max_{y \in \mathbb{R}^{n_y}} \left\{ f(x, y) - \frac{H_1}{2} \|y - z_{k-1}\|_2^2 \right\}. \qquad (25)$$

This saddle-point problem (25) can be considered as an optimization problem for a certain function. Indeed, let us introduce a function

$$\xi(x) = \max_{y \in \mathbb{R}^{n_y}} \psi(x, y), \qquad (26)$$

and rewrite the initial problem (25) as follows:

$$\min_{x \in \mathcal{X}} \xi(x). \qquad (27)$$

To solve problem (25), we solve the outer minimization problem with respect to the variable x by the fast adaptive gradient method with inexact oracle. In each iteration of this method, to find the inexact first-order oracle for the outer problem, we solve the inner problem (26). Since for this inner problem we have access only to the zeroth-order oracle, we use accelerated gradient-free method ARDDsc [7]. Our approach is summarized as follows.

Approach 3. *The outer problem* (17) *is solved via Catalyst Algorithm 3. The subproblem* (23) *is solved as saddle-point problem* (25)*. The outer problem* (27) *is solved via Fast Adaptive Gradient Method (Algorithm 4). At each iteration of Algorithm 4 the inner problem* (26) *is solved via ARDDsc, see Algorithm 2, for case $p = 2$ (that is, prox-function $d(x) = \frac{1}{2}\|x\|_2^2$, see Definition 2, Bregman divergence $V_x(x') = \frac{1}{2}\|x' - x\|_2^2$, see Definition 3, and $\Omega_p = \Omega_2 = 2$).*

Analysis of fast adaptive gradient method with (δ, σ, L, μ)-oracle (Algorithm 4) For our analysis, due to the fact that Algorithm 2 is randomized, we need not just a fast gradient method for solving the outer problem (27), but a fast adaptive gradient method for (δ, σ, L, μ)-oracle. This is the extension of the fast adaptive gradient method from [28]. To understand this problem in depth and in detail, we need to carefully consider the concept of (δ, σ, L, μ)- oracle and perform a deep and thorough convergence analysis of the fast gradient method using such a seemingly unusual oracle. To that end, we consider the following general minimization problem:

$$\min_{x \in \mathcal{X}} f(x). \qquad (28)$$

Definition 4 ((δ, σ, L, μ)-oracle). *Let function f be convex on convex set $\mathcal{X}$. We say that it is equipped with a first-order (δ, σ, L, μ)-oracle if, for any $x' \in \mathcal{X}$, we can compute a pair $(f_{\delta, L, \mu}(x'), g_{\delta, L, \mu}(x')) \in \mathbb{R} \times \mathbb{R}^{n_x}$ such that with probability at least $1 - \sigma$*

$$\frac{\mu}{2}\|x' - x\|^2 \le f(x) - (f_{\delta, L, \mu}(x') + \langle g_{\delta, L, \mu}(x'), x - x' \rangle) \le \frac{L}{2}\|x - x'\|^2 + \delta. \quad (29)$$

Definition 5 ((ε, σ)-solution). *Let $\varepsilon > 0$ be the target accuracy of the solution and $\sigma \in (0, 1)$ be the target confidence level. We say that a random point $\hat{x} \in \mathcal{X}$ is (ε, σ)-solution to problem (28) if*

$$\mathbb{P}\left\{ f(\hat{x}) - \min_{x \in \mathcal{X}} f(x) \leq \varepsilon \right\} \geq 1 - \sigma. \qquad (30)$$

If $\sigma = 0$, we say that $\hat{x} \in \mathcal{X}$ is an ε-solution to problem (28).

Algorithm 4 Fast adaptive gradient method with (δ, σ, L, μ)-oracle [28]

1: **Input:** starting point x_0, $L_0 > 0$, $\mu \geq 0$, sequence $\{\delta\}_{k \geq 0}$.
2: $y_0 := x_0$, $u_0 := x_0$, $\alpha_0 := 0$, $A_0 := \alpha_0$
3: **for** $k \geq 0$ **do**
4:     Find the smallest integer $i_k \geq 0$ such that

$$f_{\delta_k, L, \mu}(x_{k+1}) \leq f_{\delta_k, L, \mu}(y_{k+1}) + \langle g_{\delta, L, \mu}(y_{k+1}), x_{k+1} - y_{k+1} \rangle + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|_2^2 + \delta_k,$$

   where $L_{k+1} = 2^{i_k - 1} L_k$.
5:     Compute α_{k+1} such that α_{k+1} is the largest root of

$$A_{k+1}(1 + A_k \mu) = L_{k+1}\alpha_{k+1}^2, \quad \text{where } A_{k+1} := A_k + \alpha_{k+1}$$

6:     $y_{k+1} = \frac{\alpha_{k+1} u_k + A_k x_k}{A_{k+1}}$

$$\phi_{k+1}(x) = \alpha_{k+1}\langle g_{\delta, L, \mu}(y_{k+1}), x - y_{k+1}\rangle + \frac{(1 + A_k \mu)}{2}\|x - u_k\|_2^2 + \frac{\alpha_{k+1}\mu}{2}\|x - y_{k+1}\|_2^2$$

7:     $u_{k+1} := \underset{x \in \mathcal{X}}{\operatorname{argmin}} \, \phi_{k+1}(x)$
8:     $x_{k+1} = \frac{\alpha_{k+1} u_{k+1} + A_k x_k}{A_{k+1}}$
9: **end for**
10: **Output:** x_{k+1}.

Note that the problem $\underset{x \in \mathcal{X}}{\operatorname{argmin}} \, \phi_{k+1}(x)$ is solved exactly in each iteration.

Theorem 6. *Let function f be convex on convex set $\mathcal{X}$ and be equipped with a first-order (δ, σ, L, μ)-oracle. Then, after N iterations of Algorithm 4 applied to problem (28), we have that with probability at least $(1 - N\sigma)$:*

$$f(x_N) - f(x^*) \leq 2L \exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu}{L}}\right) R_2^2 + \frac{2\sum_{k=0}^{N-1} A_{k+1}\delta_k}{A_N}, \qquad (31)$$

where R_2 is such that $\frac{1}{2}\|x_0 - x^*\|_2^2 \leq R_2^2$ and x_0 is the starting point.

Corollary 1. *Let function f be convex on convex set $\mathcal{X}$ and be equipped with a first-order $(\delta, \frac{\sigma}{N}, L, \mu)$-oracle. If the sequence $\{\delta\}_{k \geq 0}$ is bounded by δ, we have with probability at least $(1 - \sigma)$:*

$$f(x_N) - f(x^*) \leq 2L \exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu}{L}}\right) R_2^2 + \left(1 + \sqrt{\frac{L}{\mu}}\right)\delta \qquad (32)$$

where R_2 is such that $\frac{1}{2}\|x_0 - x^*\|_2^2 \leq R_2^2$ and x_0 is the starting point.

Proof. To prove this statement, we give an auxiliary lemma

Lemma 1 (see [6]). *The sequence $\{A_k\}_{k \leq 0}$ satisfies*

$$\frac{\sum_{i=0}^{k} A_i}{A_k} \leq 1 + \sqrt{\frac{L}{\mu}} \qquad (33)$$

We get the result of the Corollary immediately using (31) and Lemma 1.

Analysis of Approach 3 For further analysis, we present the main lemma of this subsection

Lemma 4 (see [1]).   *We denote*

$$y_f^*(x) = \operatorname*{argmax}_{y \in \mathbb{R}^{n_y}} f(x, y), \ \ x_f^*(y) = \operatorname*{argmin}_{x \in \mathcal{X}} f(x, y).$$

Under assumption 1(s), 2 we have

- *Function $x_f^*(y)$ is $\frac{L_{xy}}{\mu_x}$ -Lipschitz continuous w.r.t. the norm $\|\cdot\|_2$.*
- *Function $y_f^*(x)$ is $\frac{L_{xy}}{\mu_y}$ -Lipschitz continuous w.r.t. the norm $\|\cdot\|_2$*
- *Function $g(x)$ (see (16)) is $L_g := \left(L_{xx} + \frac{2L_{xy}^2}{\mu_y}\right)$-smooth w.r.t. the norm $\|\cdot\|_2$.*
- *Let $\widetilde{y}_\delta(x)$ be a (δ, σ)-solution to problem $\max\limits_{y \in \mathbb{R}^{n_y}} f(x, y)$. Then, for any $x', x \in \mathcal{X}$, with probability at least $1 - \sigma$ we have:*

$$\frac{\mu_x}{2}\|x - x'\|_2^2 \leq g(x') - f(x, \widetilde{y}_\delta(x)) - \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x \rangle \leq \frac{2L_g}{2}\|x' - x\|_2^2 + 2\delta.$$

- *We define*
$$g(x) = \max_{y \in \mathbb{R}^{n_y}} f(x, y), \ \ h(y) = \min_{x \in \mathcal{X}} f(x, y)$$

Let $\hat{x}$ be $(\varepsilon_x, \sigma_x)$-solution of problem $\min\limits_{x \in \mathcal{X}} g(x)$, let $\widetilde{y}_{\varepsilon_y}(\hat{x})$ be $(\varepsilon_y, \sigma_y)$-solution of problem $\max\limits_{y \in \mathbb{R}^{n_y}} f(\hat{x}, y)$. Then $\widetilde{y}_{\varepsilon_y}(\hat{x})$ is $(\tilde{\varepsilon}, 1 - \sigma_x - \sigma_y)$-solution to problem (15), where

$$\tilde{\varepsilon} = \left(\frac{L_{yy}}{\mu_y} + \frac{2L_{xy}^2}{\mu_x \mu_y}\right)\varepsilon_y + \left(\frac{L_{xy}^2 L_{yy}}{\mu_x \mu_y^2} + \frac{2L_{xy}^4}{\mu_x^2 \mu_y^2}\right)\varepsilon_x.$$

Now we are ready to present the main result of this section

Theorem 7. *Let $\varepsilon > 0$ be the target accuracy of the solution to the problem (15) and $\sigma \in (0, 1)$ be the target confidence level. Let the auxiliary problems (16), (17) be solved with accuracies*

$$\varepsilon_x = \widetilde{O}\left(\varepsilon\left(\frac{L_{xy}^2 L_{yy}}{\mu_x(\mu_y + L_{yy})^2} + \frac{2L_{xy}^4}{\mu_x^2(\mu_y + L_{yy})^2}\right)^{-1}\right);$$

$$\varepsilon_y = \widetilde{O}\left(\varepsilon\left(\frac{L_{yy}}{\mu_y + L_{yy}} + \frac{2L_{xy}^2}{\mu_x(\mu_y + L_{yy})}\right)^{-1}\left(\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x(\mu_y + L_{yy})}\right)^{-1/2}\right)$$

and confidence levels

$$\sigma_x = \widetilde{O}\left(\sigma\sqrt{\frac{\mu_y}{L_{yy}}}\right);$$

$$\sigma_y = \widetilde{O}\left(\sigma\left(\frac{L_{xx}L_{yy}}{\mu_x\mu_y} + \frac{2L_{xy}^2}{\mu_x\mu_y}\right)^{-1/2}\right),$$

that is, a $(\varepsilon_x, \sigma_x)-$solution to the problem (17) and a $(\varepsilon_y, \sigma_y)-$solution to the problem (16) are found (see Definition 5). Then, under assumptions 1(s), 2, the proposed Approach 3 guarantees to find an (ε, σ)-solution to the problem (15). Moreover, the required number of calls to the first-order oracle $\nabla_x f(x, y)$ and the zeroth-order oracle $f(x, y)$ satisfy the following bounds

$$\text{Total Number of Calls for } \nabla_x f(x, y) \text{ is } \widetilde{O}\left(\sqrt{\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x\mu_y}}\right),$$

$$\text{Total Number of Calls for } f(x, y) \text{ is } \widetilde{O}\left(n_y\sqrt{\frac{L_{xx}L_{yy}}{\mu_x\mu_y} + \frac{2L_{xy}^2}{\mu_x\mu_y}}\right).$$

References

1. Alkousa, M., Dvinskikh, D., Stonyakin, F., Gasnikov, A., Kovalev, D.: Accelerated methods for composite non-bilinear saddle point problem (2020)
2. Beznosikov, A., Sadiev, A., Gasnikov, A.: Gradient-free methods for saddle-point problem. arXiv preprint arXiv:2005.05913 (2020)
3. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision **40**(1), 120–145 (2011)
4. Chen, C.P., Lin, L.: Inequalities for the volume of the unit ball in $\mathbb{R}^n$ r n. Mediterranean Journal of Mathematics **11**(2), 299–314 (2014)
5. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. Society for Industrial and Applied Mathematics (2009). https://doi.org/10.1137/1.9780898718768

6. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods with inexact oracle: the strongly convex case (2013)
7. Dvurechensky, P., Gorbunov, E., Gasnikov, A.: An accelerated directional derivative method for smooth stochastic convex optimization. European Journal of Operational Research **290**(2), 601–621 (Apr 2021). https://doi.org/10.1016/j.ejor.2020.08.027, http://dx.doi.org/10.1016/j.ejor.2020.08.027
8. Fu, M.C., et al.: Handbook of simulation optimization, vol. 216. Springer (2015)
9. Gasnikov, A., Dvurechensky, P., Kamzolov, D., Nesterov, Y., Spokoiny, V., Stetsyuk, P., Suvorikova, A., Chernov, A.: Universal method with inexact oracle and its applications for searching equillibriums in multistage transport problems. arXiv preprint arXiv:1506.00292 (2015)
10. Gasnikov, A., Dvurechensky, P., Usmanova, I.: About accelerated randomized methods. arXiv preprint arXiv:1508.02182 (2015)
11. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014)
12. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014)
13. Ivanova, A., Pasechnyuk, D., Grishchenko, D., Shulgin, E., Gasnikov, A., Matyukhin, V.: Adaptive catalyst for smooth convex optimization (2020)
14. Korpelevich, G.M.: The extragradient method for finding saddle points and other problems (1976)
15. Lin, H., Mairal, J., Harchaoui, Z.: A universal catalyst for first-order optimization. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper/2015/file/c164bbc9d6c72a52c599bbb43d8db8e1-Paper.pdf
16. Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., O'Reilly, U.M.: Min-max optimization without gradients: Convergence and applications to adversarial ml (2019)
17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings (2018)
18. Narodytska, N., Kasiviswanathan, S.P.: Simple black-box adversarial attacks on deep neural networks. In: CVPR Workshops. pp. 1310–1318. IEEE Computer Society (2017), http://doi.ieeecomputersociety.org/10.1109/CVPRW.2017.172
19. Nedić, A., Ozdaglar, A.: Subgradient methods for saddle-point problems. Journal of Optimization Theory and Applications **142**(1), 205–228 (Jul 2009)
20. Nemirovski, A.: Prox-method with rate of convergence o (1/ t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM Journal on Optimization **15**, 229–251 (01 2004). https://doi.org/10.1137/S1052623403425629
21. Nesterov, Y.: Lectures on convex optimization, vol. 137. Springer International Publishing (2018)
22. Nesterov, Y., Stich, S.U.: Efficiency of the accelerated coordinate descent method on structured optimization problems. SIAM Journal on Optimization **27**(1), 110–123 (2017)
23. Pinto, L., Davidson, J., Sukthankar, R., Gupta, A.: Robust adversarial reinforcement learning. Proceedings of Machine Learning Research, vol. 70, pp. 2817–2826. PMLR, International Convention Centre, Sydney, Australia (06–11 Aug 2017), http://proceedings.mlr.press/v70/pinto17a.html

24. Polyak, B.T.: Introduction to optimization. Inc., Publications Division, New York (1987)
25. Renegar, J.: A polynomial-time algorithm, based on newton's method, for linear programming. Mathematical programming **40**(1), 59–93 (1988)
26. Sadiev, A., Beznosikov, A., Dvurechensky, P., Gasnikov, A.: Zeroth-order algorithms for smooth saddle-point problems. arXiv:2009.09908 (2020)
27. Shashaani, S., Hashemi, F.S., Pasupathy, R.: Astro-df: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. SIAM Journal on Optimization **28**(4), 3145–3176 (2018). https://doi.org/10.1137/15M1042425, https://doi.org/10.1137/15M1042425
28. Stonyakin, F., Tyurin, A., Gasnikov, A., Dvurechensky, P., Agafonov, A., Dvinskikh, D., Pasechnyuk, D., Artamonov, S., Piskunova, V.: Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model (2020)
29. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses (2017)
30. Vaidya, P.M.: A new algorithm for minimizing convex functions over convex sets. In: 30th Annual Symposium on Foundations of Computer Science. pp. 338–343. IEEE Computer Society (1989)
31. Vaidya, P.M.: A new algorithm for minimizing convex functions over convex sets. Mathematical programming **73**(3), 291–341 (1996)
32. Wang, Z., Balasubramanian, K., Ma, S., Razaviyayn, M.: Zeroth-order algorithms for nonconvex minimax problems with improved complexities (2020)

A   Proofs for Section 2 (Solving Min-Min Problems)

Proposition 1. *The volumetric center for P_0 is $\boldsymbol{\omega} = \omega \mathbf{1}_n$, where $\omega := \frac{n-1}{n+1}\mathcal{R}$ and $\mathbf{1}_n$ denotes the vector $(1, \ldots, 1)^\top \in \mathbb{R}^n$.*

Proof. P_0 can be viewed as $P_0 = \{x \in \mathbb{R}^n : A_0 x \geq b_0\}$, where $A_0 \in \mathbb{R}^{(n+1)\times n}$ and $b_0 \in \mathbb{R}^{n+1}$ are defined as

$$b_0 = -\mathcal{R} \begin{bmatrix} \mathbf{1}_n \\ n \end{bmatrix} \text{ and } A_0 = \begin{bmatrix} e_1^\top \\ \vdots \\ e_n^\top \\ -\mathbf{1}_n^\top \end{bmatrix} \tag{34}$$

It is known that the analytic center of simplex (i.e. minimizer of $L(x)$) is also its volumetric center [25]. Note that $L(x)$ is a convex function.

$$\nabla L(x) = -\sum_{i=1}^{n+1} \frac{a_i}{a_i^\top x - b_i} = -\sum_{i=1}^{n} \frac{e_i}{x_i + \mathcal{R}} + \frac{\mathbf{1}_n}{n\mathcal{R} - \sum_{j=1}^{n} x_j}$$

$$\nabla L(\bar{x}) = 0 \iff \frac{1}{n\mathcal{R} - \sum_{j=1}^{n} \bar{x}_j} = \frac{1}{\bar{x}_i + \mathcal{R}}, i = \overline{1, n}$$

$$\bar{x}_i = \frac{n-1}{n+1}\mathcal{R}, \ i = \overline{1, n}, \text{ and } \bar{x} \in P_0 \Rightarrow \boldsymbol{\omega} = \bar{x}.$$

Theorem 2. *Let $\mathcal{B}_\rho$ and $\mathcal{B}_\mathcal{R}$ be some Euclidean balls of radii ρ and $\mathcal{R}$, respectively, such that $\mathcal{B}_\rho \subseteq \mathcal{X} \subseteq \mathcal{B}_\mathcal{R}$, and let a number $B > 0$ be such that $|g(x) - g(x')| \leq B \ \forall x, x' \in \mathcal{X}$. After $N \geq \frac{2n}{\gamma} \ln\left(\frac{n^{1.5}\mathcal{R}}{\gamma\rho}\right) + \frac{1}{\gamma} \ln \pi$ iterations Vaidya's method with δ-subgradient for the problem* (4) *returns a point x^N such that*

$$g(x^N) - g(x_*) \leq \frac{Bn^{1.5}\mathcal{R}}{\gamma\rho} \exp\left(\frac{\ln \pi - \gamma N}{2n}\right) + \delta, \tag{8}$$

where $\gamma > 0$ is the parameter of the algorithm and x_* is a solution of the problem (4).

Proof. The proof goes as follows. First, we consider the set $\mathcal{X}^\varepsilon := \{(1-\varepsilon)x_* + \varepsilon x, \ x \in \mathcal{X}\}$ for some $\varepsilon \in [0, 1]$. Then, we calculate the number of iterations N sufficient for the volume of the polytope P_N to go below the volume of $\mathcal{X}^\varepsilon$. Finally, we show that after N iterations the algorithm arrives at a $(B\varepsilon + \delta)$-suboptimal point.

Fix $\varepsilon \in [0, 1]$ and consider the set $\mathcal{X}^\varepsilon := \{(1-\varepsilon)x_* + \varepsilon x, x \in \mathcal{X}\}$. Note that $\mathcal{X}^\varepsilon \subseteq P_0$ and

$$\text{vol}\,\mathcal{X}^\varepsilon = \varepsilon^n \,\text{vol}\,\mathcal{X} \geq \varepsilon^n \,\text{vol}\,\mathcal{B}_\rho. \tag{35}$$

The lower bound on the volume of n-ball [4] is

$$\text{vol}\,\mathcal{B}_\rho \geq \frac{\rho^n}{\sqrt{\pi \left(n + \frac{e}{2} - 1\right)}} \left(\frac{2\pi e}{n}\right)^{n/2}. \tag{36}$$

We will now derive an upper bound on the volume of polytope P_N at the beginning of the N-th iteration. According to the Vaidya's paper [30], the following inequality holds:

$$\ln\left(\text{vol}\, P_N\right) \le n\ln\left(\frac{2n}{\gamma}\right) - \rho^0 - \frac{\gamma}{2}N, \tag{37}$$

where γ is the parameter of the algorithm and ρ^0 denotes the minimum of the volumetric barrier at the beginning of the 0-th iteration. To calculate it, we will use proposition 1, which states that $\boldsymbol{\omega} = \omega\mathbf{1}_n$, where $\omega := \frac{n-1}{n+1}\mathcal{R}$. Observe that

$$\mathcal{R} - \omega = \frac{2}{n+1}\mathcal{R}, \ \omega + \mathcal{R} = \frac{2n}{n+1}\mathcal{R} = n(\mathcal{R} - \omega). \tag{38}$$

Let us calculate $H(\boldsymbol{\omega})$ and $F(\boldsymbol{\omega})$:

$$H(\boldsymbol{\omega}) \overset{(5)}{=} \frac{1}{(\omega + \mathcal{R})^2}I_n + \frac{1}{(n\mathcal{R} - n\omega)^2}\mathbf{1}_n\mathbf{1}_n^\top \overset{(38)}{=} \frac{1}{(\omega + \mathcal{R})^2}\left(I_n + \mathbf{1}_n\mathbf{1}_n^\top\right)$$

$$\det H(\boldsymbol{\omega}) = \frac{1}{(\omega + \mathcal{R})^{2n}}\left(1 + n\right) = \frac{(n+1)^{2n+1}}{(2n\mathcal{R})^{2n}}.$$

$$\rho_0 = F(\boldsymbol{\omega}) = \left(n + \frac{1}{2}\right)\ln(n+1) - n\ln(2n\mathcal{R}) \tag{39}$$

Using the formulas (37) and (39), we obtain

$$\ln\left(\text{vol}\, P_N\right) \le n\ln\left(\frac{4n^2\mathcal{R}}{\gamma(n+1)}\right) - \frac{1}{2}\ln(n+1) - \frac{\gamma}{2}N. \tag{40}$$

Putting together (35), (36) and (40), we will determine a sufficient number of iterations to assure $\text{vol}\, P_N < \text{vol}\, \mathcal{X}^\varepsilon$:

$$n\ln\left(\frac{4n^2\mathcal{R}}{\gamma(n+1)}\right) - \frac{1}{2}\ln(n+1) - \frac{\gamma}{2}N < n\ln\left(\sqrt{\frac{2\pi e}{n}}\varepsilon\rho\right) - \frac{1}{2}\ln\left(\pi\left(n + \frac{e}{2} - 1\right)\right) \tag{41}$$

$$\gamma N > 2n\ln\left(\frac{4n^{2.5}\mathcal{R}}{\gamma(n+1)\sqrt{2\pi e}\varepsilon\rho}\right) + \ln\left(\frac{\pi\left(n + \frac{e}{2} - 1\right)}{n+1}\right). \tag{42}$$

It is sufficient to do

$$N = \frac{2n}{\gamma}\ln\left(\frac{n^{1.5}\mathcal{R}}{\gamma\varepsilon\rho}\right) + \frac{1}{\gamma}\ln\pi \tag{43}$$

iterations. Since $\text{vol}\, P_N < \text{vol}\, \mathcal{X}^\varepsilon$, there will be an iteration $j \in \{0, ..., N-1\}$ and a point $x_\varepsilon \in \mathcal{X}^\varepsilon$ such that $x_\varepsilon \in P_j$ and $x_\varepsilon \notin P_{j+1}$. Thus, on the j-th iteration we added a new constraint to the polytope. There exists a point $x \in \mathcal{X}$ such that $x_\varepsilon = (1 - \varepsilon)x_* + \varepsilon x$, therefore it follows from the convexity of g that

$$g(x_\varepsilon) \leqslant (1 - \varepsilon)g(x_*) + \varepsilon g(x) \leqslant (1 - \varepsilon)g(x_*) + \varepsilon\left((g(x_*) + B\right) =$$
$$= g(x_*) + B\varepsilon. \tag{44}$$

We will now show that the following inequality holds for the query point z_j at the j-th iteration:

$$g(z_j) < g(x_\varepsilon) + \delta. \tag{45}$$

First, note that z_j belongs to $\mathcal{X}$, otherwise we would cut out a part of P_j that doesn't intersect with $\mathcal{X}$, which is a contradiction since $x_\varepsilon \in P_j \setminus P_{j+1}$ and $x_\varepsilon \in \mathcal{X}$. Recall that the constraint $c_j^\top x \geq \beta_j$ was added to the polytope on the j-th iteration, where $-c_j \in \partial_\delta g(z_j)$ and $\beta_j \leq c_j^\top z_j$. According to the definition of δ-subgradient,

$$g(x) \geq g(z_j) - c_j^\top (x - z_j) - \delta \quad \forall x \in \mathcal{X}.$$

Therefore,

$$(P_j \setminus P_{j+1}) \cap \mathcal{X} \subseteq \{x \in \mathcal{X} : c_j^\top (x - z_j) < 0\} \subseteq \{x \in \mathcal{X} : g(x) > g(z_j) - \delta\},$$

which results in (45) since $x_\varepsilon \in (P_j \setminus P_{j+1}) \cap \mathcal{X}$. Combining (43), (44) and (45), we conclude the proof.

Theorem 4. *Let $\tilde\delta > 0$ and $\tilde y \in \mathbb{R}^{n_y}$ satisfy $f(x', \tilde y) - g(x') \leq \tilde\delta$, then $\partial_x f(x', \tilde y) \in \partial_\delta g(x')$ with*

$$\delta = 2 \left(\sqrt{\tilde\delta} + 2\sqrt{D} \right) \sqrt{\frac{L_{yy}\tilde\delta}{\mu_y}}, \tag{12}$$

where $D := \max\limits_{x \in \mathcal{X}} (f(x, \mathbf{0}) - g(x)) < +\infty$.

Proof. Using Lemmas 1, 2 and triangle inequality, we get $\partial_x f(x', \tilde y) \in \partial_\delta g(x')$ with

$$\delta = \left(\|\tilde y\|_2 + \max_{x \in \mathcal{X}} \|y(x)\|_2 \right) \sqrt{2L_{yy}\tilde\delta}. \tag{46}$$

Let us bound the two terms in parentheses. For any $x \in \mathcal{X}$, it follows from the strong convexity of $f(x, \cdot)$ that

$$\|y' - y(x)\|_2 \leq \sqrt{\frac{2(f(x, y') - g(x))}{\mu_y}} \quad \forall y' \in \mathbb{R}^{n_y}. \tag{47}$$

In particular, putting $y' = \mathbf{0}$ leads to

$$\|y(x)\|_2 \leq \sqrt{\frac{2(f(x, \mathbf{0}) - g(x))}{\mu_y}}.$$

Observe that $\Delta(x) := f(x, \mathbf{0}) - g(x)$ is a continuous function defined on the compact set $\mathcal{X}$. Therefore, it is bounded, i.e. $\Delta(x) \leq D$ for some $D < +\infty$. Thus,

$$\max_{x \in \mathcal{X}} \|y(x)\|_2 \leq \sqrt{\frac{2D}{\mu_y}}. \tag{48}$$

Finally, let us bound $\|\tilde{y}\|_2$ by again using triangle inequality, then inequalities (47), (48) and $f(x', \tilde{y}) - g(x') \leq \tilde{\delta}$:

$$\|\tilde{y}\|_2 \leq \|\tilde{y} - y(x')\|_2 + \|y(x')\|_2 \leq \|\tilde{y} - y(x')\|_2 + \sqrt{\frac{2D}{\mu_y}} \leq \sqrt{\frac{2\tilde{\delta}}{\mu_y}} + \sqrt{\frac{2D}{\mu_y}}. \quad (49)$$

The statement of the theorem immediately follows from (46), (48) and (49).

Lemma 1. *Let $h : \mathbb{R}^{n_y} \to \mathbb{R}$ be an L-smooth convex function, and let the point $\tilde{y} \in \mathbb{R}^{n_y}$ satisfy $h(\tilde{y}) - h(y_*) \leq \tilde{\delta}$ for some delta > 0, where $y_* \in \underset{y \in \mathbb{R}^{n_y}}{\mathrm{Argmin}}\ h(y)$.*

Then

$$\langle \nabla h(\tilde{y}), \tilde{y} - y \rangle \leq \|\tilde{y} - y\|_2 \sqrt{2L\tilde{\delta}} \quad \forall y \in \mathbb{R}^{n_y}.$$

Proof. Due to the Cauchy–Bunyakovsky–Schwarz inequality,

$$\langle \nabla h(\tilde{y}), \tilde{y} - y \rangle \leq \|\nabla h(\tilde{y})\|_2 \|\tilde{y} - y\|_2 \quad (50)$$

Since h is L-smooth and convex,

$$h(\tilde{y}) \geq h(y_*) + \langle \nabla h(y_*), \tilde{y} - y_* \rangle + \frac{1}{2L}\|\nabla h(y_*) - \nabla h(\tilde{y})\|_2^2$$

Using $\nabla h(y_*) = 0$, we obtain

$$\|\nabla h(\tilde{y})\|_2^2 \leq 2L\left(h(\tilde{y}) - h(y_*)\right). \quad (51)$$

Combining (50) and (51), we conclude the proof.

Lemma 2 (see [9], p.12). *Let $\tilde{y} \in \mathbb{R}^{n_y}$ satisfy*

$$\langle \nabla_y f(x, \tilde{y}), \tilde{y} - y(x') \rangle \leq \delta \quad \forall x' \in \mathcal{X}, \quad (13)$$

then $\partial_x f(x, \tilde{y}) \in \partial_\delta g(x)$.

Proof. From convexity of f we get for any $x' \in \mathcal{X}$

$$f(x', y(x')) \geq f(x, \tilde{y}) + \langle \nabla_x f(x, \tilde{y}), x' - x \rangle + \langle \nabla_y f(x, \tilde{y}), y(x') - \tilde{y} \rangle$$

Using (13) and $g(x') = f(x', y(x'))$, we obtain

$$g(x') \geq f(x, \tilde{y}) + \langle \nabla_x f(x, \tilde{y}), x' - x \rangle - \delta.$$

Note that $g(x) \leq f(x, \tilde{y})$, therefore,

$$g(x') \geq g(x) + \langle \nabla_x f(x, \tilde{y}), x' - x \rangle - \delta \quad \forall x' \in \mathcal{X}.$$

Thus, by definition we have $\partial_x f(x, \tilde{y}) \in \partial_\delta g(x)$.

B Proofs for Section 3 (Solving Min-Max Saddle-Point Problems)

Theorem 6. *Let function f be convex on convex set $\mathcal{X}$ and be equipped with a first-order (δ, σ, L, μ)-oracle. Then, after N iterations of Algorithm 4 applied to problem (28), we have that with probability at least $(1 - N\sigma)$:*

$$f(x_N) - f(x^*) \leq 2L \exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu}{L}}\right) R_2^2 + \frac{2\sum_{k=0}^{N-1} A_{k+1}\delta_k}{A_N}, \quad (31)$$

where R_2 is such that $\frac{1}{2}\|x_0 - x^*\|_2^2 \leq R_2^2$ and x_0 is the starting point.

Proof. We almost completely repeat the proof of result from [28].

Lemma 2 (see [28]). *Let $\phi(x)$ be convex function on the convex set $\mathcal{X}$ and*

$$y = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \phi(x) + \frac{\beta}{2}\|x - z\|_2^2 + \frac{\gamma}{2}\|x - u\|_2^2 \right\},$$

where $\beta \geq 0$, $\gamma \geq 0$. Then

$$\psi(x) + \frac{\beta}{2}\|x-z\|_2^2 + \frac{\gamma}{2}\|x-u\|_2^2 \geq \psi(y) + \frac{\beta}{2}\|y-z\|_2^2 + \frac{\gamma}{2}\|y-u\|_2^2 + \frac{\gamma+\beta}{2}\|y-x\|_2^2 \quad (52)$$

Lemma 3 (see [28]). *For all $x \in \mathcal{X}$ we have with probability with $1 - \sigma$*

$$A_{k+1}f(x_{k+1}) - A_k f(x_k) + \frac{(1 + A_{k+1}\mu)}{2}\|x - u_{k+1}\|_2^2 - \frac{(1 + A_k\mu)}{2}\|x - u_k\|_2^2$$
$$\leq \alpha_{k+1}f(x) + 2\delta_{k+1}A_{k+1} \quad (53)$$

Proof Using definition of (δ, σ, L, μ)-oracle (29), we have with probability with $1 - \sigma$ inequalities $f(x) - \delta \leq f_{\delta, L, \mu}(x) \leq f(x)$. Then with probability at least $1 - \sigma$:

$$f(x_{k+1}) \leq f_{\delta_k, L, \mu}(y_{k+1}) + \langle g_{\delta, L, \mu}(y_{k+1}), x_{k+1} - y_{k+1} \rangle + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|_2^2 + 2\delta_k,$$

Using definitions of sequences x_{k+1} and y_{k+1}, we have with probability at least $1 - \sigma$:

$$f(x_{k+1}) \le f_{\delta_k, L, \mu}(y_{k+1}) + \langle g_{\delta, L, \mu}(y_{k+1}), x_{k+1} - y_{k+1} \rangle + \frac{L_{k+1}}{2} \|x' - x\|_2^2 + 2\delta_k$$

$$= f_{\delta_k, L, \mu}(y_{k+1}) + \left\langle g_{\delta, L, \mu}(y_{k+1}), \frac{\alpha_{k+1} u_{k+1} + A_k x_k}{A_{k+1}} - y_{k+1} \right\rangle$$

$$+ \frac{L_{k+1}}{2} \left\| \frac{\alpha_{k+1} u_{k+1} + A_k x_k}{A_{k+1}} - \frac{\alpha_{k+1} u_k + A_k x_k}{A_{k+1}} \right\|_2^2 + 2\delta_k$$

$$= \frac{A_k}{A_{k+1}} \left(f_{\delta_k, L, \mu}(y_{k+1}) + \langle g_{\delta, L, \mu}(y_{k+1}), x_k - y_{k+1} \rangle \right)$$

$$+ \frac{\alpha_{k+1}}{A_{k+1}} \left(f_{\delta_k, L, \mu}(y_{k+1}) + \langle g_{\delta, L, \mu}(y_{k+1}), u_{k+1} - y_{k+1} \rangle \right) + \frac{L_{k+1} \alpha_{k+1}^2}{2 A_{k+1}^2} \|u_{k+1} - u_k\|_2^2 + 2\delta_k$$

$$= \frac{A_k}{A_{k+1}} \left(f_{\delta_k, L, \mu}(y_{k+1}) + \langle g_{\delta, L, \mu}(y_{k+1}), x_k - y_{k+1} \rangle \right)$$

$$+ \frac{\alpha_{k+1}}{A_{k+1}} \left(f_{\delta_k, L, \mu}(y_{k+1}) + \langle g_{\delta, L, \mu}(y_{k+1}), u_{k+1} - y_{k+1} \rangle + \frac{1 + A_k \mu}{2 \alpha_{k+1}} \|u_{k+1} - u_k\|_2^2 \right) + 2\delta_k$$

$$\le \frac{A_k}{A_{k+1}} f_{\delta_k, L, \mu}(x_k)$$

$$+ \frac{\alpha_{k+1}}{A_{k+1}} \left(f_{\delta_k, L, \mu}(y_{k+1}) + \langle g_{\delta, L, \mu}(y_{k+1}), u_{k+1} - y_{k+1} \rangle + \frac{1 + A_k \mu}{2 \alpha_{k+1}} \|u_{k+1} - u_k\|_2^2 \right) + 2\delta_k$$

Due to u_{k+1} is the solution to $\underset{x \in \mathcal{X}}{\operatorname{argmin}} \, \phi_{k+1}(x)$, using (52) we have for any $x \in \mathcal{X}$ with probability at least $1 - \sigma$:

$$\alpha_{k+1} \langle g_{\delta, L, \mu}(y_{k+1}), u_{k+1} - y_{k+1} \rangle + \frac{(1 + A_k \mu)}{2} \|u_{k+1} - u_k\|_2^2$$

$$+ \frac{\alpha_{k+1} \mu}{2} \|u_{k+1} - y_{k+1}\|_2^2 + \frac{(1 + A_{k+1} \mu)}{2} \|u_{k+1} - x\|_2^2$$

$$\le \alpha_{k+1} \langle g_{\delta, L, \mu}(y_{k+1}), x - y_{k+1} \rangle + \frac{(1 + A_k \mu)}{2} \|x - u_k\|_2^2 + \frac{\alpha_{k+1} \mu}{2} \|x - y_{k+1}\|_2^2$$

Then for all $x \in \mathcal{X}$ we have with probability at least $1 - \sigma$:

$$\alpha_{k+1} \langle g_{\delta, L, \mu}(y_{k+1}), u_{k+1} - y_{k+1} \rangle + \frac{(1 + A_k \mu)}{2} \|u_{k+1} - u_k\|_2^2$$

$$\le \alpha_{k+1} \langle g_{\delta, L, \mu}(y_{k+1}), x - y_{k+1} \rangle + \frac{(1 + A_k \mu)}{2} \|x - u_k\|_2^2 - \frac{(1 + A_{k+1} \mu)}{2} \|u_{k+1} - x\|_2^2$$

$$+ \frac{\alpha_{k+1} \mu}{2} \|x - y_{k+1}\|_2^2$$

Then for all $x \in \mathcal{X}$ we have with probability at least $1 - \sigma$:

$$f(x_{k+1}) \le \frac{A_k}{A_{k+1}} f(x_k) + \frac{\alpha_{k+1}}{A_{k+1}} f(x) + 2\delta_k + \frac{\alpha_{k+1}}{A_{k+1}} \left(\frac{1 + A_{k+1} \mu}{2 \alpha_{k+1}} \|u_{k+1} - x\|_2^2 + \frac{1 + A_k \mu}{2 \alpha_{k+1}} \|x - u_k\|_2^2 \right)$$

$$= \frac{A_k}{A_{k+1}} f(x_k) + \frac{\alpha_{k+1}}{A_{k+1}} f(x) + 2\delta_k + \frac{1 + A_{k+1} \mu}{2 A_{k+1}} \|u_{k+1} - x\|_2^2 + \frac{1 + A_k \mu}{2 A_{k+1}} \|x - u_k\|_2^2$$

Now we prove the statement of this theorem. First, we note that if we sum the inequality (53) by k from 0 to $N-1$, for any $x \in \mathcal{X}$ we get

$$A_N f(x_N) - A_N f(x^*) \leq \sum_{k=0}^{N-1} 2\delta_{k+1} A_{k+1} - \frac{1 + A_N \mu}{2} \|x - u_N\|_2^2 + \frac{1}{2} \|x - u_0\|_2^2$$

Then using the resulting inequality above for any $x \in \mathcal{X}$ we get

$$\mathbb{P}\left\{ A_N f(x_N) - A_N f(x^*) > \sum_{k=0}^{N-1} 2\delta_{k+1} A_{k+1} - \frac{1 + A_N \mu}{2} \|x - u_N\|_2^2 + \frac{1}{2} \|x - u_0\|_2^2 \right\}$$

$$\leq \sum_{k=0}^{N-1} \mathbb{P}\left\{ A_{k+1} f(x_{k+1}) - A_k f(x_k) + \frac{(1 + A_{k+1}\mu)}{2} \|x - u_{k+1}\|_2^2 \right.$$

$$\left. - \frac{(1 + A_k \mu)}{2} \|x - u_k\|_2^2 \leq \alpha_{k+1} f(x) + 2\delta_{k+1} A_{k+1} \right\} < N\sigma$$

In this way, for any $x \in \mathcal{X}$ with probability at least $(1 - N\sigma)$ we have

$$A_N f(x_N) - A_N f(x^*) \leq \sum_{k=0}^{N-1} 2\delta_{k+1} A_{k+1} + \frac{1}{2} \|x - u_0\|_2^2$$

We finish proof with next inequality (see [28]):

$$A_N \geq \frac{1}{2L} \exp\left(\frac{N-1}{2} \sqrt{\frac{\mu}{L}} \right)$$

Lemma 4 (see [1]). *We denote*

$$y_f^*(x) = \operatorname*{argmax}_{y \in \mathbb{R}^{n_y}} f(x, y), \quad x_f^*(y) = \operatorname*{argmin}_{x \in \mathcal{X}} f(x, y).$$

Under assumption 1(s), 2 we have

- *Function $x_f^*(y)$ is $\frac{L_{xy}}{\mu_x}$-Lipschitz continuous w.r.t. the norm $\|\cdot\|_2$.*
- *Function $y_f^*(x)$ is $\frac{L_{xy}}{\mu_y}$-Lipschitz continuous w.r.t. the norm $\|\cdot\|_2$*
- *Function $g(x)$ (see (16)) is $L_g := \left(L_{xx} + \frac{2L_{xy}^2}{\mu_y} \right)$-smooth w.r.t. the norm $\|\cdot\|_2$.*
- *Let $\widetilde{y}_\delta(x)$ be a (δ, σ)-solution to problem $\max_{y \in \mathbb{R}^{n_y}} f(x, y)$. Then, for any $x', x \in \mathcal{X}$, with probability at least $1 - \sigma$ we have:*

$$\frac{\mu_x}{2} \|x - x'\|_2^2 \leq g(x') - f(x, \widetilde{y}_\delta(x)) - \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x \rangle \leq \frac{2L_g}{2} \|x' - x\|_2^2 + 2\delta.$$

- *We define*
$$g(x) = \max_{y \in \mathbb{R}^{n_y}} f(x, y), \quad h(y) = \min_{x \in \mathcal{X}} f(x, y)$$

Let $\hat{x}$ be $(\varepsilon_x, \sigma_x)$-solution of problem $\min\limits_{x \in \mathcal{X}} g(x)$, let $\widetilde{y}_{\varepsilon_y}(\hat{x})$ be $(\varepsilon_y, \sigma_y)$-solution of problem $\max\limits_{y \in \mathbb{R}^{n_y}} f(\hat{x}, y)$. Then $\widetilde{y}_{\varepsilon_y}(\hat{x})$ is $(\tilde{\varepsilon}, 1 - \sigma_x - \sigma_y)$-solution to problem (15), where

$$\tilde{\varepsilon} = \left(\frac{L_{yy}}{\mu_y} + \frac{2L_{xy}^2}{\mu_x \mu_y} \right) \varepsilon_y + \left(\frac{L_{xy}^2 L_{yy}}{\mu_x \mu_y^2} + \frac{2L_{xy}^4}{\mu_x^2 \mu_y^2} \right) \varepsilon_x.$$

Proof. To justify the statement of the lemma, we give a more general statement.

Lemma 4 (see lemma 4). *We denote*

$$y_f^*(x) = \operatorname*{argmax}_{y \in \mathbb{R}^{n_y}} g(x, y); \quad x_f^*(y) = \operatorname*{argmin}_{x \in \mathcal{X}} g(x, y).$$

Under assumption 1(s) for $\| \cdot \|_{p_x}$ and $\| \cdot \|_{p_y}$ norms, assumption 2 for $\| \cdot \|_{p_x}$ and $\| \cdot \|_{p_y}$ norms we have

- *Function $x_g^*(y)$ is $\frac{L_{xy}}{\mu_x}$-Lipschitz continuous w.r.t. the norm $\| \cdot \|_{p_x}$.*
- *Function $y_g^*(x)$ is $\frac{L_{xy}}{\mu_y}$-Lipschitz continuous w.r.t. the norm $\| \cdot \|_{p_x}$*
- *Function $g(x)$ is $L_g = \left(L_{xx} + \frac{2L_{xy}^2}{\mu_y} \right)$-smooth w.r.t. the norm $\| \cdot \|_{p_x}$.*
- *Let $\widetilde{y}_\delta(x)$ - (δ, σ) be solution to problem $\max\limits_{y \in \mathbb{R}^{n_y}} f(x, y)$. Then for any $x', x \in \mathcal{X}$ with probability at least $1 - \sigma$ we have:*

$$\frac{\mu_x}{2} \|x - x'\|_{p_x}^2 \le g(x') - f(x, \widetilde{y}_\delta(x)) - \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x \rangle \le \frac{2L_g}{2} \|x' - x\|_{p_x}^2 + 2\delta$$

- *We define*

$$g(x) = \max\limits_{y \in \mathbb{R}^{n_y}} f(x, y), \quad h(y) = \min\limits_{x \in \mathcal{X}} f(x, y)$$

Let $\hat{x}$ be $(\varepsilon_x, \sigma_x)$-solution of problem $\min\limits_{x \in \mathcal{X}} g(x)$, let $\widetilde{y}_{\varepsilon_y}(\hat{x})$ be $(\varepsilon_y, \sigma_y)$-solution of problem $\max\limits_{y \in \mathbb{R}^{n_y}} f(\hat{x}, y)$. Then $\widetilde{y}_{\varepsilon_y}(\hat{x})$ is $(\tilde{\varepsilon}, 1 - \sigma_x - \sigma_y)$-solution to problem, where

$$\tilde{\varepsilon} = \left(\frac{L_{yy}}{\mu_y} + \frac{2L_{xy}^2}{\mu_x \mu_y} \right) \varepsilon_y + \left(\frac{L_{xy}^2 L_{yy}}{\mu_x \mu_y^2} + \frac{2L_{xy}^4}{\mu_x^2 \mu_y^2} \right) \varepsilon_x$$

Proof

1. Estimate $\|y^*(x) - y^*(x')\|_{p_y}$:

$$\|y^*(x) - y^*(x')\|_{p_y}^2 \leq \frac{2}{\mu_y}(f(x, y^*(x)) - f(x, y^*(x')))$$

$$\leq \frac{2}{\mu_y}(f(x, y^*(x)) - f(x, y^*(x'))) - \frac{2}{\mu_y}(f(x', y^*(x)) - f(x', y^*(x')))$$

$$= \frac{2}{\mu_y}\int_0^1 \langle \nabla_x f(x' + t(x - x'), y^*(x)) - \nabla_x f(x' + t(x - x'), y^*(x')), x' - x \rangle dt$$

$$\leq \frac{2}{\mu_y}\int_0^1 \|\nabla_x f(x' + t(x - x'), y^*(x)) - \nabla_x f(x' + t(x - x'), y^*(x'))\|_{q_x} \|x' - x\|_{p_x} dt$$

$$\leq \frac{2}{\mu_y} L_{xy} \|x' - x\|_{p_x} \|y^*(x) - y^*(x')\|_{p_y}$$

In thus way, we have

$$\|y^*(x) - y^*(x')\|_{p_y} \leq \frac{2L_{xy}}{\mu_y}\|x' - x\|_{p_x}$$

2. Estimate $\|x^*(y) - x^*(y')\|_{p_x}$. Due to μ_x-strong convexity function $f(x, y)$:

$$\|x^*(y) - x^*(y')\|_{p_x}^2 \leq \frac{2}{\mu_x}(f(x^*(y), y') - f(x^*(y'), y'))$$

$$\leq \frac{2}{\mu_x}(f(x^*(y), y') - f(x^*(y'), y')) - \frac{2}{\mu_x}(f(x^*(y), y) - f(x^*(y'), y))$$

$$= \frac{2}{\mu_x}\int_0^1 \langle \nabla_y f(x^*(y), y + t(y' - y)) - \nabla_y f(x^*(y'), y + t(y' - y)), y' - y \rangle dt$$

$$\leq \frac{2}{\mu_x}\int_0^1 \|\nabla_y f(x^*(y), y + t(y' - y)) - \nabla_y f(x^*(y'), y + t(y' - y))\|_{q_y} \|y' - y\|_{p_y} dt$$

$$\leq \frac{2}{\mu_x} L_{xy} \|y' - y\|_{p_y} \|x^*(y) - x^*(y')\|_{p_x}$$

We have
$$\|x^*(y) - x^*(y')\|_{p_x} \leq \frac{2L_{xy}}{\mu_x}\|y' - y\|_{p_y}$$

3. Consider the following value $\|\nabla g(x) - \nabla g(x')\|_{q_x}$:

$$\|\nabla g(x) - \nabla g(x')\|_{q_x} = \|\nabla_x f(x, y^*(x)) - \nabla_x f(x', y^*(x'))\|_{q_x}$$
$$\leq \|\nabla_x f(x, y^*(x)) - \nabla_x f(x, y^*(x'))\|_{q_x}$$
$$+ \|\nabla_x f(x, y^*(x')) - \nabla_x f(x', y^*(x'))\|_{q_x}$$
$$\leq L_{xx}\|x - x'\|_{p_x} + L_{xy}\|y^*(x) - y^*(x')\|_{p_y}$$

Then we have:

$$\|\nabla g(x) - \nabla g(x')\|_{q_x} \leq \left(L_{xx} + \frac{2L_{xy}^2}{\mu_y}\right)\|x - x'\|_{p_x}$$

4. Let $\widetilde{y}_\delta(x)$ - (δ, σ) be solution to problem $\max\limits_{y \in \mathbb{R}^{n_y}} f(x, y)$, and $y^*(x)$ be exact solution. We prove the left side of inequality. For any $x, x' \in \mathcal{X}$ we have

$$g(x') = f(x', y^*(x')) \geq f(x', \widetilde{y}_\delta(x)) \geq f(x, \widetilde{y}_\delta(x)) + \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x\rangle + \frac{\mu_x}{2}\|x - x'\|_{p_x}^2$$

Then we have

$$\frac{\mu_x}{2}\|x - x'\|_{p_x}^2 \leq g(x') - f(x, \widetilde{y}_\delta(x)) - \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x\rangle$$

Now we prove the left side of inequality. For any $x, x' \in \mathcal{X}$ we have with probability at least $1 - \sigma$

$$g(x') \leq g(x) + \langle \nabla g(x), x' - x\rangle + \frac{L_g}{2}\|x - x'\|_{p_x}^2$$

$$\leq f(x, \widetilde{y}_\delta(x)) + \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x\rangle + \frac{L_g}{2}\|x - x'\|_{p_x}^2 + \delta$$

$$+ \langle \nabla_x f(x, \widetilde{y}_\delta(x)) - \nabla g(x), x - x'\rangle$$

$$\leq f(x, \widetilde{y}_\delta(x)) + \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x\rangle + \frac{L_g}{2}\|x - x'\|_{p_x}^2 + \delta$$

$$+ \|\nabla_x f(x, \widetilde{y}_\delta(x)) - \nabla g(x)\|_{q_x}\|x - x'\|_{p_x}$$

Estimate $\|\nabla_x f(x, \widetilde{y}_\delta(x)) - \nabla g(x)\|_{q_x}$. we have with probability at least $1 - \sigma$

$$\|\nabla_x f(x, \widetilde{y}_\delta(x)) - \nabla g(x)\|_{q_x}^2 = \|\nabla_x f(x, \widetilde{y}_\delta(x)) - \nabla_x f(x, y^*(x))\|_{q_x}^2 \leq L_{xy}^2\|\widetilde{y}_\delta(x) - y^*(x)\|_{p_y}^2$$

$$\leq \frac{2L_{xy}^2}{\mu_y}\left(f(x, y^*(x)) - f(x, \widetilde{y}_\delta(x))\right) \leq \frac{2L_{xy}^2}{\mu_y}\delta$$

Then for any $x, x' \in \mathcal{X}$ we have with probability at least $1 - \sigma$

$$g(x') \leq f(x, \widetilde{y}_\delta(x)) + \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x\rangle + \frac{L_g}{2}\|x - x'\|_{p_x}^2 + \delta + \sqrt{\frac{2L_{xy}^2}{\mu_y}\delta}\|x - x'\|_{p_x}$$

$$\leq f(x, \widetilde{y}_\delta(x)) + \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x\rangle + \left(\frac{L_g}{2} + \frac{2L_{xy}^2}{\mu_y}\right)\|x - x'\|_{p_x}^2 + 2\delta$$

$$\leq f(x, \widetilde{y}_\delta(x)) + \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x\rangle + \left(\frac{L_{xx}}{2} + \frac{3L_{xy}^2}{\mu_y}\right)\|x - x'\|_{p_x}^2 + 2\delta$$

$$\leq f(x, \widetilde{y}_\delta(x)) + \langle \nabla_x f(x, \widetilde{y}_\delta(x)), x' - x\rangle + 2\left(L_{xx} + \frac{2L_{xy}^2}{\mu_y}\right)\|x - x'\|_{p_x}^2 + 2\delta$$

I this way, we have function $g(x)$ equips $(2\delta, \sigma, 2L_g, \mu_x)$-oracle.

5. According to μ_x-strong convexity of function $g(x)$, we have with probability at least $1 - \sigma_x$

$$\|\hat{x} - x^*\|_{p_x}^2 \leq \frac{2}{\mu_x}\varepsilon_x$$

According to μ_x-strong concavity of function $h(x)$, using union bound we have with probability at least $1 - \sigma_y - \sigma_x$

$$\|\widetilde{y}_{\varepsilon_y}(\hat{x}) - y^*\|_{p_y}^2 \leq 2\|\widetilde{y}_{\varepsilon_y}(\hat{x}) - y^*(\hat{x})\|_{p_y}^2 + 2\|y^*(\hat{x}) - y^*\|_{p_y}^2$$

$$\leq \frac{2}{\mu_y}\varepsilon_y + \frac{L_{xy}^2}{\mu_y^2}\|\hat{x} - x^*\|_{p_x}^2 \leq \frac{2}{\mu_y}\varepsilon_y + \frac{2L_{xy}^2}{\mu_x\mu_y^2}\varepsilon_x$$

We have with probability at least $1 - \sigma_y - \sigma_x$

$$h(y^*) - h\left(\widetilde{y}_{\varepsilon_y}(\hat{x})\right) \leq \frac{L_{yy} + \frac{2L_{xy}^2}{\mu_x}}{2}\|y^* - \widetilde{y}_{\varepsilon_y}(\hat{x})\|_{p_y}^2$$

$$\leq \left(\frac{L_{yy}}{\mu_y} + \frac{2L_{xy}^2}{\mu_x\mu_y}\right)\varepsilon_y + \left(\frac{L_{xy}^2 L_{yy}}{\mu_x\mu_y^2} + \frac{2L_{xy}^4}{\mu_x^2\mu_y^2}\right)\varepsilon_x$$

Theorem 7. *Let $\varepsilon > 0$ be the target accuracy of the solution to the problem (15) and $\sigma \in (0, 1)$ be the target confidence level. Let the auxiliary problems (16), (17) be solved with accuracies*

$$\varepsilon_x = \widetilde{O}\left(\varepsilon\left(\frac{L_{xy}^2 L_{yy}}{\mu_x(\mu_y + L_{yy})^2} + \frac{2L_{xy}^4}{\mu_x^2(\mu_y + L_{yy})^2}\right)^{-1}\right);$$

$$\varepsilon_y = \widetilde{O}\left(\varepsilon\left(\frac{L_{yy}}{\mu_y + L_{yy}} + \frac{2L_{xy}^2}{\mu_x(\mu_y + L_{yy})}\right)^{-1}\left(\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x(\mu_y + L_{yy})}\right)^{-1/2}\right)$$

and confidence levels

$$\sigma_x = \widetilde{O}\left(\sigma\sqrt{\frac{\mu_y}{L_{yy}}}\right);$$

$$\sigma_y = \widetilde{O}\left(\sigma\left(\frac{L_{xx}L_{yy}}{\mu_x\mu_y} + \frac{2L_{xy}^2}{\mu_x\mu_y}\right)^{-1/2}\right),$$

that is, a $(\varepsilon_x, \sigma_x)-$solution to the problem (17) and a $(\varepsilon_y, \sigma_y)-$solution to the problem (16) are found (see Definition 5). Then, under assumptions 1(s), 2, the proposed Approach 3 guarantees to find an (ε, σ)-solution to the problem (15). Moreover, the required number of calls to the first-order oracle $\nabla_x f(x, y)$ and the zeroth-order oracle $f(x, y)$ satisfy the following bounds

Total Number of Calls for $\nabla_x f(x, y)$ is $\widetilde{O}\left(\sqrt{\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x\mu_y}}\right),$

Total Number of Calls for $f(x, y)$ is $\widetilde{O}\left(n_y\sqrt{\frac{L_{xx}L_{yy}}{\mu_x\mu_y} + \frac{2L_{xy}^2}{\mu_x\mu_y}}\right).$

Proof. **Oracles Complexities Analysis**

Now we can analyze the oracle complexity of our approach

1. The number of iterations of the Catalyst algorithm 3 (see [15]) to solve the problem (22) with the accuracy of ε:

$$\widetilde{O}\left(\sqrt{\frac{H_1}{\mu_y}}\right) \tag{54}$$

2. Now we need to determine the number of iterations of the $\mathcal{M}$ method to solve the problem (23). According to theorem 3, the number of iterations of algorithm 2 required to find the solution with accuracy ε_y is

$$N_{\text{ARDDsc}} = \widetilde{O}\left(n_y\sqrt{\frac{L_{yy} + H_1}{\mu_y + H_1}}\log\left(\frac{(\mu_y + H_1)R_{p_y}^2}{\varepsilon_y}\right)\right). \tag{55}$$

3. Now we solve the outer problem (27) with the algorithm 4. We need to determine the Lipschitz constant and the strong convexity constant. The strong convexity constant remains the same, the Lipschitz constant, according to Lemma 4, is

$$L_g = L_{xx} + \frac{2L_{xy}^2}{\mu_y + H_1}.$$

Complexity of solution of the outer problem (27) (number of calculations of the gradient value $\nabla_x f(x, y)$) with accuracy ε_x:

$$N_{\text{outer}} = \widetilde{O}\left(\sqrt{\frac{L_g}{\mu_x}}\right) = \widetilde{O}\left(\sqrt{\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x(\mu_y + H_1)}}\right) \tag{56}$$

4. Thus, the total complexity of calls of zeroth-order oracle $f(x, y)$, solving the inner problem (15) by the algorithm 2 with accuracy ε, taking $H_1 = L_{yy}$, is equal to

$$\begin{aligned}
N_{\text{inner}} &= \widetilde{O}\left(n_y\sqrt{\frac{L_{yy} + H_1}{\mu_y + H_1}}\sqrt{\frac{H_1}{\mu_y}}\sqrt{\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x(\mu_y + H_1)}}\right) \\
&= \widetilde{O}\left(n_y\sqrt{\frac{L_{yy} + L_{yy}}{\mu_y + L_{yy}}}\sqrt{\frac{L_{yy}}{\mu_y}}\sqrt{\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x(\mu_y + L_{yy})}}\right) \\
&= \widetilde{O}\left(n_y\sqrt{\frac{L_{yy} + L_{yy}}{L_{yy}}}\sqrt{\frac{L_{xx}L_{yy}}{\mu_x\mu_y} + \frac{2L_{xy}^2 L_{yy}}{\mu_x\mu_y L_{yy}}}\right) \\
&= \widetilde{O}\left(n_y\sqrt{\frac{L_{xx}L_{yy}}{\mu_x\mu_y} + \frac{2L_{xy}^2}{\mu_x\mu_y}}\right)
\end{aligned}$$

Inexactness Complexities Analysis

Now it remains to determine with what accuracy ε_x, ε_y it is necessary to solve inner (26) and outer (27) problems, so that the number of calls to the oracles satisfies this ratio (**??**). In order to answer this question, we will introduce new designations and redefine some things. First, we denote

$$\xi(x) = \max_{y \in \mathbb{R}^{n_y}} \psi(x, y), \ \zeta(y) = \min_{x \in \mathcal{X}} \psi(x, y)$$

Let $\hat{x} \in \mathcal{X}$ be $(\varepsilon_x, \sigma_x)$-solution to problem $\min_{x \in \mathcal{X}} \xi(x)$, in other words, we have with probability at least $1 - \sigma_x$:

$$0 \leq \xi(\hat{x}) - \min_{x \in \mathcal{X}} \xi(x) = \psi(\hat{x}, y^*(\hat{x})) - \psi(x^*, y^*) \leq \varepsilon_x$$

Let $\widetilde{y}_{\varepsilon_y}(\hat{x})$ - $(\varepsilon_y, \sigma_y)$-solution to problem $\max_{y \in \mathbb{R}^{n_y}} \psi(\hat{x}, y)$, in other words, we have with probability at least $1 - \sigma_y$:

$$0 \leq \psi(\hat{x}, y^*(\hat{x})) - \psi(\hat{x}, \widetilde{y}_{\varepsilon_y}(\hat{x})) \leq \varepsilon_y$$

First, according to corollary 1 for $\min_{x \in \mathcal{X}} \xi(x)$, we have a limit on noise ε_y:

$$\varepsilon_y \left(1 + \sqrt{\frac{L_g}{\mu_x}} \right) = O\left(\varepsilon_x\right)$$

Thus, we have

$$\varepsilon_y = O\left(\varepsilon_x \sqrt{\frac{\mu_x}{L_g}} \right)$$

It remains only to understand how to define ε_x via ε. To answer this question using the main lemma 4, replacing μ_y with $\mu_y + H_1$ and L_{yy} with $L_{yy} + H_1$, we obtain inequality:

$$\widetilde{\varepsilon} \geq \left(\frac{L_{yy} + H_1}{\mu_y + H_1} + \frac{2L_{xy}^2}{\mu_x \mu_y} \right) \varepsilon_y + \left(\frac{L_{xy}^2 (L_{yy} + H_1)}{\mu_x (\mu_y + H_1)^2} + \frac{2L_{xy}^4}{\mu_x^2 (\mu_y + H_1)^2} \right) \varepsilon_x$$

According to results from [13], we have $\widetilde{\varepsilon} = O(\varepsilon)$. Then we have

$$\varepsilon_x = \widetilde{O}\left(\varepsilon \left(\frac{L_{xy}^2 (L_{yy} + H_1)}{\mu_x (\mu_y + H_1)^2} + \frac{2L_{xy}^4}{\mu_x^2 (\mu_y + H_1)^2} \right)^{-1} \right)$$

$$\varepsilon_y = \widetilde{O}\left(\varepsilon \left(\frac{L_{yy} + H_1}{\mu_y + H_1} + \frac{2L_{xy}^2}{\mu_x (\mu_y + H_1)} \right)^{-1} \sqrt{\frac{\mu_x}{L_g}} \right)$$

Using that $H_1 = L_{yy}$, we have

$$\varepsilon_x = \widetilde{O}\left(\varepsilon\left(\frac{L_{xy}^2 L_{yy}}{\mu_x(\mu_y + L_{yy})^2} + \frac{2L_{xy}^4}{\mu_x^2(\mu_y + L_{yy})^2}\right)^{-1}\right)$$

$$\varepsilon_y = \widetilde{O}\left(\varepsilon\left(\frac{L_{yy}}{\mu_y + L_{yy}} + \frac{2L_{xy}^2}{\mu_x(\mu_y + L_{yy})}\right)^{-1}\sqrt{\frac{\mu_x}{L_{xx} + \frac{2L_{xy}^2}{\mu_y + L_{yy}}}}\right)$$

According to theorem 6 and approach 3, we have that

$$\sigma_x = N_{\text{outer}}\sigma_y, \text{where } N_{\text{outer}} = \widetilde{O}\left(\sqrt{\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x(\mu_y + L_{yy})}}\right)$$

Having obtained this ratio, we now need to determine how they relate to each other. Using the catalyst algorithm and the convergence proof, one can easily show that

$$\sigma \geq N_{\text{Catalyst}}(\sigma_y + \sigma_x) > N_{\text{Catalyst}} N_{\text{outer}} \sigma_y, \text{ where } N_{\text{Catalyst}} = \widetilde{O}\left(\sqrt{\frac{L_{yy}}{\mu_y}}\right)$$

Thus we have

$$\sigma_y = \widetilde{O}\left(\frac{\sigma}{N_{\text{Catalyst}} N_{\text{outer}}}\right) = \widetilde{O}\left(\sigma\sqrt{\frac{\mu_y}{L_{yy}}\left(\frac{L_{xx}}{\mu_x} + \frac{2L_{xy}^2}{\mu_x(\mu_y + L_{yy})}\right)^{-1}}\right)$$

$$= \widetilde{O}\left(\sigma\sqrt{\left(\frac{L_{xx}L_{yy}}{\mu_x\mu_y} + \frac{2L_{xy}^2 L_{yy}}{\mu_x(\mu_y + L_{yy})\mu_y}\right)^{-1}}\right)$$

$$= \widetilde{O}\left(\sigma\left(\frac{L_{xx}L_{yy}}{\mu_x\mu_y} + \frac{2L_{xy}^2}{\mu_x\mu_y}\right)^{-1/2}\right)$$

$$\sigma_x = \sigma_y N_{\text{outer}} = \widetilde{O}\left(\frac{\sigma N_{\text{outer}}}{N_{\text{Catalyst}} N_{\text{outer}}}\right) = \widetilde{O}\left(\frac{\sigma}{N_{\text{Catalyst}}}\right) = \widetilde{O}\left(\sigma\sqrt{\frac{\mu_y}{L_{yy}}}\right)$$