Mathematical Programming
https://doi.org/10.1007/s10107-019-01420-0

FULL LENGTH PAPER

Series A

Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems

Yuyuan Ouyang[1] · **Yangyang Xu**[2]

Received: 6 August 2018 / Accepted: 2 August 2019

Abstract
On solving a convex-concave bilinear saddle-point problem (SPP), there have been many works studying the complexity results of first-order methods. These results are all about upper complexity bounds, which can determine at most how many iterations would guarantee a solution of desired accuracy. In this paper, we pursue the opposite direction by deriving lower complexity bounds of first-order methods on large-scale SPPs. Our results apply to the methods whose iterates are in the linear span of past first-order information, as well as more general methods that produce their iterates in an arbitrary manner based on first-order information. We first work on the affinely constrained smooth convex optimization that is a special case of SPP. Different from gradient method on unconstrained problems, we show that first-order methods on affinely constrained problems generally cannot be accelerated from the known convergence rate $O(1/t)$ to $O(1/t^2)$, and in addition, $O(1/t)$ is optimal for convex problems. Moreover, we prove that for strongly convex problems, $O(1/t^2)$ is the best possible convergence rate, while it is known that gradient methods can have linear convergence on unconstrained problems. Then we extend these results to general SPPs. It turns out that our lower complexity bounds match with several established upper complexity bounds in the literature, and thus they are tight and indicate the optimality of several existing first-order methods.

Keywords Convex optimization · Saddle point problems · First-order methods · Information-based complexity · Lower complexity bound

Mathematics Subject Classification 90C25 · 90C06 · 90C60 · 49M37 · 68Q25

Ouyang's research is partly supported by NSF grant DMS-1913006 and ONR award N00014-19-1-2295, and Xu's research is partly supported by NSF grant DMS-1719549.

✉ Yangyang Xu
xuy21@rpi.edu

Extended author information available on the last page of the article

Published online: 07 August 2019

1 Introduction

In recent years, first-order methods have been particularly popular partly due to the huge scale of many modern applications. These methods only access the function value and gradient information of the underlying problems, and possibly also other "simple" operations. For example, on solving the constrained optimization problem $f^* := \min_{\mathbf{x}\in X} f(\mathbf{x})$, the projected gradient (PG) method

$$\mathbf{x}^{(t+1)} \leftarrow \mathrm{Proj}_X\left(\mathbf{x}^{(t)} - \alpha\nabla f(\mathbf{x}^{(t)})\right)$$

is a first-order method if the projection operator Proj_X is easy to evaluate such as projection onto a box constraint set. For convex problems, if ∇f is Lipschitz continuous and α is appropriately chosen, the PG method can have convergence rate in the order of $\frac{1}{t}$, namely, $f(\mathbf{x}^{(t)}) - f^* = O(\frac{1}{t})$, where t is the number of gradient evaluations. Through smart extrapolation, the rate can be improved to $O(\frac{1}{t^2})$; see [3,36]. In addition, there exists an instance showing that the order $\frac{1}{t^2}$ cannot be further improved (see [31–34]) and thus is optimal.

In this paper, we consider the bilinear saddle-point problem (SPP):

$$\min_{\mathbf{x}\in X}\max_{\mathbf{y}\in Y} \mathcal{L}(\mathbf{x},\mathbf{y}) := f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle - g(\mathbf{y}). \tag{1.1}$$

Here, $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ are closed convex sets, $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ are closed convex functions, $\mathbf{A} \in \mathbb{R}^{m\times n}$, and $\mathbf{b} \in \mathbb{R}^m$. We assume that the function f is L_f-smooth, namely, f is differentiable, and ∇f is L_f-Lipschitz continuous:

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \le L_f\|\mathbf{x}_1 - \mathbf{x}_2\|,\ \forall\, \mathbf{x}_1, \mathbf{x}_2 \in X. \tag{1.2}$$

In addition, we assume that g is simple such that its proximal mapping can be easily computed. The scale of the problem is large, so it is expensive to form the Hessian of f and also, it is expensive to solve or project onto a linear system of size $m \times n$.

Two optimization problems are associated with (1.1). One is called the primal problem

$$\phi^* := \min_{\mathbf{x}\in X}\left\{\phi(\mathbf{x}) := f(\mathbf{x}) + \max_{\mathbf{y}\in Y}\langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle - g(\mathbf{y})\right\}, \tag{1.3}$$

and the other is the dual problem

$$\psi^* := \max_{\mathbf{y}\in Y}\left\{\psi(\mathbf{y}) := -g(\mathbf{y}) + \min_{\mathbf{x}\in X}\langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle + f(\mathbf{x})\right\}. \tag{1.4}$$

The weak duality always holds, i.e., $\psi^* \le \phi^*$. Under certain mild assumptions (e.g., X and Y are compact [38]), the strong duality holds, i.e., $\psi^* = \phi^*$, and in this case,

(1.1) has a saddle point $(\mathbf{x}^*, \mathbf{y}^*)$, namely,

$$\mathcal{L}(\mathbf{x}^*, \mathbf{y}) \le \mathcal{L}(\mathbf{x}^*, \mathbf{y}^*) \le \mathcal{L}(\mathbf{x}, \mathbf{y}^*),\ \forall \mathbf{x} \in X,\ \forall \mathbf{y} \in Y. \tag{1.5}$$

Many applications can be formulated into an SPP. For instance, it includes as special cases all affinely constrained smooth convex optimization problems. To see this, let $Y = \mathbb{R}^m$ and $g \equiv 0$. Then $\max_{\mathbf{y}} \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle = 0$ if $\mathbf{A}\mathbf{x} = \mathbf{b}$ and ∞ otherwise, and thus (1.3) becomes

$$f^* := \min_{\mathbf{x} \in X} \big\{ f(\mathbf{x}),\ \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b} \big\}. \tag{1.6}$$

1.1 Main goal

We aim at answering the following question:

> For any deterministic first-order method, what is the best possible performance on solving a general large scale saddle-point problem (1.1)?

More precisely, our goal is to study the *lower information-based complexity bound* of *first-order methods* on solving the class of problems that can be formulated into (1.1). In the literature, all existing works about first-order methods on solving saddle-point problems only provide upper complexity bounds. Establishing lower complexity bounds is important because they can tell us whether the existing methods are improvable and also because they can guide us to design "optimal" algorithms that have the best performance. To achieve this goal, we will construct worst-case SPP instances such that the complexity result of a first-order method to reach a desired accuracy is lower bounded by a problem-dependent quantity.

In the above question, we say an iterative algorithm for solving (1.1) is a *first-order method* if it accesses the information of the function f and the matrix $\mathbf{A}$ through a *first-order oracle*, denoted by $\mathcal{O} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$. For an inquiry on any point $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^m$, the oracle returns

$$\mathcal{O}(\mathbf{x}, \mathbf{y}) := \big(\nabla f(\mathbf{x}), \mathbf{A}\mathbf{x}, \mathbf{A}^\top \mathbf{y}\big). \tag{1.7}$$

Given an initial point $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)})$, a first-order method $\mathcal{M}$ for solving SPPs, at the t-th iteration, calls the oracle on $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ to collect the oracle information $\mathcal{O}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ and then obtains a new point $(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)})$ by a rule $\mathcal{I}_t$. The complete method $\mathcal{M}$ can be described by the initial point $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) \in X \times Y$ and a sequence of rules $\{\mathcal{I}_t\}_{t=0}^{\infty}$ such that

$$\big(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, \bar{\mathbf{x}}^{(t+1)}, \bar{\mathbf{y}}^{(t+1)}\big) = \mathcal{I}_t\Big(\boldsymbol{\vartheta}; \mathcal{O}(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}), \ldots, \mathcal{O}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\Big),\ \forall t \ge 0, \tag{1.8}$$

where $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) \in X \times Y$ denotes the inquiry point, and $(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}) \in X \times Y$ is the approximate solution by the method. We are interested at the performance of $\mathcal{M}$ for

solving very large scale instances of (1.1), namely, m and n are so large that we can only afford $t \ll m$ oracle calls. In (1.8), ϑ contains all rest information in an SPP, including the sets X and Y, the function g, the vector $\mathbf{b}$, and the Lipschitz constant L_f and its associated norm. Given a maximum number T of iterations, we assume without loss of generality that the output by $\mathcal{M}$ coincides with the last inquiry point, namely, $\bar{\mathbf{x}}^{(T+1)} = \mathbf{x}^{(T+1)}$ and $\bar{\mathbf{y}}^{(T+1)} = \mathbf{y}^{(T+1)}$.

1.2 Literature review

Among existing works on complexity analysis of numerical methods, many more are about showing upper complexity bounds instead of lower bounds. Usually, the upper complexity bounds are established on solving problems with specific structures. They are important because they can tell the users at most how many iterations would guarantee a desired solution. On the contrary, lower complexity bounds, which were first studied in the seminal work [31], are usually information-based and shown on solving a general class of problems. Their importance lies in telling if a certain numerical method can still be improved for a general purpose and also in guiding the algorithm designers to make "optimal" methods. Although there are not many works along this line, each of them sets a base for designing numerical approaches. Below we review these lower complexity bound results on different classes of problems.

Proximal gradient methods On solving convex problems in the form of $F^* := \min_{\mathbf{x}}\{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$, the proximal gradient method (PGM) iteratively updates the estimated solution by acquiring information of ∇f and $\mathbf{prox}_{\eta g}$ at certain points, where $\eta > 0$ is the stepsize, and the proximal mapping of ηg is defined as

$$\mathbf{prox}_{\eta g}(\mathbf{z}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{z}\|^2.$$

For the problem class that has L_f-smooth f, the lower bound has been established in [15,31,32,34]. For example, [34, Theorem 2.1.7] establishes a lower convergence rate bound: $F(\bar{\mathbf{x}}^{(t)}) - F^* \geq \frac{3L_f\|\mathbf{x}^{(0)}-\mathbf{x}^*\|^2}{32(t+1)^2}$, where $\bar{\mathbf{x}}^{(t)}$ is the approximate solution output by PGM after t iterations, and $\mathbf{x}^*$ is one optimal solution. In addition, setting $\eta = \frac{1}{L_f}$, [3,36] show that the PGM can achieve $O(L_f/t^2)$ convergence rate, and more precisely, $F(\bar{\mathbf{x}}^{(t)}) - F^* \leq \frac{2L_f\|\mathbf{x}^{(0)}-\mathbf{x}^*\|^2}{(t+1)^2}$. Comparing the lower and upper bounds, one can easily see that they differ only by a constant multiple. Hence, the lower bound is tight in terms of the dependence on t, L_f, and $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|$, and also the method given in [3,36] is optimal[1] among all methods that only access the information of ∇f and $\mathbf{prox}_{\eta g}$.

For the class of problems where f is L_f-smooth and also μ-strongly convex (μ-SC), namely,

$$\langle \nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2\rangle \geq \mu\|\mathbf{x}_1 - \mathbf{x}_2\|^2, \ \forall \mathbf{x}_1, \mathbf{x}_2,$$

[1] By "optimal", we mean that the convergence rate cannot be further improved for the considered problem class.

the lower bound has been established in [31–34]. For example, [34, Theorem 2.1.13] establishes a lower convergence rate bound: $F(\bar{\mathbf{x}}^{(t)}) - F^* \geq \frac{\mu\|\mathbf{x}^{(0)}-\mathbf{x}^*\|^2}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t}$, where $\kappa = \frac{L_f}{\mu}$ denotes the condition number. In addition, assuming the knowledge of μ and L_f, [36, Theorem 6] shows the convergence rate: $F(\bar{\mathbf{x}}^{(t)}) - F^* \leq \frac{L_f\|\mathbf{x}^{(0)}-\mathbf{x}^*\|^2}{4}\left(1+\frac{1}{\sqrt{2\kappa}}\right)^{-2t}$. Note that both lower and upper bounds of convergence rate are linear, and they have the same dependence on $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|$ and κ. In this sense, the lower bound is tight, and the method is optimal.

Inexact gradient methods On the convex problem $f^* := \min_{\mathbf{x}} f(\mathbf{x})$ for which only approximation of ∇f is available, there have been several studies on the corresponding lower complexity bound. For example, on solving the convex stochastic program $f^* := \min_{\mathbf{x}}\{f(\mathbf{x}) := \mathbb{E}_\xi f_\xi(\mathbf{x})\}$, the stochastic gradient method (SGM) performs iterative update to the solution by accessing the stochastic approximation of subgradient $\tilde{\nabla} f$ at a certain point. For the class of problems whose f is Lipschitz continuous[2], [31] shows that to find a stochastic ε-optimal solution $\bar{\mathbf{x}}$, i.e., $\mathbb{E} f(\bar{\mathbf{x}}) - f^* \leq \varepsilon$, the algorithm needs to run $O(1/\varepsilon^2)$ iterations. On the other hand, as shown in [30], the order $1/\varepsilon^2$ is achievable with appropriate setting of algorithm parameters. Hence, the lower complexity bound $O(1/\varepsilon^2)$ is tight, and the stochastic gradient method is optimal on finding an approximate solution to the convex stochastic program. Further study of lower complexity bound of inexact gradient methods is also performed in [10]. When $f(\mathbf{x})$ has a special finite-sum structure, the lower complexity bound of randomized gradient method is studied in [1,26,40].

Primal-dual first-order methods On an affinely constrained problem (1.6) or the more general saddle-point problem (1.1), many works have studied primal-dual first-order methods, e.g., [6,7,9,11–13,16,18,24,37,42,44]. To obtain an ε-optimal solution in a certain measure, an $O(1/\varepsilon)$ complexity result is established by many of them for convex problems. In addition, for strongly convex cases, an improved result of $O(1/\sqrt{\varepsilon})$ has been shown in a few works such as [14,16,42,43]. All these results are about upper complexity bounds and none about lower bounds. Hence, it is unclear if these methods achieve the optimal order of convergence rate. Our results fill the missing part and can be used to determine the optimality of these existing algorithms.

Others In adddition to the above list of lower complexity bounds, there are also a few results on special types of problems. For convex quadratic minimization, [39] gives a high-probability lower complexity bound of randomized first-order method. The lower complexity bound of subgradient methods for uniformly convex optimization has been studied in [20]. Under the assumption that an algorithm has access to gradient information and is only allowed to perform linear optimization (instead of computing a projection), the lower complexity bounds have been studied in [19,21]. The lower complexity bounds of oblivious algorithms are studied in [2], where the way to generate new iterates by the algorithms is restricted. To find stationary points of smooth convex

[2] f is Lipschitz continuous on a set X if there is a constant L such that $|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|$ for any $\mathbf{x}_1, \mathbf{x}_2 \in X$.

or nonconvex problems [4,5], study the lower complexity bounds of first-order and also higher-order methods.

We list in Table 1 several results that are reviewed above and also the results we establish in this paper.

1.3 Research tools, main results, and contributions

In this subsection, without specifying many technical details, we state the main results obtained in this paper, and the research tools that lead to such results. We start with a brief review of the seminal research tools developed in [31–34] that lead to the classical lower complexity bound result of first-order methods for unconstrained smooth convex optimization $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$. Then, we describe our efforts adapting their research tools to derive lower complexity bounds of first-order methods for SPPs, and state our main results.

The work in [32,33] constructs a worst-case instance of unconstrained smooth convex optimization in the form of a quadratic problem:

$$f^* := \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{q}^\top \mathbf{x} \right\}, \tag{1.9}$$

which is also equivalent to solving the linear system $\mathbf{Q}\mathbf{x} = \mathbf{q}$. Briefly speaking, the main research tool in [32,33] is rotating into a certain linear subspace the iterates given by a deterministic first-order method for solving (1.9). Ignoring most of the technical details, given $(\mathbf{Q}, \mathbf{q})$ and a deterministic first-order method $\mathcal{M}$, the tool allows us to analyze the performance of $\mathcal{M}$ by simply assuming that the output $\bar{\mathbf{x}}^{(t)}$ lies in $\mathcal{K}_{2t+1} := \text{span}\{\mathbf{q}, \mathbf{Q}\mathbf{q}, \ldots, \mathbf{Q}^{2t}\mathbf{q}\}$, the Krylov subspace generated by $\mathbf{Q}$ and $\mathbf{q}$. To construct a worst-case instance of (1.9), it then suffices to maximize the objective value difference $\min_{\mathbf{x}\in\mathcal{K}_{2t+1}} f(\mathbf{x}) - f^*$ with respect to $(\mathbf{Q}, \mathbf{q})$. In [32,33], it is proved that the worst-case $\mathbf{Q}$ can be any matrix (e.g., diagonal matrix) that has certain specified eigenvalues, which are computed through Chebyshev Equioscillation theorem. In [34], a tri-diagonal worst-case matrix $\mathbf{Q}$ is constructed along with a worst-case vector $\mathbf{q}$. The tri-diagonal $\mathbf{Q}$ has eigenvalues specified in [32,33]. Also, with the $(\mathbf{Q}, \mathbf{q})$, $\mathcal{K}_{2t+1}$ is spanned by standard basis vectors. This makes it easier to prove that the constructed $\mathbf{Q}$ and $\mathbf{q}$ yield worst-case performance of first-order methods. The analysis in [34] focuses only on first-order methods whose iterates lie in the Krylov subspace. However, combining with the aforementioned research tool in [32,33], the result can be extended to an arbitrary deterministic first-order method.

Our main contribution is to establish lower complexity bounds of deterministic first-order methods on solving bilinear saddle-point problem (1.1), through adapting the techniques in [32,33] and [34]. First, we design a worst-case instance of the convex constrained problem (1.6). It is an affinely constrained convex quadratic program:

$$f^* := \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x} - \mathbf{h}^\top \mathbf{x}, \ \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{b} \right\}.$$

Our design of $(\mathbf{H}, \mathbf{h}, \mathbf{A}, \mathbf{b})$ is inspired by the setting in [34] of $(\mathbf{Q}, \mathbf{q})$ in a worst-case instance of (1.9). Also, we follow the constructive proof in [34] and focus on

Table 1 Lower and upper complexity bounds of first-order methods for different problem classes on producing an ε-solution

Problem	Conditions	Oracle	lower compl. bound	upper compl. bound
$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$	f is L_f-smooth	$(\nabla f, \mathbf{prox}_{\eta g})$	$O(\sqrt{\frac{L_f}{\varepsilon}})$ [34]	$O(\sqrt{\frac{L_f}{\varepsilon}})$ [3,36]
Where g is closed convex	f is L_f-smooth and μ-SC	$(\nabla f, \mathbf{prox}_{\eta g})$	$O(\sqrt{\frac{L_f}{\mu}}\log\frac{1}{\varepsilon})$ [34]	$O(\sqrt{\frac{L_f}{\mu}}\log\frac{1}{\varepsilon})$ [36]
$\min_{\mathbf{x}} f(\mathbf{x}) := \mathbb{E} f_\xi(\mathbf{x})$	f is Lipschitz continuous	$\tilde{\nabla} f_\xi$	$O(\frac{1}{\varepsilon^2})$ [31,34]	$O(\frac{1}{\varepsilon^2})$ [30]
$\min_{\mathbf{x}\in X} f(\mathbf{x})$, s.t. $\mathbf{Ax} = \mathbf{b}$	f is L_f-smooth	$(\nabla f, \mathbf{Ax}, \mathbf{A}^\top\mathbf{y})$	$O(\sqrt{\frac{L_f}{\varepsilon}} + \frac{\|\mathbf{A}\|}{\varepsilon})$ [this paper]	$O(\sqrt{\frac{L_f}{\varepsilon}} + \frac{\|\mathbf{A}\|}{\varepsilon})$[37]
SPP (1.1)	f is L_f-smooth	$(\nabla f, \mathbf{Ax}, \mathbf{A}^\top\mathbf{y})$	$O(\sqrt{\frac{L_f}{\varepsilon}} + \frac{\|\mathbf{A}\|}{\varepsilon})$ [this paper]	$O(\sqrt{\frac{L_f}{\varepsilon}} + \frac{\|\mathbf{A}\|}{\varepsilon})$ [35]

All problems are convex. First through third rows: ε-solution $\bar{\mathbf{x}}$ if $\text{obj}(\bar{\mathbf{x}}) - \text{obj}^* \le \varepsilon$; fourth row: ε-solution $\bar{\mathbf{x}} \in X$ if $f(\bar{\mathbf{x}}) - f^* \le \varepsilon$ and $\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| \le \varepsilon$; fifth row: ε-solution $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in X \times Y$ if primal dual gap $\phi(\bar{\mathbf{x}}) - \psi(\bar{\mathbf{y}}) \le \varepsilon$. Here, f is L_f-smooth if (1.2) holds, and SC stands for strongly convex

first-order methods whose iterates lie in a certain linearly spanned subspace. This way, we establish the lower complexity bound of deterministic first-order methods on solving (1.6). Secondly, we adapt the rotation technique in [32,33] to relax the linear span restriction and show lower complexity of any deterministic first-order method on solving (1.6). Due to the additional linear constraint, the analysis in [32,33] cannot be directly applied. It may not be true that for any ($\mathbf{H}$, $\mathbf{h}$, $\mathbf{A}$, $\mathbf{b}$), the iterates by a given first-order method $\mathcal{M}$ can be rotated into a Krylov subspace. However, with our specific designed ($\mathbf{H}$, $\mathbf{h}$, $\mathbf{A}$, $\mathbf{b}$), we are able to do so. Finally, we use as a bridge the designed worse-case instance of (1.6) to design a worst-case instance of (1.1) and thus address our main question posed in Sect. 1.1.

The main results we obtain in this paper are summarized in the following two theorems. Throughout this paper, by $a_t = \Omega(b_t)$, we mean that there is a positive constant C independent of t such that $a_t \geq C \cdot b_t$.

Theorem 1.1 (Lower complexity bounds for affinely constrained problems) Let t be a positive integer, $L_f > 0$, and $L_A > 0$. For any first-order method $\mathcal{M}$ that is described in (1.8), there exists a problem instance of (1.6) such that f is L_f-smooth, $\|\mathbf{A}\| = L_A$, the instance has a primal-dual solution $(\mathbf{x}^*, \mathbf{y}^*)$, and

$$\left|f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*)\right| = \Omega\left(\frac{L_f\|\mathbf{x}^*\|^2}{t^2} + \frac{L_A\|\mathbf{x}^*\| \cdot \|\mathbf{y}^*\|}{t}\right),$$
$$\|\mathbf{A}\bar{\mathbf{x}}^{(t)} - \mathbf{b}\| = \Omega\left(\frac{L_A\|\mathbf{x}^*\|}{t}\right),$$

where $\bar{\mathbf{x}}^{(t)}$ is the approximate solution output by $\mathcal{M}$. In addition, given $\mu > 0$, there exists an instance of (1.6) with μ-strongly convex function f, and it has a primal-dual solution $(\mathbf{x}^*, \mathbf{y}^*)$ satisfying

$$\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 = \Omega\left(\frac{L_A^2\|\mathbf{y}^*\|^2}{\mu^2 t^2}\right).$$

Theorem 1.2 (Lower complexity bounds for bilinear saddle-point problems) Let t be a positive integer, $L_f > 0$, and $L_A > 0$. For any first-order method $\mathcal{M}$ that is described in (1.8), there exists a problem instance of (1.1) such that f is L_f-smooth, $\|\mathbf{A}\| = L_A$, X and Y are Euclidean balls with radii R_X and R_Y respectively, and

$$\phi(\bar{\mathbf{x}}^{(t)}) - \psi(\bar{\mathbf{y}}^{(t)}) = \Omega\left(\frac{L_f R_X^2}{t^2} + \frac{L_A R_X R_Y}{t}\right),$$

where ϕ and ψ are the associated primal and dual objective functions in (1.3) and (1.4), and $(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)})$ is the approximate solution output by $\mathcal{M}$. In addition, given $\mu > 0$, there exists an instance of (1.1) such that f is μ-strongly convex, X and Y are Euclidean balls with radii R_X and R_Y respectively, and

$$\phi(\bar{\mathbf{x}}^{(t)}) - \psi(\bar{\mathbf{y}}^{(t)}) = \Omega\left(\frac{L_A^2 R_Y^2}{\mu t^2}\right).$$

Comparing to upper complexity bounds of several existing first-order methods, we find that our lower complexity bounds are tight, up to the difference of constant multiples and/or logarithmic terms.

1.4 Notation and outline

We use bold lower-case letters $\mathbf{x}, \mathbf{y}, \mathbf{c}, \ldots$ for vectors and bold upper-case letters $\mathbf{A}, \mathbf{Q}, \ldots$ for matrices. For any vector $\mathbf{x} \in \mathbb{R}^n$, we use x_i to denote its i-th component. When describing an algorithm, we use $\mathbf{x}^{(k)}$ for the k-th iterate. $\mathbf{A}^\top$ denotes the transpose of a matrix $\mathbf{A}$. We use $\mathbf{0}$ for all-zero vector and $\mathbf{1}$ for all-one vector, and we use $\mathbf{O}$ for a zero matrix and $\mathbf{I}$ for the identity matrix. Their sizes will be specified by a subscript, if necessary, and otherwise are clear from the context. We adopt MATLAB's operations to concatenate matrices and vectors. For example, $\mathbf{e}_{j,p} = [\mathbf{0}_{j-1}; 1; \mathbf{0}_{p-j}] \in \mathbb{R}^p$ denotes the j-th standard basis vector in $\mathbb{R}^p$. We use $\mathbb{Z}_{++}$ for the set of positive integers and $\mathbb{S}^n_+$ for the set of all $n \times n$ symmetric positive semidefinite matrices. Without further specification, $\|\cdot\|$ is used for the Euclidean norm of a vector and the spectral norm of a matrix.

The rest of the paper is organized as follows. In Sect. 2, for affinely constrained problems, we present lower complexity bounds of first-order methods that satisfy a linear span requirement. We drop the linear span assumption in Sect. 3 and show lower complexity bounds of first-order methods that are described in (1.8). Section 4 is about the bilinear saddle-point problems. Lower complexity bounds are established there for first-order methods described in (1.8). In Sect. 5, we show the tightness of the established lower complexity bounds by comparing them with existing upper complexity bounds. Finally, Sect. 6 proposes a few interesting topics for future work and concludes the paper.

2 Lower complexity bounds under linear span assumption for affinely constrained problems

In this and the next sections, we study lower complexity bounds of first-order methods on solving the affinely constrained problem (1.6). Our approach is to design a "hard" problem instance such that the convergence rate of any first-order method is lower bounded. The designed instances are convex quadratic programs in the form of

$$f^* := \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x} - \mathbf{h}^\top \mathbf{x}, \ \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{b} \right\}, \tag{2.1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, and $\mathbf{H} \in \mathbb{S}^n_+$. Note that the above problem is a special case of (1.6).

Throughout this section, we assume that

- the dimensions $m, n \in \mathbb{Z}_{++}$ are given and satisfy $m \le n$, and
- a fixed positive integer number $k < \frac{m}{2}$ is specified.

Our lower complexity analysis will be based on the performance of the k-th iterate of a first-order method on solving the designed instance. It should be noted that the

assumption $k < \frac{m}{2}$ is valid if the problem dimensions m and n are very big and we do not run too many iterations of the algorithm.

To have a relatively simple start, we focus on a special class of first-order methods in this section. More precisely, we make the following assumption.

Assumption 2.1 (Linear span) The iterate sequence $\{\mathbf{x}^{(t)}\}_{t=0}^{\infty}$ satisfies $\mathbf{x}^{(0)} = \mathbf{0}$ and

$$\mathbf{x}^{(t)} \in \text{span}\left\{\nabla f(\mathbf{x}^{(0)}), \mathbf{A}^\top \mathbf{r}^{(0)}, \nabla f(\mathbf{x}^{(1)}), \mathbf{A}^\top \mathbf{r}^{(1)}, \ldots, \nabla f(\mathbf{x}^{(t-1)}), \mathbf{A}^\top \mathbf{r}^{(t-1)}\right\}, \ t \geq 1,$$

where $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{b}$ denotes the residual.

In the context, we refer to the above assumption as the linear span assumption. It is not difficult to find rules $\{\mathcal{I}_t\}_{t=0}^{\infty}$ such that the iterate sequence $\{\mathbf{x}^{(t)}\}$ in Assumption 2.1 can be obtained by (1.8). Note that we do not lose generality by assuming $\mathbf{x}^{(0)} = \mathbf{0}$, because otherwise we can consider a shifted problem

$$\min_{\mathbf{x}} f(\mathbf{x} - \mathbf{x}^{(0)}), \ \text{s.t.} \ \mathbf{A}(\mathbf{x} - \mathbf{x}^{(0)}) = \mathbf{b}.$$

It should be noted that Assumption 2.1 may not always hold for a first-order method, e.g., when there is projection involved in the algorithm. The lower complexity bound analysis can be performed without the linear span assumption, thanks to a technique introduced in [32,33] that utilizes a certain rotational invariance of quadratic functions over a Euclidean ball. To facilitate reading, we defer the incorporation of such a technique to Sect. 3, where we will elaborate on the technical details and perform the lower complexity bound analysis without Assumption 2.1.

2.1 Special linear constraints

In this subsection, we describe a set of special linear constraints, which will be used to study the lower complexity bound of first-order methods satisfying Assumption 2.1.

We let the matrix $\boldsymbol{\Lambda}$ and vector $\mathbf{c}$ be

$$\boldsymbol{\Lambda} = \begin{bmatrix} \mathbf{B} & \mathbf{O} \\ \mathbf{O} & \mathbf{G} \end{bmatrix} \in \mathbb{R}^{m\times n} \text{ and } \mathbf{c} = \begin{bmatrix} \mathbf{1}_{2k} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^m, \tag{2.2}$$

where $\mathbf{G} \in \mathbb{R}^{(m-2k)\times(n-2k)}$ is any matrix of full row rank such that $\|\mathbf{G}\| = 2$, and

$$\mathbf{B} := \begin{bmatrix} & & & & -1 & 1 \\ & & & \iddots & \iddots & \\ & & -1 & 1 & & \\ & -1 & 1 & & & \\ -1 & & & & & \\ 1 & & & & & \end{bmatrix} \in \mathbb{R}^{2k\times 2k}. \tag{2.3}$$

All the designed "hard" instances in this paper are built upon $\boldsymbol{\Lambda}$ and $\mathbf{c}$ given in (2.2). Two immediate observations regarding (2.2) and (2.3) are as follows. First, for any

$\mathbf{u} = (u_1; \cdots; u_{2k}) \in \mathbb{R}^{2k}$, we have

$$\begin{aligned}\|\mathbf{B}\mathbf{u}\|^2 = (u_{2k} - u_{2k-1})^2 + \cdots + (u_2 - u_1)^2 + u_1^2 \le 2(u_{2k}^2 + u_{2k-1}^2) \\ + \cdots + 2(u_2^2 + u_1^2) + u_1^2 \le 4\|\mathbf{u}\|^2,\end{aligned}$$

so

$$\|\mathbf{B}\| \le 2. \tag{2.4}$$

Consequently, noting $\|\mathbf{G}\| = 2$ and the block diagonal structure of $\boldsymbol{\Lambda}$, we have

$$\|\boldsymbol{\Lambda}\| = \max\{\|\mathbf{B}\|, \|\mathbf{G}\|\} = 2. \tag{2.5}$$

Second, it is straightforward to verify that

$$\mathbf{B}^{-1} = \begin{bmatrix} & & 1 \\ & 1 & 1 \\ \iddots & \vdots & \vdots \\ 1 & \cdots & 1 & 1 \end{bmatrix}. \tag{2.6}$$

Krylov subspaces We study two Krylov subspaces that are associated with the matrix $\boldsymbol{\Lambda}$ and vector $\mathbf{c}$ described in (2.2). In particular, we consider the Krylov subspaces

$$\begin{aligned}\mathcal{J}_i &:= \text{span}\left\{\mathbf{c}, (\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)\mathbf{c}, (\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^2\mathbf{c}, \ldots, (\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^i\mathbf{c}\right\} \subseteq \mathbb{R}^m \\ &\text{and } \mathcal{K}_i := \boldsymbol{\Lambda}^\top \mathcal{J}_i \subseteq \mathbb{R}^n, \quad \text{for } i \ge 0.\end{aligned} \tag{2.7}$$

As shown below in (2.17), restricting on the first $2k$ entries, the above two Krylov subspaces reduce to

$$\mathcal{F}_i := \text{span}\left\{\mathbf{1}_{2k}, \mathbf{B}^2\mathbf{1}_{2k}, \ldots, \mathbf{B}^{2i}\mathbf{1}_{2k}\right\} \text{ and } \mathcal{R}_i := \text{span}\left\{\mathbf{B}\mathbf{1}_{2k}, \ldots, \mathbf{B}^{2i+1}\mathbf{1}_{2k}\right\}. \tag{2.8}$$

We first establish some important properties of $\mathcal{F}_i$ and $\mathcal{R}_i$ as follows.

Lemma 2.1 *Let $\mathcal{F}_i$ and $\mathcal{R}_i$ be defined in (2.8). For any $0 \le i \le 2k-1$, we have*

$$\mathcal{F}_i = \text{span}\{\mathbf{1}_{2k}, \mathbf{e}_{1,2k}, \mathbf{e}_{2,2k}, \ldots, \mathbf{e}_{i,2k}\}, \quad \mathcal{R}_i = \text{span}\{\mathbf{e}_{2k-i,2k}, \mathbf{e}_{2k-i+1,2k}, \ldots, \mathbf{e}_{2k,2k}\}, \tag{2.9}$$

and

$$\mathbf{B}\mathcal{R}_i = \text{span}\{\mathbf{e}_{1,2k}, \mathbf{e}_{2,2k}, \ldots, \mathbf{e}_{i+1,2k}\} \subseteq \mathcal{F}_{i+1}, \tag{2.10}$$

where we have used the convention $\mathbf{e}_{0,2k} = \mathbf{0}$.

Proof From the definition of $\mathbf{B}$ in (2.3), we have

$$\begin{aligned} \mathbf{B}\mathbf{1}_{2k} &= \mathbf{e}_{2k,2k}, \mathbf{B}\mathbf{e}_{2k,2k} = \mathbf{e}_{1,2k}, \\ \mathbf{B}\mathbf{e}_{i,2k} &= \mathbf{e}_{2k-i+1,2k} - \mathbf{e}_{2k-i,2k}, \ \forall i = 1, \ldots, 2k-1. \end{aligned} \tag{2.11}$$

Hence, from (2.8) and (2.11), it holds that

$$\begin{aligned} \mathcal{F}_0 &= \text{span}\{\mathbf{1}_{2k}\}, \\ \mathcal{R}_0 &= \text{span}\{\mathbf{B}\mathbf{1}_{2k}\} = \text{span}\{\mathbf{e}_{2k,2k}\}, \\ \mathbf{B}\mathcal{R}_0 &= \text{span}\{\mathbf{B}\mathbf{e}_{2k,2k}\} = \text{span}\{\mathbf{e}_{1,2k}\}, \\ \mathcal{F}_1 &= \text{span}\{\mathbf{1}_{2k}, \mathbf{B}^2\mathbf{1}_{2k}\} = \text{span}\{\mathbf{1}_{2k}, \mathbf{e}_{1,2k}\}, \\ \mathcal{R}_1 &= \text{span}\{\mathbf{B}\mathbf{1}_{2k}, \mathbf{B}^3\mathbf{1}_{2k}\} = \text{span}\{\mathbf{e}_{2k,2k}, \mathbf{B}\mathbf{e}_{1,2k}\} = \text{span}\{\mathbf{e}_{2k-1,2k}, \mathbf{e}_{2k,2k}\}, \\ \mathbf{B}\mathcal{R}_1 &= \text{span}\{\mathbf{B}^2\mathbf{1}_{2k}, \mathbf{B}^4\mathbf{1}_{2k}\} = \text{span}\{\mathbf{e}_{1,2k}, \mathbf{B}^2\mathbf{e}_{1,2k}\} = \text{span}\{\mathbf{e}_{1,2k}, \mathbf{e}_{2,2k}\}. \end{aligned}$$

Therefore, the results in (2.9) and (2.10) hold for $i = 0$ and $i = 1$.

Below we prove the results by induction. Assume that there is a positive integer $s < 2k$ and (2.9) holds for $i = s - 1$, namely,

$$\mathcal{F}_{s-1} = \text{span}\{\mathbf{1}_{2k}, \mathbf{e}_{1,2k}, \mathbf{e}_{2,2k}, \ldots, \mathbf{e}_{s-1,2k}\}, \ \mathcal{R}_{s-1} = \text{span}\{\mathbf{e}_{2k-s+1,2k}, \ldots, \mathbf{e}_{2k,2k}\}. \tag{2.12}$$

From (2.11) and (2.12), it follows that

$$\mathbf{B}\mathcal{R}_{s-1} = \mathbf{B}\,\text{span}\{\mathbf{e}_{2k-s+1,2k}, \ldots, \mathbf{e}_{2k,2k}\} \subseteq \text{span}\{\mathbf{e}_{s,2k}, \mathbf{e}_{s-1,2k}, \ldots, \mathbf{e}_{1,2k}\}. \tag{2.13}$$

Since $\mathbf{B}$ is nonsingular, $\dim\left(\mathbf{B}\mathcal{R}_{s-1}\right) = \dim\left(\mathcal{R}_{s-1}\right) = s$. Hence, from (2.13) and also noting

$$\dim\left(\text{span}\{\mathbf{e}_{s,2k}, \mathbf{e}_{s-1,2k}, \ldots, \mathbf{e}_{1,2k}\}\right) = s,$$

we have

$$\mathbf{B}\mathcal{R}_{s-1} = \text{span}\{\mathbf{e}_{s,2k}, \mathbf{e}_{s-1,2k}, \ldots, \mathbf{e}_{1,2k}\}. \tag{2.14}$$

Observing $\text{span}\{\mathbf{B}^2\mathbf{1}_{2k}, \ldots, \mathbf{B}^{2s}\mathbf{1}_{2k}\} = \mathbf{B}\mathcal{R}_{s-1}$, we have

$$\mathcal{F}_s = \text{span}\{\mathbf{1}_{2k}, \mathbf{B}^2\mathbf{1}_{2k}, \ldots, \mathbf{B}^{2s}\mathbf{1}_{2k}\} = \text{span}\{\mathbf{1}_{2k}, \mathbf{e}_{1,2k}, \mathbf{e}_{2,2k}, \ldots, \mathbf{e}_{s,2k}\},$$

and thus by (2.14), it follows that $\mathbf{B}\mathcal{R}_{s-1} \subseteq \mathcal{F}_s$. Through essentially the same arguments, one can use (2.11), the above equation, and the fact $\mathcal{R}_s = \mathbf{B}\mathcal{F}_s$ to conclude $\mathcal{R}_s = \text{span}\{\mathbf{e}_{2k-s,2k}, \ldots, \mathbf{e}_{2k,2k}\}$, and thus we complete the proof. □

Through relating $\mathcal{J}_i$ (resp. $\mathcal{K}_i$) to $\mathcal{F}_i$ (resp. $\mathcal{R}_i$), we have the following result.

Lemma 2.2 *Let $\mathcal{J}_i$ and $\mathcal{K}_i$ be defined in (2.7). For any $0 \le i \le 2k-1$, it holds*

$$\mathcal{J}_i = \text{span}\{\mathbf{c}, \mathbf{e}_{1,m}, \mathbf{e}_{2,m}, \ldots, \mathbf{e}_{i,m}\},\ \mathcal{K}_i = \text{span}\{\mathbf{e}_{2k-i,n}, \mathbf{e}_{2k-i+1,n}, \ldots, \mathbf{e}_{2k,n}\}, \tag{2.15}$$

and

$$\boldsymbol{\Lambda}\mathcal{K}_i = \text{span}\{\mathbf{e}_{1,m}, \mathbf{e}_{2,m}, \ldots, \mathbf{e}_{i+1,m}\} \subseteq \mathcal{J}_{i+1}. \tag{2.16}$$

Proof Observe that for any $i = 0, \ldots, 2k-1$ we have

$$(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^i \mathbf{c} = \begin{bmatrix} \mathbf{B}^{2i}\mathbf{1}_{2k} \\ \mathbf{0}_{m-2k} \end{bmatrix} \text{ and } \boldsymbol{\Lambda}^\top(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^i \mathbf{c} = \begin{bmatrix} \mathbf{B}^{2i+1}\mathbf{1}_{2k} \\ \mathbf{0}_{n-2k} \end{bmatrix}. \tag{2.17}$$

Consequently, the definitions in (2.7) becomes

$$\mathcal{J}_i = \mathcal{F}_i \times \{\mathbf{0}_{m-2k}\} \text{ and } \mathcal{K}_i = \mathcal{R}_i \times \{\mathbf{0}_{n-2k}\}.$$

Therefore, the results in (2.15) and (2.16) immediately follow from Lemma 2.1. □

Two remarks are in place for the Krylov subspaces $\mathcal{K}_i$ and $\mathcal{J}_i$. First, by the definitions of $\mathcal{K}_i$ and $\mathcal{J}_i$ in (2.7) and the relation (2.16), we have

$$\boldsymbol{\Lambda}\mathcal{K}_i \subseteq \mathcal{J}_{i+1} \text{ and } \boldsymbol{\Lambda}^\top \mathcal{J}_i = \mathcal{K}_i,\ \forall i = 1, \ldots, 2k-1. \tag{2.18}$$

Second, by (2.15) we have

$$\mathcal{K}_{i-1} \subsetneq \mathcal{K}_i \text{ and } \mathcal{J}_{i-1} \subsetneq \mathcal{J}_i,\ \forall i = 1, \ldots, 2k-1. \tag{2.19}$$

An important lemma We conclude this subsection by showing a lemma that will be used a few times in our analysis. It specifies the conditions on (2.1) to guarantee that any iterate sequence $\{\mathbf{x}^{(t)}\}_{t=1}^k$ satisfying Assumption 2.1 is in the subspace $\mathcal{K}_{k-1}$.

Lemma 2.3 *Let $\boldsymbol{\Lambda}$ and $\mathbf{c}$ be given in (2.2). Given any $L_A \in \mathbb{R}$, let*

$$\mathbf{A} = \frac{L_A}{2}\boldsymbol{\Lambda} \text{ and } \mathbf{b} = \frac{L_A}{2}\mathbf{c}. \tag{2.20}$$

Consider (2.1) with $\mathbf{A}$ and $\mathbf{b}$ defined as above, $\mathbf{h} \in \mathcal{K}_0$ and $\mathbf{H}$ satisfying $\mathbf{H}\mathcal{K}_{t-1} \subseteq \mathcal{K}_t$ for any $1 \le t \le k$, where $\mathcal{K}_i$ is defined in (2.7). Then under Assumption 2.1, we have $\mathbf{x}^{(t)} \in \mathcal{K}_{t-1}$ for any $1 \le t \le k$.

Proof It suffices to prove that for any $t = 1, \ldots, k$,

$$\text{span}\left\{\nabla f(\mathbf{x}^{(0)}), \mathbf{A}^\top \mathbf{r}^{(0)}, \nabla f(\mathbf{x}^{(1)}), \mathbf{A}^\top \mathbf{r}^{(1)}, \ldots, \nabla f(\mathbf{x}^{(t-1)}), \mathbf{A}^\top \mathbf{r}^{(t-1)}\right\} \subseteq \mathcal{K}_{t-1}. \tag{2.21}$$

We prove the result by induction. First, since $\mathbf{x}^{(0)} = \mathbf{0}$, from (2.2) and (2.20) we have $\mathbf{A}^\top \mathbf{r}^{(0)} = -\mathbf{A}^\top \mathbf{b} \in \text{span}\{\mathbf{e}_{2k,n}\} = \mathcal{K}_0$. In addition, from the condition $\mathbf{h} \in \mathcal{K}_0$, it follows that $\nabla f(\mathbf{x}^{(0)}) = -\mathbf{h} \in \mathcal{K}_0$. Therefore, (2.21) holds when $t = 1$. Assume that for a certain $1 \leq s < k$, (2.21) holds for $t = s$, and consequently

$$\mathbf{x}^{(s)} \in \mathcal{K}_{s-1}. \tag{2.22}$$

We go to prove the result in (2.21) for $t = s+1$, or equivalently $\nabla f(\mathbf{x}^{(s)}), \mathbf{A}^\top \mathbf{r}^{(s)} \in \mathcal{K}_s$, and finish the induction. From (2.19) we have $\mathcal{K}_0 \subseteq \mathcal{K}_s$. By this observation, noting $\mathbf{x}^{(s)} \in \mathcal{K}_{s-1}$, and using the conditions $\mathbf{h} \in \mathcal{K}_0$ and $\mathbf{H}\mathcal{K}_{s-1} \subseteq \mathcal{K}_s$, we have $\nabla f(\mathbf{x}^{(s)}) = \mathbf{H}\mathbf{x}^{(s)} - \mathbf{h} \in \mathcal{K}_s$. In addition, from (2.18) and (2.22), we have $\mathbf{A}^\top \mathbf{A}\mathbf{x}^{(s)} \in \mathcal{K}_s$. Since $\mathbf{A}^\top \mathbf{b} \in \mathcal{K}_0 \subseteq \mathcal{K}_s$, then $\mathbf{A}^\top \mathbf{r}^{(s)} = \mathbf{A}^\top \mathbf{A}\mathbf{x}^{(s)} - \mathbf{A}^\top \mathbf{b} \in \mathcal{K}_s$. Therefore, $\nabla f(\mathbf{x}^{(s)})$ and $\mathbf{A}^\top \mathbf{r}^{(s)}$ are both in $\mathcal{K}_s$, and by induction (2.21) holds for any $1 \leq t \leq k$. This completes the proof. □

2.2 A lower complexity bound for convex case

In this subsection, we establish a lower complexity bound of any first-order method that satisfies the linear span assumption (Assumption 2.1) on solving (2.1). Our approach is to build an instance such that the iterate $\mathbf{x}^{(t)} \in \mathcal{K}_{k-1}$, $\forall t \leq k$ and then estimate the values

$$\min_{\mathbf{x} \in \mathcal{K}_{k-1}} f(\mathbf{x}) - f^*, \text{ and } \min_{\mathbf{x} \in \mathcal{K}_{k-1}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|. \tag{2.23}$$

In the above equation, the former value is used to measure the performance of an algorithm by the objective value difference and the latter by the feasibility error. The instance we construct is in the form of (2.1) with

$$\mathbf{H} = \frac{L_f}{4} \begin{bmatrix} \mathbf{B}^\top \mathbf{B} & \\ & \mathbf{I}_{n-2k} \end{bmatrix} \in \mathbb{R}^{n \times n}, \ \mathbf{h} = \frac{L_f}{2} \mathbf{e}_{2k,n}, \ \mathbf{A} = \frac{L_A}{2} \boldsymbol{\Lambda}, \ \mathbf{b} = \frac{L_A}{2} \mathbf{c}, \tag{2.24}$$

where L_f and L_A are given nonnegative numbers, $\mathbf{B}$, $\boldsymbol{\Lambda}$ and $\mathbf{c}$ are those given in (2.2) and (2.3). From (2.4), (2.5), and the block diagonal structure of $\mathbf{H}$ above, we have $\|\mathbf{H}\| = (L_f/4)\|\mathbf{B}\|^2 \leq L_f$ and $\|\mathbf{A}\| = L_A$. Therefore, (2.1) with data specified in (2.24) provides an instance of (1.6) whose objective function f is L_f-smooth.

To establish the lower complexity bound, we first present three technical lemmas. We show in Lemma 2.4 below that under Assumption 2.1, the iterates generated by a first-order method on solving the designed instance would satisfy $\mathbf{x}^{(t)} \in \mathcal{K}_{t-1}$ for any $1 \leq t \leq k$. In Lemma 2.5, we give a pair of optimal primal-dual solution and also the optimal objective value of the instance. Then, in Lemma 2.6, we provide lower bounds of the values in (2.23).

Lemma 2.4 *Consider the instance of* (2.1) *with data described in* (2.24). *Under Assumption 2.1, we have* $\mathbf{x}^{(t)} \in \mathcal{K}_{t-1}$ *for any* $1 \leq t \leq k$, *where* $\mathcal{K}_{t-1}$ *is defined in* (2.7).

Proof To prove the lemma, it suffices to verify that $\mathbf{h} \in \mathcal{K}_0$ and $\mathbf{H}\mathcal{K}_{t-1} \subseteq \mathcal{K}_t$ for any $1 \leq t \leq k$ and then apply Lemma 2.3. Since $\mathbf{h}$ is a multiple of $\mathbf{e}_{2k,n}$, from (2.15) we immediately have $\mathbf{h} \in \mathcal{K}_0$. Using the definition of $\mathbf{H}$ and the second line of equation in (2.11), one can easily verify that $\mathbf{H}\,\mathrm{span}\{\mathbf{e}_{2k-t+1,n}, \ldots, \mathbf{e}_{2k,n}\} = \mathrm{span}\{\mathbf{e}_{2k-t,n}, \mathbf{e}_{2k-t+1,n}, \ldots, \mathbf{e}_{2k,n}\}$ for any $1 \leq t \leq k$. Hence we have all the conditions required by Lemma 2.3, and thus $\mathbf{x}^{(t)} \in \mathcal{K}_{t-1}$, which completes the proof. □

The next lemma gives the primal-dual solution and optimal objective value of the considered instance.

Lemma 2.5 *Let $L_f > 0$ and $L_A > 0$ be given. The instance of (2.1) with data given in (2.24) has a unique optimal solution $\mathbf{x}^*$ with a unique associated Lagrange multiplier $\mathbf{y}^*$ given by*

$$\mathbf{x}^* = (1; 2; \cdots; 2k; \mathbf{0}_{n-2k}), \quad \mathbf{y}^* = -\frac{L_f}{2L_A}(\mathbf{1}_{2k}; \mathbf{0}_{m-2k}). \tag{2.25}$$

In addition, the optimal objective value is

$$f^* = -\frac{3kL_f}{4}, \tag{2.26}$$

and the norm of the dual solution is

$$\|\mathbf{y}^*\| = \frac{L_f}{2L_A}\sqrt{2k}. \tag{2.27}$$

Proof We split $\mathbf{x}$ into two parts as $\mathbf{x} = (\mathbf{u}; \mathbf{v})$ with $\mathbf{u} \in \mathbb{R}^{2k}$ and $\mathbf{v} \in \mathbb{R}^{n-2k}$. Then from the block structure of $\mathbf{H}$ and $\mathbf{A}$ in (2.24), we obtain the following two optimization problems with respect to $\mathbf{u}$ and $\mathbf{v}$:

$$\min_{\mathbf{u}} \frac{1}{2}\mathbf{u}^\top \mathbf{S}\mathbf{u} - \mathbf{s}^\top \mathbf{u}, \ \text{s.t.} \ \frac{L_A}{2}\mathbf{B}\mathbf{u} = \frac{L_A}{2}\mathbf{1}_{2k}, \tag{2.28}$$

$$\min_{\mathbf{v}} \frac{L_f}{8}\|\mathbf{v}\|^2, \ \text{s.t.} \ \frac{L_A}{2}\mathbf{G}\mathbf{v} = \mathbf{0}, \tag{2.29}$$

where

$$\mathbf{S} = \frac{L_f}{4}\mathbf{B}^\top\mathbf{B} \text{ and } \mathbf{s} = \frac{L_f}{2}\mathbf{e}_{2k,2k}.$$

Since $L_A > 0$ and $\mathbf{B}$ is nonsingular, we have that $\mathbf{u}^* = (1; 2; \cdots; 2k)$ is the unique feasible and thus optimal solution of (2.28). In addition, since $L_f > 0$, (2.29) clearly has a unique solution $\mathbf{v}^* = \mathbf{0}$. Hence, $\mathbf{x}^*$ is unique and given in (2.25). Consequently,

$$f^* = \frac{1}{2}(\mathbf{u}^*)^\top \mathbf{S}\mathbf{u}^* - \mathbf{s}^\top \mathbf{u}^* = \frac{L_f}{8}\|\mathbf{B}\mathbf{u}^*\|^2 - \mathbf{s}^\top \mathbf{u}^* = -\frac{3kL_f}{4}.$$

To derive the corresponding dual variable, we split $\mathbf{y} = (\boldsymbol{\lambda}; \boldsymbol{\pi})$ with $\boldsymbol{\lambda} \in \mathbb{R}^{2k}$ and $\boldsymbol{\pi} \in \mathbb{R}^{m-2k}$. It follows from the KKT conditions of (2.28) that

$$\frac{L_A}{2}\mathbf{B}^\top \boldsymbol{\lambda}^* = \mathbf{S}\mathbf{u}^* - \mathbf{s}, \quad \frac{L_A}{2}\mathbf{G}^\top \boldsymbol{\pi}^* = \frac{L_f}{4}\mathbf{v}^* = \mathbf{0}.$$

Since $\mathbf{G}$ has full row rank, we have $\boldsymbol{\pi}^* = \mathbf{0}$. In addition, noting the description of $\mathbf{B}^{-1}$ in (2.6), we have

$$\boldsymbol{\lambda}^* = \frac{2}{L_A}\left(\mathbf{B}^\top\right)^{-1}(\mathbf{S}\mathbf{u}^* - \mathbf{s}) = -\frac{L_f}{2L_A}\mathbf{1}_{2k}.$$

Therefore, (2.25) follows immediately, and it is straightforward to have (2.27). □

Using Lemma 2.5, we have the following estimate.

Lemma 2.6 *Let $L_f > 0$ and $L_A > 0$ be given. Then for the instance of (2.1) with data given in (2.24), we have*

$$\min_{\mathbf{x}\in\mathcal{K}_{k-1}} f(\mathbf{x}) - f^* = \frac{kL_f}{4}, \tag{2.30a}$$

$$\min_{\mathbf{x}\in\mathcal{K}_{k-1}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\| \geq \frac{\sqrt{3}L_A\|\mathbf{x}^*\|}{4\sqrt{2}(k+1)}, \tag{2.30b}$$

where $\mathbf{x}^*$ is given in (2.25), and $\mathcal{K}_{k-1}$ is defined in (2.7).

Proof Using the formula

$$\sum_{i=1}^{p} i^2 = \frac{p(p+1)(2p+1)}{6}, \ \forall\, p \in \mathbb{Z}_{++}, \tag{2.31}$$

and the description of $\mathbf{x}^*$ in (2.25), we have

$$\|\mathbf{x}^*\|^2 = \sum_{i=1}^{2k} i^2 = \frac{k(2k+1)(4k+1)}{3}. \tag{2.32}$$

For any $\mathbf{x} \in \mathcal{K}_{k-1}$, we observe from (2.2), (2.16) and (2.20) that $\mathbf{A}\mathbf{x}$ can only have nonzeros on its first k components. Since the first $2k$ components of $\mathbf{b}$ all equal $\frac{L_A}{2}$, we have

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \geq \frac{L_A^2}{4}k \overset{(2.32)}{=} \frac{3L_A^2\|\mathbf{x}^*\|^2}{4(2k+1)(4k+1)} \geq \frac{3L_A^2\|\mathbf{x}^*\|^2}{32(k+1)^2}, \tag{2.33}$$

and hence (2.30b) holds.

To prove (2.30a), we need to compute the minimal objective value of $f(\mathbf{x})$ over $\mathcal{K}_{k-1}$. By (2.15) we have $\mathcal{K}_{k-1} = \mathrm{span}\{\mathbf{e}_{k+1,n}, \ldots, \mathbf{e}_{2k,n}\}$. Hence, for any $\mathbf{x} \in \mathcal{K}_{k-1}$,

we can write it as $\mathbf{x} = (\mathbf{0}_k; \mathbf{z}; \mathbf{0}_{n-2k})$ where $\mathbf{z} \in \mathbb{R}^k$. Recalling (2.24), we have

$$\mathbf{h}^\top \mathbf{x} = \frac{L_f}{2} \mathbf{e}_{2k,n}^\top \mathbf{x} = \frac{L_f}{2} z_k, \ \mathbf{x}^\top \mathbf{H} \mathbf{x} = \frac{L_f}{4} \|\mathbf{B}(\mathbf{0}_k; \mathbf{z})\|^2 = \frac{L_f}{4} \|\bar{\mathbf{B}}\mathbf{z}\|^2,$$

where

$$\bar{\mathbf{B}} := \begin{bmatrix} & & & -1 & 1 \\ & & \iddots & \iddots & \\ & -1 & 1 & & \\ -1 & 1 & & & \\ 1 & & & & \end{bmatrix} \in \mathbb{R}^{k \times k}$$

is a $k \times k$ submatrix of $\mathbf{B}$. Therefore,

$$\min_{\mathbf{x} \in \mathcal{K}_{k-1}} f(\mathbf{x}) = \min_{\mathbf{z} \in \mathbb{R}^k} \frac{L_f}{8} \|\bar{\mathbf{B}}\mathbf{z}\|^2 - \frac{L_f}{2} z_k. \tag{2.34}$$

Let $\mathbf{z}^*$ be the optimal solution to the right hand side minimization problem in (2.34). Then it must satisfy the optimality condition: $\frac{L_f}{4} \bar{\mathbf{B}}^2 \mathbf{z}^* = \frac{L_f}{2} \mathbf{e}_{k,k}$, which has the unique solution $\mathbf{z}^* = 2(1; \cdots; k)$. Plugging $\mathbf{z} = \mathbf{z}^*$ into the right hand side of (2.34) yields

$$\min_{\mathbf{x} \in \mathcal{K}_{k-1}} f(\mathbf{x}) = -\frac{k L_f}{2}. \tag{2.35}$$

From the above result and (2.26), we have (2.30a) and complete the proof. □

Using Lemmas 2.4 through 2.6, we are ready to establish the following lower complexity bound results.

Theorem 2.1 (Lower complexity bound for convex case under linear span assumption) Let $m \leq n$ be positive integers, $L_f > 0$, and $L_A > 0$. For any positive integer $t < \frac{m}{2}$, there exists an instance of (1.6) such that f is L_f-smooth, $\|\mathbf{A}\| = L_A$, and it has a unique primal-dual solution $(\mathbf{x}^*, \mathbf{y}^*)$. In addition, on solving (1.6), if the algorithm satisfies Assumption 2.1, then

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \geq \frac{3 L_f \|\mathbf{x}^*\|^2}{64(t+1)^2} + \frac{\sqrt{3} L_A \|\mathbf{x}^*\| \cdot \|\mathbf{y}^*\|}{16(t+1)}, \tag{2.36a}$$

$$\|\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\| \geq \frac{\sqrt{3} L_A \|\mathbf{x}^*\|}{4\sqrt{2}(t+1)}. \tag{2.36b}$$

Proof Set $k = t < \frac{m}{2}$ and consider the instance (2.1) with data given in (2.24). Clearly, this instance is in the form of (1.6), f is L_f-smooth, and $\|\mathbf{A}\| = L_A$.

Lemma 2.5 indicates that the considered instance has a unique primal-dual solution $(\mathbf{x}^*, \mathbf{y}^*)$ given in (2.25). By Lemma 2.4 and noting $t = k$, we have $\mathbf{x}^{(t)} \in \mathcal{K}_{k-1}$.

Consequently,

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \geq \min_{\mathbf{x}\in\mathcal{K}_{k-1}} f(\mathbf{x}) - f^*, \text{ and } \|\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\| \geq \min_{\mathbf{x}\in\mathcal{K}_{k-1}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|.$$

From (2.30a), (2.32) and (2.27), it follows that

$$\min_{\mathbf{x}\in\mathcal{K}_{k-1}} f(\mathbf{x}) - f^* = \frac{kL_f}{8} + \frac{kL_f}{8} = \frac{3L_f\|\mathbf{x}^*\|^2}{8(2k+1)(4k+1)} + \frac{\sqrt{3}L_A\|\mathbf{x}^*\|\cdot\|\mathbf{y}^*\|}{4\sqrt{2}\sqrt{(2k+1)(4k+1)}}.$$

Since $k = t$, we conclude (2.36a) from the above relation, and (2.36b) from (2.30b). □

***Remark* 2.1** The norm $\|\mathbf{y}^*\|$ in (2.27) depends on the ratio $\frac{L_f}{L_A}$. With the assumption $L_f \geq L_A$, we can remove such a dependence. In particular, setting $\mathbf{h} = \left(\frac{L_f}{4} + \frac{L_A}{4\sqrt{2}}\right)\mathbf{e}_{2k,n}$ in (2.24), we can obtain that

$$\|\mathbf{y}^*\| = \frac{\sqrt{k}}{2}, \text{ and } \min_{\mathbf{x}\in\mathcal{K}_{k-1}} f(\mathbf{x}) - f^* = \frac{L_f}{16}k + \frac{\sqrt{2}L_A}{8}k + \frac{L_f^2 - L_A^2}{16L_f}k.$$

Hence, assuming $L_f \geq L_A$ and taking $k = t$ we have

$$\begin{aligned} f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) &\geq \min_{\mathbf{x}\in\mathcal{K}_{t-1}} f(\mathbf{x}) - f^* \geq \frac{L_f}{16}t + \frac{\sqrt{2}L_A}{8}t \\ &= \frac{3L_f\|\mathbf{x}^*\|^2}{16(2t+1)(4t+1)} + \frac{\sqrt{6}L_A\|\mathbf{x}^*\|\cdot\|\mathbf{y}^*\|}{4\sqrt{(2t+1)(4t+1)}}. \end{aligned}$$

The proof of the above claim follows the same lines of arguments throughout this subsection. We do not repeat it here but leave the details to interested readers.

2.3 A lower complexity bound for strongly convex case

In this subsection, we develop a lower complexity bound for solving (1.6) when f is μ-strongly convex. The measure we use is different from those in (2.36). Instead of bounding the objective and feasibility error, we directly bound the distance of generated iterate to the unique optimal solution. Similar to the previous subsection, the "hard" instance we design is also a quadratic program in the form of (2.1). The following theorem summarizes our result.

Theorem 2.2 (Lower complexity bound for strongly convex case under linear span assumption) Let $m \leq n$ be positive integers, $\mu > 0$, and $L_A > 0$. For any positive integer $t < \frac{m}{2}$, there exists an instance of (1.6) such that f is differentiable and μ-strongly convex, $\|\mathbf{A}\| = L_A$, and it has a unique primal-dual solution pair $(\mathbf{x}^*, \mathbf{y}^*)$. In

addition, for any algorithm on solving (1.6), if it satisfies Assumption 2.1, then

$$\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2 \geq \frac{5L_A^2\|\mathbf{y}^*\|^2}{256\mu^2(t+1)^2}.$$

Proof Set $k = t$ and consider an instance of (2.1) with $\mathbf{H} = \mu\mathbf{I}$, $\mathbf{h} = \mathbf{0}$, and $\mathbf{A}$ and $\mathbf{b}$ given in (2.20). Clearly, f is differentiable and μ-strongly convex, and $\|\mathbf{A}\| = L_A$. It is easy to verify that Lemma 2.3 applies to this instance, and thus $\mathbf{x}^{(t)} \in \mathcal{K}_{t-1}$. Also, by the KKT condition $\mu\mathbf{x}^* = \mathbf{A}^\top\mathbf{y}^*$, we can easily verify that the system has a unique primal-dual solution $(\mathbf{x}^*, \mathbf{y}^*)$ with $\mathbf{x}^*$ given in (2.25) and $\mathbf{y}^*$ given by

$$y_i^* = \begin{cases} \dfrac{\mu}{L_A} i(4k - i + 1), & \text{if } 1 \leq i \leq 2k, \\ 0, & \text{if } i \geq 2k+1. \end{cases} \tag{2.37}$$

From the formula of $\mathcal{K}_i$ in (2.15), it follows that for any $\mathbf{x} \in \mathcal{K}_{k-1}$,

$$\|\mathbf{x} - \mathbf{x}^*\|^2 \geq \sum_{i=1}^{k} i^2 \overset{(2.31)}{=} \frac{k(k+1)(2k+1)}{6}. \tag{2.38}$$

Moreover, by (2.31) and also the formulas

$$\sum_{i=1}^{p} i^3 = \frac{p^2(p+1)^2}{4}, \quad \sum_{i=1}^{p} i^4 = \frac{p(p+1)(2p+1)(3p^2+3p-1)}{30},$$

we have from (2.37) that

$$\begin{aligned} \|\mathbf{y}^*\|^2 &= \frac{\mu^2}{L_A^2} \sum_{i=1}^{2k} i^2(4k - i + 1)^2 \\ &= \frac{\mu^2}{L_A^2} \left((4k+1)^2 \sum_{i=1}^{2k} i^2 - 2(4k+1) \sum_{i=1}^{2k} i^3 + \sum_{i=1}^{2k} i^4 \right) \\ &= \frac{2k(2k+1)(4k+1)\mu^2}{L_A^2} \left(\frac{(4k+1)^2}{6} - k(2k+1) + \frac{12k^2+6k-1}{30} \right) \\ &= \frac{2k(2k+1)(4k+1)\mu^2}{15L_A^2}(16k^2+8k+2). \end{aligned}$$

Since $t = k$ and $\mathbf{x}^{(t)} \in \mathcal{K}_{t-1}$, it is not difficult to verify the desired result from (2.38) and the above equation, and thus we complete the proof. □

3 Lower complexity bounds of general deterministic first-order methods for affinely constrained problems

In this section, we drop the linear span assumption (see Assumption 2.1) and establish lower complexity bounds of general first-order methods described in (1.8) on solving (1.6). The key idea is to utilize certain rotational invariance of quadratic functions and linear systems, a technique that was introduced in [32,33]. Specifically, we first establish a key proposition (i.e., Proposition 3.1 below) as our main tool and then derive the lower complexity bounds by the results obtained in the previous section.

For ease of notation, we define a specific class of SPPs as follows.

Definition 3.1 *(A special class of SPPs)* Given $\mathbf{H} \in \mathbb{S}^n_+$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, and $\boldsymbol{\theta} = (\mathbf{h}, \mathbf{b}, R_X, R_Y, \lambda)$ where $R_X, R_Y \in [0, +\infty]$ and $\lambda \geq 0$, $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ is defined as one instance of (1.3) with

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x} - \mathbf{h}^\top \mathbf{x},\ g(\mathbf{y}) = \lambda\|\mathbf{y}\|,\ X = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq R_X\},$$
$$\text{and } Y = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y}\| \leq R_Y\}. \tag{3.1}$$

Hence, by $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ or more specifically $P\big((\mathbf{h}, \mathbf{b}, R_X, R_Y, \lambda); \mathbf{H}, \mathbf{A}\big)$, we mean the instance

$$\phi^* := \min_{\|\mathbf{x}\|\leq R_X} \left\{\phi(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x} - \mathbf{h}^\top \mathbf{x} + \max_{\|\mathbf{y}\|\leq R_Y} \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle - \lambda\|\mathbf{y}\|\right\}. \tag{3.2}$$

***Remark* 3.1** We will call $(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ as the data in the instance $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$. Given $\mathbf{H} \in \mathbb{S}^n_+$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{h} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$, then an instance of (2.1) can be denoted as $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ with $\boldsymbol{\theta} = (\mathbf{h}, \mathbf{b}, +\infty, +\infty, 0)$.

Proposition 3.1 *Let $m \leq n$, $k < \frac{m}{2}$, and $t \leq \frac{k}{2} - 1$ be positive integers, and let L_f and L_A be nonnegative numbers. Suppose that we have an instance $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$, called* original instance*, where $\|\mathbf{H}\| \leq L_f$, and $\mathbf{A}$ and $\mathbf{b}$ are those given in (2.20). Moreover, assume that $\mathbf{H} \in \mathbb{S}^n_+$ and satisfies $\mathbf{H}\mathcal{K}_{2s-1} \subseteq \mathcal{K}_{2s}$ for any $s \leq \frac{k}{2}$ and $\mathbf{h} \in \mathcal{K}_0$, where $\mathcal{K}_i$ is defined in (2.7). Then for any deterministic first-order method $\mathcal{M}$ that is described in (1.8), there exists another instance $P(\boldsymbol{\theta}; \tilde{\mathbf{H}}, \tilde{\mathbf{A}})$, called* rotated instance*, where $\tilde{\mathbf{H}} = \mathbf{U}^\top \mathbf{H}\mathbf{U}$, $\tilde{\mathbf{A}} = \mathbf{V}^\top \mathbf{A}\mathbf{V}$, $\mathbf{U}$ and $\mathbf{V}$ are certain orthogonal matrices dependent on t such that $\mathbf{U}\mathbf{h} = \mathbf{h}$ and $\mathbf{V}\mathbf{b} = \mathbf{b}$, and*

1. *In addition, $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point that satisfies (1.5) to the* original instance *if and only if $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) := (\mathbf{U}^\top \mathbf{x}^*, \mathbf{V}^\top \mathbf{y}^*)$ is a saddle point to the* rotated instance.
2. *Furthermore, when $\mathcal{M}$ is applied to solve $P(\boldsymbol{\theta}; \tilde{\mathbf{H}}, \tilde{\mathbf{A}})$, its t-th computed approximate solution $\bar{\mathbf{x}}^{(t)}$ satisfies*

$$\tilde{\phi}(\bar{\mathbf{x}}^{(t)}) - \tilde{\phi}^* \geq \min_{\mathbf{x}\in\mathcal{K}_{k-1}} \phi(\mathbf{x}) - \phi^*, \tag{3.3}$$

$$\tilde{f}(\bar{\mathbf{x}}^{(t)}) - \tilde{f}(\hat{\mathbf{x}}) \geq \min_{\mathbf{x}\in\mathcal{K}_{k-1}} f(\mathbf{x}) - f(\mathbf{x}^*), \tag{3.4}$$

$$\|\tilde{\mathbf{A}}\bar{\mathbf{x}}^{(t)} - \mathbf{b}\| \geq \min_{\mathbf{x}\in\mathcal{K}_{k-1}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|, \tag{3.5}$$

$$\|\bar{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}\|^2 \geq \min_{\mathbf{x}\in\mathcal{K}_{k-1}} \|\mathbf{x} - \mathbf{x}^*\|^2, \tag{3.6}$$

where ϕ and f are the functions in the original instance (see the definitions in (3.1) and (3.2)), and $\tilde{\phi}$ and $\tilde{f}$ are those in the rotated instance.

The proof of Proposition 3.1 is rather technical and deferred after we present the lower complexity bound results. Here we give a few remarks on this proposition. First, in Proposition 3.1 there are two problem instances, which have been distinguished as *original* and *rotated* instances, respectively. Second, the results in (3.3) through (3.6) establish an important relation between the original and rotated instances. Specifically, by this relation, we are able to study the best possible performance of general first-order methods through the linear subspace $\mathcal{K}_{k-1}$.

3.1 Lower complexity bounds

In this subsection, we apply Proposition 3.1 together with Theorems 2.1 and 2.2 to establish the lower complexity bounds of general first-order methods on solving (1.6). The theorem below extends the results in Theorem 2.1.

Theorem 3.1 (Lower complexity bound of general first-order methods) Let $8 < m \leq n$ be positive integers, $L_f > 0$, and $L_A > 0$. For any positive integer $t < \frac{m}{4} - 1$ and any deterministic first-order method $\mathcal{M}$ that is described in (1.8), there exists an instance of (1.6) such that f is L_f-smooth and $\|\mathbf{A}\| = L_A$. In addition, the instance has a unique primal-dual solution $(\mathbf{x}^*, \mathbf{y}^*)$, and

$$f(\bar{\mathbf{x}}^{(t)}) - f^* \geq \frac{3L_f\|\mathbf{x}^*\|^2}{64(2t+5)^2} + \frac{\sqrt{3}L_A\|\mathbf{x}^*\| \cdot \|\mathbf{y}^*\|}{16(2t+5)}, \tag{3.7a}$$

$$\|\mathbf{A}\bar{\mathbf{x}}^{(t)} - \mathbf{b}\| \geq \frac{\sqrt{3}L_A\|\mathbf{x}^*\|}{4\sqrt{2}(2t+5)}, \tag{3.7b}$$

where $\bar{\mathbf{x}}^{(t)}$ is the output by $\mathcal{M}$.

Proof Set $k = 2t + 2 < \frac{m}{2}$ in the definition of $\boldsymbol{A}$ and $\mathbf{c}$ given in (2.2). Consider (2.1) with data given in (2.24). By Remark 3.1, this problem instance is $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ with $\boldsymbol{\theta} = (\mathbf{h}, \mathbf{b}, +\infty, +\infty, 0)$. It is easy to check that the data satisfy the conditions required in Proposition 3.1. Hence, there exists a *rotated* instance $P(\boldsymbol{\theta}; \tilde{\mathbf{H}}, \tilde{\mathbf{A}})$, i.e.,

$$\tilde{f}^* := \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \tilde{f}(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top\tilde{\mathbf{H}}\mathbf{x} - \mathbf{h}^\top\mathbf{x}, \ \text{ s.t. } \ \tilde{\mathbf{A}}\mathbf{x} = \mathbf{b} \right\}, \tag{3.8}$$

where $\tilde{\mathbf{H}} = \mathbf{U}^\top\mathbf{H}\mathbf{U}$ and $\tilde{\mathbf{A}} = \mathbf{V}^\top\mathbf{A}\mathbf{V}$ with orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ dependent on t, and in addition (3.4) and (3.5) hold. From the proof of Theorem 2.1 together with

these two inequalities, we have

$$\begin{aligned} \tilde{f}(\bar{\mathbf{x}}^{(t)}) - \tilde{f}^* &\geq \min_{\mathbf{x}\in\mathcal{K}_{k-1}} f(\mathbf{x}) - f^* \geq \frac{3L_f\|\mathbf{x}^*\|^2}{64(k+1)^2} + \frac{\sqrt{3}L_A\|\mathbf{x}^*\|\cdot\|\mathbf{y}^*\|}{16(k+1)}, \\ \|\tilde{\mathbf{A}}\bar{\mathbf{x}}^{(t)} - \mathbf{b}\| &\geq \min_{\mathbf{x}\in\mathcal{K}_{k-1}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\| \geq \frac{\sqrt{3}L_A\|\mathbf{x}^*\|}{4\sqrt{2}(k+1)}, \end{aligned} \tag{3.9}$$

where $(\mathbf{x}^*, \mathbf{y}^*)$ is the unique primal-dual solution to the original instance. By item 1 of Proposition 3.1, the rotated instance (3.8) also has a unique primal-dual solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ given by $\hat{\mathbf{x}} = \mathbf{U}^\top\mathbf{x}^*$ and $\hat{\mathbf{y}} = \mathbf{V}^\top\mathbf{y}^*$. Since $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, it holds that $\|\mathbf{x}^*\| = \|\hat{\mathbf{x}}\|$ and $\|\mathbf{y}^*\| = \|\hat{\mathbf{y}}\|$. Therefore, noting that $k = 2t + 2$ and (3.8) is an instance of (1.6), we obtain the desired results from the two inequalities in (3.9) and abusing the notation $(f, \mathbf{A}, \mathbf{x}^*, \mathbf{y}^*)$ for $(\tilde{f}, \tilde{\mathbf{A}}, \hat{\mathbf{x}}, \hat{\mathbf{y}})$. □

For strongly convex case, we below generalize Theorem 2.2 to any first-order method given in (1.8).

Theorem 3.2 (Lower complexity bound of general first-order methods for strongly convex case) Let $8 < m \leq n$ be positive integers, and μ and L_A be positive numbers. For any positive integer $t < \frac{m}{4} - 1$ and any deterministic first-order method $\mathcal{M}$ that is described in (1.8), there exists an instance of (1.6) such that f is μ-strongly convex, and $\|\mathbf{A}\| = L_A$. In addition, the instance has a unique primal-dual solution $(\mathbf{x}^*, \mathbf{y}^*)$, and

$$\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 \geq \frac{5L_A^2\|\mathbf{y}^*\|^2}{256\mu^2(2t+5)^2}, \tag{3.10}$$

where $\bar{\mathbf{x}}^{(t)}$ is the output by $\mathcal{M}$.

The proof of Theorem 3.2 is similar to that of Theorem 3.1: one can use (3.6) together with Theorem 2.2. We omit the details.

3.2 Proof of Proposition 3.1

This subsection is dedicated to the technical details on the proof of Proposition 3.1. On an instance $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ defined in Definition 3.1, the first-order method $\mathcal{M}$ described in (1.8) can be written as

$$\left(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, \bar{\mathbf{x}}^{(t+1)}, \bar{\mathbf{y}}^{(t+1)}\right)$$
$$= \mathcal{I}_t\left(\boldsymbol{\theta}; \mathbf{H}\mathbf{x}^{(0)}, \mathbf{A}\mathbf{x}^{(0)}, \mathbf{A}^\top\mathbf{y}^{(0)}, \ldots, \mathbf{H}\mathbf{x}^{(t)}, \mathbf{A}\mathbf{x}^{(t)}, \mathbf{A}^\top\mathbf{y}^{(t)}\right), \ \forall\, t \geq 0. \tag{3.11}$$

We start our proof with several technical lemmas. The following lemma is an elementary result of linear subspaces and will be used several times in our analysis.

Lemma 3.1 *Let $\mathcal{X} \subsetneq \bar{\mathcal{X}} \subseteq \mathbb{R}^p$ be two linear subspaces. Then for any $\bar{\mathbf{x}} \in \mathbb{R}^p$, there exists an orthogonal matrix $\mathbf{V} \in \mathbb{R}^{p\times p}$ such that*

$$\mathbf{V}\mathbf{x} = \mathbf{x}, \ \forall \mathbf{x} \in \mathcal{X}, \ \textit{and}\ \mathbf{V}\bar{\mathbf{x}} \in \bar{\mathcal{X}}. \tag{3.12}$$

Proof If $\bar{\mathbf{x}} \in \mathcal{X}$, then we can simply choose $\mathbf{V} = \mathbf{I}$. Otherwise, we decompose $\bar{\mathbf{x}} = \mathbf{y} + \mathbf{z}$, where $\mathbf{z} \in \mathcal{X}$ and $\mathbf{y} \neq \mathbf{0}$ is in the complement subspace $\mathcal{X}^\perp$. Let $s = \dim(\mathcal{X})$ and $t = \dim(\bar{\mathcal{X}}) > s$. Assume $\mathbf{u}_1, \ldots, \mathbf{u}_s$ to be an orthonormal basis of $\mathcal{X}$. We extend it to $\mathbf{u}_1, \ldots, \mathbf{u}_t$, an orthonormal basis of $\bar{\mathcal{X}}$. The desired result in (3.12) is then obtained by choosing $\mathbf{V}$ as an orthogonal matrix such that $\mathbf{V}\mathbf{u}_i = \mathbf{u}_i$, $\forall\, i = 1, \ldots, s$, and $\mathbf{V}\mathbf{y} = \|\mathbf{y}\|\mathbf{u}_{s+1}$. □

By Lemma 3.1, we show the results below.

Lemma 3.2 *Given $m \leq n$ and $k < \frac{m}{2}$, let $\boldsymbol{\Lambda}$ be the matrix in (2.2). Let $s \leq \frac{k}{2}$ be a positive integer, $\mathbf{H} \in \mathbb{S}^n_+$, and $\mathbf{U}, \boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$ and $\mathbf{V}, \boldsymbol{\Psi} \in \mathbb{R}^{m\times m}$ be orthogonal matrices. If $\mathbf{H}\mathcal{K}_{2s-1} \subseteq \mathcal{K}_{2s}$, and*

$$\boldsymbol{\Phi}\mathbf{x} = \mathbf{x}, \forall\, \mathbf{x} \in \mathbf{U}^\top\mathcal{K}_{2s}, \ \text{and}\ \boldsymbol{\Psi}\mathbf{y} = \mathbf{y}, \ \forall\, \mathbf{y} \in \mathbf{V}^\top\mathcal{J}_{2s}, \tag{3.13}$$

then for any $\mathbf{x} \in \mathbf{U}^\top\mathcal{K}_{2s-1}$ and any $\mathbf{y} \in \mathbf{V}^\top\mathcal{J}_{2s-1}$, it holds:

$$\tilde{\mathbf{U}}^\top\mathbf{H}\tilde{\mathbf{U}}\mathbf{x} = \mathbf{U}^\top\mathbf{H}\mathbf{U}\mathbf{x}, \ \tilde{\mathbf{V}}^\top\boldsymbol{\Lambda}\tilde{\mathbf{U}}\mathbf{x} = \mathbf{V}^\top\boldsymbol{\Lambda}\mathbf{U}\mathbf{x}, \ \textit{and}\ \tilde{\mathbf{U}}^\top\boldsymbol{\Lambda}^\top\tilde{\mathbf{V}}\mathbf{y} = \mathbf{U}^\top\boldsymbol{\Lambda}^\top\mathbf{V}\mathbf{y},$$

where $\tilde{\mathbf{U}} = \mathbf{U}\boldsymbol{\Phi}$ and $\tilde{\mathbf{V}} = \mathbf{V}\boldsymbol{\Psi}$.

Proof Let $\mathbf{x} \in \mathbf{U}^\top\mathcal{K}_{2s-1}$ and $\mathbf{y} \in \mathbf{V}^\top\mathcal{J}_{2s-1}$. Since $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, it holds that $\mathbf{U}\mathbf{x} \in \mathcal{K}_{2s-1}$ and $\mathbf{V}\mathbf{y} \in \mathcal{J}_{2s-1}$. Hence, from the assumption on $\mathbf{H}$, the properties of $\mathcal{J}_i$ and $\mathcal{K}_i$ in (2.18) and (2.19), and noting $2s - 1 \leq k - 1$, we have

$$\mathbf{H}\mathbf{U}\mathbf{x} \in \mathcal{K}_{2s}, \ \boldsymbol{\Lambda}\mathbf{U}\mathbf{x} \in \mathcal{J}_{2s}, \ \text{and}\ \boldsymbol{\Lambda}^\top\mathbf{V}\mathbf{y} \in \mathcal{K}_{2s-1} \subsetneq \mathcal{K}_{2s},$$

which implies

$$\mathbf{U}^\top\mathbf{H}\mathbf{U}\mathbf{x} \in \mathbf{U}^\top\mathcal{K}_{2s}, \ \mathbf{V}^\top\boldsymbol{\Lambda}\mathbf{U}\mathbf{x} \in \mathbf{V}^\top\mathcal{J}_{2s}, \ \text{and}\ \mathbf{U}^\top\boldsymbol{\Lambda}^\top\mathbf{V}\mathbf{y} \in \mathbf{U}^\top\mathcal{K}_{2s}.$$

From (3.13), we obtain

$$\boldsymbol{\Phi}\mathbf{U}^\top\mathbf{H}\mathbf{U}\mathbf{x} = \mathbf{U}^\top\mathbf{H}\mathbf{U}\mathbf{x}, \ \boldsymbol{\Psi}\mathbf{V}^\top\boldsymbol{\Lambda}\mathbf{U}\mathbf{x} = \mathbf{V}^\top\boldsymbol{\Lambda}\mathbf{U}\mathbf{x}, \ \text{and}\ \boldsymbol{\Phi}\mathbf{U}^\top\boldsymbol{\Lambda}^\top\mathbf{V}\mathbf{y} = \mathbf{U}^\top\boldsymbol{\Lambda}^\top\mathbf{V}\mathbf{y}.$$

Because $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are orthogonal matrix, the above equations indicate that

$$\boldsymbol{\Phi}^\top \mathbf{U}^\top \mathbf{H}\mathbf{U}\mathbf{x} = \mathbf{U}^\top \mathbf{H}\mathbf{U}\mathbf{x}, \ \boldsymbol{\Psi}^\top \mathbf{V}^\top \boldsymbol{\Lambda}\mathbf{U}\mathbf{x} = \mathbf{V}^\top \boldsymbol{\Lambda}\mathbf{U}\mathbf{x}, \text{ and } \boldsymbol{\Phi}^\top \mathbf{U}^\top \boldsymbol{\Lambda}^\top \mathbf{V}\mathbf{y} = \mathbf{U}^\top \boldsymbol{\Lambda}^\top \mathbf{V}\mathbf{y}. \tag{3.14}$$

Moreover, since $\mathbf{x} \in \mathbf{U}^\top \mathcal{K}_{2s-1}$ and $\mathbf{y} \in \mathbf{V}^\top \mathcal{J}_{2s-1}$, it follows from (2.19) that $\mathbf{x} \in \mathbf{U}^\top \mathcal{K}_{2s}$ and $\mathbf{y} \in \mathbf{V}^\top \mathcal{J}_{2s}$, and thus using (3.13) again and also the definition of $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$, we have

$$\tilde{\mathbf{U}}\mathbf{x} = \mathbf{U}\boldsymbol{\Phi}\mathbf{x} = \mathbf{U}\mathbf{x}, \text{ and } \tilde{\mathbf{V}}\mathbf{y} = \mathbf{V}\boldsymbol{\Psi}\mathbf{y} = \mathbf{V}\mathbf{y}. \tag{3.15}$$

Therefore, we conclude that for any $\mathbf{x} \in \mathbf{U}^\top \mathcal{K}_{2s-1}$ and $\mathbf{y} \in \mathbf{V}^\top \mathcal{J}_{2s-1}$,

$$\begin{aligned}
\tilde{\mathbf{U}}^\top \mathbf{H}\tilde{\mathbf{U}}\mathbf{x} &\overset{(3.15)}{=} \boldsymbol{\Phi}^\top \mathbf{U}^\top \mathbf{H}\mathbf{U}\mathbf{x} \overset{(3.14)}{=} \mathbf{U}^\top \mathbf{H}\mathbf{U}\mathbf{x}, \\
\tilde{\mathbf{V}}^\top \boldsymbol{\Lambda}\tilde{\mathbf{U}}\mathbf{x} &\overset{(3.15)}{=} \boldsymbol{\Psi}^\top \mathbf{V}^\top \boldsymbol{\Lambda}\mathbf{U}\mathbf{x} \overset{(3.14)}{=} \mathbf{V}^\top \boldsymbol{\Lambda}\mathbf{U}\mathbf{x}, \\
\tilde{\mathbf{U}}^\top \boldsymbol{\Lambda}^\top \tilde{\mathbf{V}}\mathbf{y} &= \boldsymbol{\Phi}^\top \mathbf{U}^\top \boldsymbol{\Lambda}^\top \mathbf{V}\mathbf{y} \overset{(3.14)}{=} \mathbf{U}^\top \boldsymbol{\Lambda}^\top \mathbf{V}\mathbf{y}.
\end{aligned}$$

Hence, we complete the proof. □

Proposition 3.2 *Given $m \le n$ and $k < \frac{m}{2}$, let $\boldsymbol{\Lambda}$ and $\mathbf{c}$ be the matrix and vector in (2.2), and let $\mathbf{h} \in \mathcal{K}_0$, R_X, $R_Y \in [0, +\infty]$, and $\lambda \ge 0$. Suppose that $\mathbf{A}$ and $\mathbf{b}$ are respectively a multiple of $\boldsymbol{\Lambda}$ and $\mathbf{c}$ and $\mathbf{H} \in \mathbb{S}^n_+$ satisfying $\mathbf{H}\mathcal{K}_{2s-1} \subseteq \mathcal{K}_{2s}$ for all $s \le \frac{k}{2}$. Then for any $0 \le t \le \frac{k}{2} - 1$ and any deterministic first-order method $\mathcal{M}$ described in (1.8), there exist orthogonal matrices $\mathbf{U}_t \in \mathbb{R}^{n\times n}$ and $\mathbf{V}_t \in \mathbb{R}^{m\times m}$ and a problem instance $P(\boldsymbol{\theta}; \mathbf{U}_t^\top \mathbf{H}\mathbf{U}_t, \mathbf{V}_t^\top \mathbf{A}\mathbf{U}_t)$ with $\boldsymbol{\theta} = (\mathbf{h}, \mathbf{b}, R_X, R_Y, \lambda)$ such that $\mathbf{U}_t\mathbf{h} = \mathbf{h}$, $\mathbf{V}_t\mathbf{c} = \mathbf{c}$, and in addition, when $\mathcal{M}$ is applied to solve the instance, the iterates $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=0}^t$ satisfy*

$$\mathbf{x}^{(i)} \in \mathbf{U}_t^\top \mathcal{K}_{2t+1}, \ \mathbf{y}^{(i)} \in \mathbf{V}_t^\top \mathcal{J}_{2t+1}, \ \forall\, i = 0, \ldots, t,$$

where $\mathcal{K}_{2t+1}$ and $\mathcal{J}_{2t+1}$ are the Krylov subspaces defined in (2.7). Moreover, the output $\bar{\mathbf{x}}^{(t)} \in \mathbf{U}_t^\top \mathcal{K}_{2t+1}$.

Proof Note $\mathcal{K}_0 \subsetneq \mathcal{K}_1$ and $\mathcal{J}_0 \subsetneq \mathcal{J}_1$ from Lemma 2.2. Hence, by Lemma 3.1 there exist orthogonal matrices $\mathbf{U}_0$ and $\mathbf{V}_0$ such that

$$\begin{aligned}
\mathbf{U}_0\mathbf{x} = \mathbf{x}, \forall \mathbf{x} \in \mathcal{K}_0, \text{ and } \mathbf{U}_0\mathbf{x}^{(0)} \in \mathcal{K}_1 \\
\mathbf{V}_0\mathbf{y} = \mathbf{y}, \forall \mathbf{y} \in \mathcal{J}_0, \text{ and } \mathbf{V}_0\mathbf{y}^{(0)} \in \mathcal{J}_1.
\end{aligned}$$

Therefore, from the condition $\mathbf{h} \in \mathcal{K}_0$ and $\mathbf{c} \in \mathcal{J}_0$ by Lemma 2.2, we have $\mathbf{U}_0\mathbf{h} = \mathbf{h}$ and $\mathbf{V}_0\mathbf{c} = \mathbf{c}$. Consequently, the results in the lemma hold for $t = 0$. Below we prove the results for any $t < \frac{k}{2} - 1$ by induction.

Assume that for some $1 \le s < \frac{k}{2} - 1$, the results hold for $t = s - 1$, namely, there exist orthogonal matrices $\mathbf{U}_{s-1} \in \mathbb{R}^{n\times n}$ and $\mathbf{V}_{s-1} \in \mathbb{R}^{m\times m}$ such that $\mathbf{U}_{s-1}\mathbf{h} = \mathbf{h}$,

$\mathbf{V}_{s-1}\mathbf{c} = \mathbf{c}$, and when $\mathcal{M}$ is applied to the instance $P(\boldsymbol{\theta}; \mathbf{U}_{s-1}^\top \mathbf{H}\mathbf{U}_{s-1}, \mathbf{V}_{s-1}^\top \mathbf{A}\mathbf{U}_{s-1})$, the iterates $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=0}^{s-1}$ satisfy

$$\mathbf{x}^{(i)} \in \mathbf{U}_{s-1}^\top \mathcal{K}_{2s-1}, \text{ and } \mathbf{y}^{(i)} \in \mathbf{V}_{s-1}^\top \mathcal{J}_{2s-1}, \ \forall i = 0, \ldots, s-1. \tag{3.16}$$

Suppose the next inquiry point generated by $\mathcal{M}$ is $(\mathbf{x}^{(s)}, \mathbf{y}^{(s)})$. Since $s < \frac{k}{2} - 1$, it holds that $2s < k$, and from (2.19) we have $\mathbf{U}_{s-1}^\top \mathcal{K}_{2s-1} \subsetneq \mathbf{U}_{s-1}^\top \mathcal{K}_{2s} \subsetneq \mathbf{U}_{s-1}^\top \mathcal{K}_{2s+1}$ and $\mathbf{V}_{s-1}^\top \mathcal{J}_{2s-1} \subsetneq \mathbf{V}_{s-1}^\top \mathcal{J}_{2s} \subsetneq \mathbf{V}_{s-1}^\top \mathcal{J}_{2s+1}$. By Lemma 3.1, there exist orthogonal matrices $\boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{m\times m}$ such that

$$\begin{aligned} &\boldsymbol{\Phi}\mathbf{x} = \mathbf{x}, \ \forall \mathbf{x} \in \mathbf{U}_{s-1}^\top \mathcal{K}_{2s}, \text{ and } \boldsymbol{\Phi}\mathbf{x}^{(s)} \in \mathbf{U}_{s-1}^\top \mathcal{K}_{2s+1}, \\ &\boldsymbol{\Psi}\mathbf{y} = \mathbf{y}, \ \forall \mathbf{y} \in \mathbf{V}_{s-1}^\top \mathcal{J}_{2s}, \text{ and } \boldsymbol{\Psi}\mathbf{y}^{(s)} \in \mathbf{V}_{s-1}^\top \mathcal{J}_{2s+1}. \end{aligned} \tag{3.17}$$

Since $\mathbf{c} \in \mathcal{J}_{2s}$ and $\mathbf{V}_{s-1}\mathbf{c} = \mathbf{c}$, we have $\mathbf{c} \in \mathbf{V}_{s-1}^\top \mathcal{J}_{2s}$, and thus it follows from (3.17) that $\boldsymbol{\Psi}\mathbf{c} = \mathbf{c}$. Let $\mathbf{U}_s = \mathbf{U}_{s-1}\boldsymbol{\Phi}$ and $\mathbf{V}_s = \mathbf{V}_{s-1}\boldsymbol{\Psi}$. Clearly, both $\mathbf{U}_s$ and $\mathbf{V}_s$ are orthogonal matrices, and because $\mathbf{V}_{s-1}\mathbf{c} = \mathbf{c}$ and $\boldsymbol{\Psi}\mathbf{c} = \mathbf{c}$, we have $\mathbf{V}_s\mathbf{c} = \mathbf{c}$. By a similar argument we also have $\mathbf{U}_s\mathbf{h} = \mathbf{h}$. In addition, from (3.17), Lemma 3.2, and the assumptions on $\mathbf{H}$ and $\mathbf{A}$, it follows that for any $\mathbf{x} \in \mathbf{U}_{s-1}^\top \mathcal{K}_{2s-1}$ and $\mathbf{y} \in \mathbf{V}_{s-1}^\top \mathcal{J}_{2s-1}$,

$$\begin{aligned} \mathbf{U}_s^\top \mathbf{H}\mathbf{U}_s\mathbf{x} = \mathbf{U}_{s-1}^\top \mathbf{H}\mathbf{U}_{s-1}\mathbf{x}, \ \mathbf{V}_s^\top \mathbf{A}\mathbf{U}_s\mathbf{x} = \mathbf{V}_{s-1}^\top \mathbf{A}\mathbf{U}_{s-1}\mathbf{x}, \text{ and } \mathbf{U}_s^\top \mathbf{A}^\top \mathbf{V}_s\mathbf{y} \\ = \mathbf{U}_{s-1}^\top \mathbf{A}^\top \mathbf{V}_{s-1}\mathbf{y}. \end{aligned} \tag{3.18}$$

Therefore, from the induction hypothesis (3.16) and the equations in (3.18), we conclude that the first $s+1$ iterates obtained from $\mathcal{M}$ applied to $P(\boldsymbol{\theta}; \mathbf{U}_s^\top \mathbf{H}\mathbf{U}_s, \mathbf{V}_s^\top \mathbf{A}\mathbf{U}_s)$ are exactly the same as the first $s+1$ iterates obtained from $\mathcal{M}$ applied to $P(\boldsymbol{\theta}; \mathbf{U}_{s-1}^\top \mathbf{H}\mathbf{U}_{s-1}, \mathbf{V}_{s-1}^\top \mathbf{A}\mathbf{U}_{s-1})$, because exactly the same information is used to generate those iterates (cf. (3.11)). Consequently, when $\mathcal{M}$ is applied to $P(\boldsymbol{\theta}; \mathbf{U}_s^\top \mathbf{H}\mathbf{U}_s, \mathbf{V}_s^\top \mathbf{A}\mathbf{U}_s)$, the first $s+1$ iterates are $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, $i = 0, 1, \ldots, s$. Hence, from (2.19), (3.16) and (3.17), and also the facts $\mathbf{U}_s = \mathbf{U}_{s-1}\boldsymbol{\Phi}$ and $\mathbf{V}_s = \mathbf{V}_{s-1}\boldsymbol{\Psi}$, we have

$$\mathbf{x}^{(i)} \in \mathbf{U}_s^\top \mathcal{K}_{2s+1}, \ \mathbf{y}^{(i)} \in \mathbf{V}_s^\top \mathcal{J}_{2s+1}, \ \forall i = 0, \ldots, s.$$

This finishes the induction. Recalling that in the discussion below (1.8) we have assumed that the output by $\mathcal{M}$ coincides with the last inquiry point, we have $\bar{\mathbf{x}}^{(t)} = \mathbf{x}^{(t)} \in \mathbf{U}_t^\top \mathcal{K}_{2t+1}$, and hence complete the proof. □

Using Proposition 3.2, we are now ready to prove Proposition 3.1.

Proof *(of Proposition 3.1)* Note that for the original instance $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ in Proposition 3.1, its data $\mathbf{H}$, $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{h}$ satisfy the conditions in Proposition 3.2. Hence, we apply Proposition 3.2 to obtain a rotated instance $P(\boldsymbol{\theta}; \mathbf{U}^\top \mathbf{H}\mathbf{U}, \mathbf{V}^\top \mathbf{A}\mathbf{U})$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices such that $\mathbf{U}\mathbf{h} = \mathbf{h}$ and $\mathbf{V}\mathbf{b} = \mathbf{b}$, and we have used the fact that $\mathbf{b}$ is a multiple of $\mathbf{c}$.

Let $\tilde{\mathbf{A}} = \mathbf{V}^\top \mathbf{A}\mathbf{U}$, ϕ and f denote the functions in the original instance $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$, and $\tilde{\phi}$ and $\tilde{f}$ denote those in the rotated instance $P(\boldsymbol{\theta}; \mathbf{U}^\top \mathbf{H}\mathbf{U}, \mathbf{V}^\top \mathbf{A}\mathbf{U})$. Then it is straightforward to observe the relations

$$\tilde{f}(\mathbf{x}) = f(\mathbf{U}\mathbf{x}), \text{ and } \tilde{\phi}(\mathbf{x}) = \phi(\mathbf{U}\mathbf{x}). \tag{3.19}$$

By the optimality conditions (e.g., [29]) of the original instance and the rotated instance, it is also easy to show that the pair $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point to the original instance if and only if $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = (\mathbf{U}^\top \mathbf{x}^*, \mathbf{V}^\top \mathbf{y}^*)$ is a saddle point to the rotated instance, and that $\tilde{\phi}^* = \phi^*$.

It remains to prove the inequalities from (3.3) through (3.6). By Proposition 3.2, when $\mathcal{M}$ is applied to the rotated instance, the approximate solution $\bar{\mathbf{x}}^{(t)} \in \mathbf{U}^\top \mathcal{K}_{2t+1}$, which indicates $\mathbf{U}\bar{\mathbf{x}}^{(t)} \in \mathcal{K}_{2t+1}$ by the orthogonality of $\mathbf{U}$. Since $t \le \frac{k}{2} - 1$, we have $2t + 1 \le k - 1$, and thus from (2.19), it follows that $\mathbf{U}\bar{\mathbf{x}}^{(t)} \in \mathcal{K}_{2t+1} \subseteq \mathcal{K}_{k-1}$. Therefore, from the facts $\tilde{\mathbf{A}} = \mathbf{V}^\top \mathbf{A}\mathbf{U}$, $\mathbf{U}\mathbf{h} = \mathbf{h}$ and $\mathbf{V}\mathbf{b} = \mathbf{b}$, and the relations in (3.19), we have

$$\begin{aligned}
&\tilde{\phi}(\bar{\mathbf{x}}^{(t)}) - \tilde{\phi}^* = \phi(\mathbf{U}\bar{\mathbf{x}}^{(t)}) - \phi^* \ge \min_{\mathbf{x} \in \mathcal{K}_{k-1}} \phi(\mathbf{x}) - \phi^*, \\
&\tilde{f}(\bar{\mathbf{x}}^{(t)}) - \tilde{f}(\hat{\mathbf{x}}) = f(\mathbf{U}\bar{\mathbf{x}}^{(t)}) - f(\mathbf{U}\hat{\mathbf{x}}) \ge \min_{\mathbf{x} \in \mathcal{K}_{k-1}} f(\mathbf{x}) - f(\mathbf{x}^*), \\
&\|\tilde{\mathbf{A}}\bar{\mathbf{x}}^{(t)} - \mathbf{b}\| = \|\mathbf{A}(\mathbf{U}\bar{\mathbf{x}}^{(t)}) - \mathbf{b}\| \ge \min_{\mathbf{x} \in \mathcal{K}_{k-1}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|, \\
&\|\bar{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}\|^2 = \|\mathbf{U}\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 \ge \min_{\mathbf{x} \in \mathcal{K}_{k-1}} \|\mathbf{x} - \mathbf{x}^*\|^2,
\end{aligned}$$

which complete the proof. □

4 Lower complexity bounds on bilinear saddle-point problems

In this section, we derive lower complexity bounds of first-order methods on solving the bilinear saddle-point problem (1.1) through considering its associated primal problem (1.3). As we mentioned in the beginning, the affinely constrained problem (1.6) is a special case of (1.3) if $Y = \mathbb{R}^m$ and $g = 0$. Hence, the results obtained in previous sections also apply to (1.3), namely, our designed instances of (1.6) are also "hard" instances of (1.3). However, they will not be the instance of (1.3) if we require Y to be a compact set. On solving (1.1) with both X and Y being compact, [35] gives a first-order method that can be described as (1.8), and it proves

$$0 \le \phi(\bar{\mathbf{x}}^{(t)}) - \psi(\bar{\mathbf{y}}^{(t)}) \le \frac{4L_f D_X^2}{(t+1)^2} + \frac{4D_X D_Y \|\mathbf{A}\|}{t+1}, \tag{4.1}$$

where D_X and D_Y are the diameters[3] of X and Y respectively. It is an open question if the convergence rate in (4.1) can still be improved. Under the Euclidean setting, a

[3] In fact, more general results are established in [35]. It adopts general norm (that is not necessary Euclidean norm) and general prox-functions to define $\|\mathbf{A}\|$, D_X and D_Y.

lower complexity bound for the special case $L_f = 0$ has been shown in [29]. We give instances below to show a lower complexity bound under the Euclidean setting but with $L_f > 0$. The bound is in the same form as that in (4.1) and differs only at the constants, and thus the convergence rate result in [35] is optimal under the Euclidean setting. The ingredients in the designed "hard" SPP instances are the same as those used in Sect. 2.

Let $m \leq n$ and $k < \frac{m}{2}$ be positive integers, and let $L_f > 0$ and $L_A > 0$. We consider the instance $P\big((\mathbf{h}, \mathbf{b}, R_X, R_Y, \lambda); \mathbf{H}, \mathbf{A}\big)$ with $(\mathbf{H}, \mathbf{h}, \mathbf{A}, \mathbf{b})$ given in (2.24), and

$$R_X = (2k+1)\sqrt{k},\ \ R_Y = \frac{L_f}{2L_A}\sqrt{2k},\ \ \lambda = \frac{L_A\sqrt{k}}{4}. \tag{4.2}$$

Clearly the above problem is a special instance of (1.3). In the following lemma, we give a lower bound of its optimal objective value.

Lemma 4.1 *Let $m \leq n$ and $k < \frac{m}{2}$ be positive integers, and let $L_f > 0$ and $L_A > 0$. Set $(\mathbf{H}, \mathbf{h}, \mathbf{A}, \mathbf{b})$ as those in (2.24) with R_X, R_Y, and λ as in (4.2). Then the optimal objective value of the instance $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ defined in Definition 3.1 satisfies*

$$\phi^* \leq -\frac{3L_f}{4}k. \tag{4.3}$$

Proof Since $\lambda\|\mathbf{y}\| \geq 0$, we have

$$\phi^* \leq l^* := \min_{\mathbf{x}\in X}\left\{ f(\mathbf{x}) + \max_{\mathbf{y}\in Y} \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle \right\}, \tag{4.4}$$

where f, X, and Y are defined in (3.1), and thus to prove (4.3), it suffices to show $l^* = -\frac{3L_f}{4}k$. By the optimality condition (e.g., [29]), $\mathbf{x}^* \in \mathbb{R}^n$ is an optimal solution to (4.4) if $\mathbf{x}^* \in X$, and there exists $\mathbf{y}^* \in Y$ such that

$$\langle \nabla f(\mathbf{x}^*) + \mathbf{A}^\top \mathbf{y}^*, \mathbf{x}^* - \mathbf{x}\rangle \leq 0,\ \ \langle \mathbf{b} - \mathbf{A}\mathbf{x}^*, \mathbf{y}^* - \mathbf{y}\rangle \leq 0, \forall\, \mathbf{x} \in X,\ \mathbf{y} \in Y. \tag{4.5}$$

Let $\mathbf{x}^*$ and $\mathbf{y}^*$ be the vectors given in (2.25). Note from the proof of Lemma 2.5, it holds that $\nabla f(\mathbf{x}^*) = \mathbf{H}\mathbf{x}^* - \mathbf{h} = \mathbf{A}^\top \mathbf{y}^*$ and $\mathbf{A}\mathbf{x}^* - \mathbf{b} = \mathbf{0}$. Hence, $(\mathbf{x}^*, -\mathbf{y}^*)$ satisfies the optimality condition in (4.5). In addition, from (2.32) and (2.27), it follows that $\|\mathbf{x}^*\| \leq R_X$ and $\|\mathbf{y}^*\| \leq R_Y$. Therefore, $\mathbf{x}^*$ is an optimal solution, and it is straightforward to compute $l^* = f(\mathbf{x}^*) = -\frac{3L_f}{4}k$. This completes the proof. □

In the following lemma, we compute the minimum value of $\phi(\mathbf{x})$ over $\mathcal{K}_{k-1}$.

Lemma 4.2 *Let $m \leq n$ and $k < \frac{m}{2}$ be positive integers, and let $L_f > 0$ and $L_A > 0$. Set $(\mathbf{H}, \mathbf{h}, \mathbf{A}, \mathbf{b})$ as those in (2.24) with R_X, R_Y, and λ as in (4.2). Consider the instance $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ defined in Definition 3.1, i.e.,*

$$\phi^* := \min_{\|\mathbf{x}\|\leq R_X}\left\{ \phi(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x} - \mathbf{h}^\top \mathbf{x} + \max_{\|\mathbf{y}\|\leq R_Y} \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle - \lambda\|\mathbf{y}\| \right\}.$$

Then

$$\min_{\mathbf{x}\in\mathcal{K}_{k-1}} \phi(\mathbf{x}) - \phi^* \geq \frac{L_f R_X^2}{4(2k+1)^2} + \frac{L_A R_X R_Y}{4(2k+1)}. \tag{4.6}$$

Proof Let f, X, and Y be defined in (3.1). Observing $\langle \mathbf{Ax}-\mathbf{b}, \mathbf{y}\rangle \leq \|\mathbf{Ax}-\mathbf{b}\| \cdot \|\mathbf{y}\|$, we have

$$\max_{\mathbf{y}\in Y} \langle \mathbf{Ax}-\mathbf{b}, \mathbf{y}\rangle - \lambda\|\mathbf{y}\| = \begin{cases} 0 & \text{if } \|\mathbf{Ax}-\mathbf{b}\| \leq \lambda, \\ R_Y(\|\mathbf{Ax}-\mathbf{b}\| - \lambda) & \text{if } \|\mathbf{Ax}-\mathbf{b}\| > \lambda. \end{cases}$$

For any $\mathbf{x} \in \mathcal{K}_{k-1}$, we have from (2.33) and (4.2) that $\|\mathbf{Ax}-\mathbf{b}\| - \lambda \geq \frac{L_A\sqrt{k}}{4} > 0$, and thus

$$\phi(\mathbf{x}) = f(\mathbf{x}) + R_Y(\|\mathbf{Ax}-\mathbf{b}\| - \lambda) \geq f(\mathbf{x}) + \frac{L_A R_Y\sqrt{k}}{4} \overset{(4.2)}{=} f(\mathbf{x}) + \frac{L_A R_X R_Y}{4(2k+1)}. \tag{4.7}$$

In addition, note that $f(\mathbf{x})$ here is exactly the same as that discussed in Lemma 2.6. Thus by (2.35), we have that for any $\mathbf{x} \in \mathcal{K}_{k-1}$,

$$f(\mathbf{x}) \geq -\frac{L_f}{2}k. \tag{4.8}$$

Applying (4.8) to (4.7), and noting the bound of ϕ^* in Lemma 4.1, we have for any $\mathbf{x} \in \mathcal{K}_{k-1}$ that

$$\phi(\mathbf{x}) - \phi^* \geq \frac{L_f}{4}k + \frac{L_A R_X R_Y}{4(2k+1)} \overset{(4.2)}{=} \frac{L_f R_X^2}{4(2k+1)^2} + \frac{L_A R_X R_Y}{4(2k+1)},$$

which implies the desired result in (4.6). □

Using Proposition 3.1 and Lemma 4.2, we are able to show a lower complexity bound of deterministic first-order methods on (1.1) as summarized in the following theorem.

Theorem 4.1 (Lower complexity bound for SPPs) Let $8 < m \leq n$ and $t < \frac{m}{4} - 1$ be positive integers, $L_f > 0$, and $L_A > 0$. Then for any deterministic first-order method $\mathcal{M}$ described in (1.8) on solving (1.1), there exists a problem instance of (1.1) such that f is L_f-smooth, $\|\mathbf{A}\| = L_A$, and X and Y are Euclidean balls with radii R_X and R_Y respectively. In addition,

$$\phi(\bar{\mathbf{x}}^{(t)}) - \psi(\bar{\mathbf{y}}^{(t)}) \geq \frac{L_f R_X^2}{4(4t+5)^2} + \frac{L_A R_X R_Y}{4(4t+5)}, \tag{4.9}$$

where ϕ and ψ are the associated primal and dual objective functions, and $(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)})$ is the approximate solution output by $\mathcal{M}$.

Proof Set $k = 2t+2 < \frac{m}{2}$ and consider the problem instance $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ described in Lemma 4.2. Note that the data $(\mathbf{H}, \mathbf{h}, \mathbf{A}, \mathbf{b}, R_X, R_Y, \lambda)$ satisfies the conditions required by Proposition 3.1. Hence, there is a rotated instance $P(\boldsymbol{\theta}; \tilde{\mathbf{H}}, \tilde{\mathbf{A}})$, and from (3.3) and Lemma 4.2, it follows that

$$\tilde{\phi}(\bar{\mathbf{x}}^{(t)}) - \tilde{\phi}^* \geq \min_{\mathbf{x} \in \mathcal{K}_{k-1}} \phi(\mathbf{x}) - \phi^* \geq \frac{L_f R_X^2}{4(4t+5)^2} + \frac{L_A R_X R_Y}{4(4t+5)}, \tag{4.10}$$

where ϕ and $\tilde{\phi}$ are respectively the primal objective functions of the original instance $P(\boldsymbol{\theta}; \mathbf{H}, \mathbf{A})$ and the rotated instance $P(\boldsymbol{\theta}; \tilde{\mathbf{H}}, \tilde{\mathbf{A}})$. Let $\tilde{\psi}$ be the dual objective function of the rotated instance. Then since $\bar{\mathbf{y}}^{(t)} \in Y$, it holds $\tilde{\psi}(\bar{\mathbf{y}}^{(t)}) \leq \tilde{\psi}^* \leq \tilde{\phi}^*$, where the second inequality follows from the weak duality. Therefore we have the desired result from (4.10) and by abusing the notation (ϕ, ψ) for $(\tilde{\phi}, \tilde{\psi})$. □

***Remark* 4.1** The lower bound in (4.9) has exactly the same form as the upper bound in (4.1), and they differ only on the constants. Hence, the order of the convergence rate result in (4.1) is not improvable under the Euclidean setting, and one can only improve that result by possibly decreasing the constants.

We finish this section by showing a lower complexity bound for SPPs when the function $f(\mathbf{x})$ in (1.1) is strongly convex.

Theorem 4.2 (Lower complexity bound for SPPs with strong convexity) Let $8 < m \leq n$ and $t < \frac{m}{4} - 1$ be positive integers, and μ and L_A be positive numbers. Then for any deterministic first-order method $\mathcal{M}$ described in (1.8), there exists a problem instance of (1.1) such that f is μ-strongly convex, $\|\mathbf{A}\| = L_A$, X and Y are Euclidean balls with radii R_X and R_Y respectively, and the associated primal problem (1.3) has a unique optimal solution $\mathbf{x}^* \in X$. In addition,

$$\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 \geq \frac{5L_A^2 R_Y^2}{256\mu^2(4t+5)^2} \tag{4.11}$$

and

$$\phi(\bar{\mathbf{x}}^{(t)}) - \psi(\bar{\mathbf{y}}^{(t)}) \geq \frac{5L_A^2 R_Y^2}{512\mu(4t+5)^2}, \tag{4.12}$$

where ϕ and ψ are the associated primal and dual objective functions, and $(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)})$ is the output by $\mathcal{M}$.

Proof Set $k = 2t+2 < \frac{m}{2}$ and consider the problem instance of (1.1), where $f(\mathbf{x}) = \frac{\mu}{2}\|\mathbf{x}\|^2$, $\mathbf{A}$ and $\mathbf{b}$ are those in (2.20), $g \equiv 0$, and

$$\begin{aligned} X &= \left\{ \mathbf{x} \in \mathbb{R}^n \,|\, \|\mathbf{x}\|^2 \leq R_X^2 := k(2k+1)^2 \right\}, \\ Y &= \left\{ \mathbf{y} \in \mathbb{R}^m \,|\, \|\mathbf{y}\|^2 \leq R_Y^2 := \frac{128\mu^2}{15L_A^2} k(k+1)^3(2k+1) \right\}. \end{aligned} \tag{4.13}$$

From the proof of Theorem 2.2, it is easy to verify that $\mathbf{x}^*$ in (2.25) and $\mathbf{y}^*$ in (2.37) satisfy $\mathbf{x}^* \in X$, $\mathbf{y}^* \in Y$, and the optimality condition in (4.5) holds for $(\mathbf{x}^*, -\mathbf{y}^*)$. Since f is strongly convex, $\mathbf{x}^*$ must be the unique optimal solution to the instance. From (2.38) and also the definitions of X and Y in (4.13), it follows that

$$\min_{\mathbf{x}\in\mathcal{K}_{k-1}} \|\mathbf{x}-\mathbf{x}^*\|^2 \geq \frac{5L_A^2R_Y^2}{256\mu^2(2k+1)^2}. \tag{4.14}$$

Note that the above instance can be represented as $P\big((\mathbf{0}, \mathbf{b}, R_X, R_Y, 0); \mu\mathbf{I}, \mathbf{A}\big)$ by Definition 3.1, and the data in the instance satisfy all the conditions in Proposition 3.1. Hence, we can obtain a rotated instance $P\big((\mathbf{0}, \mathbf{b}, R_X, R_Y, 0); \mu\mathbf{I}, \tilde{\mathbf{A}}\big)$, and it has a unique optimal solution $\hat{\mathbf{x}} \in X$. Now use (3.6) and (4.14) to obtain (4.11) by recalling $k = 2t + 2$ and abusing $\mathbf{x}^*$ for $\hat{\mathbf{x}}$.

Let $\tilde{\phi}$ and $\tilde{\psi}$ be the primal and dual objective functions of the rotated instance. By the strong convexity and the optimality of $\hat{\mathbf{x}}$, we have

$$\tilde{\phi}(\bar{\mathbf{x}}^{(t)}) - \tilde{\phi}^* \geq \frac{\mu}{2}\|\bar{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}\|^2.$$

Together with (4.11) and the fact $\tilde{\psi}(\bar{\mathbf{y}}^{(t)}) \leq \tilde{\psi}^* \leq \tilde{\phi}^*$, the above inequality gives (4.12) by abusing the notation $(\phi, \psi, \mathbf{x}^*)$ for $(\tilde{\phi}, \tilde{\psi}, \hat{\mathbf{x}})$. Therefore, we complete the proof. □

***Remark* 4.2** In the proof of Theorem 4.2, we have $g = 0$ in the obtained rotated instance. Similar to Theorem 4.1, we can have an instance with a nonzero g and have a result similar to that in (4.12). Specifically, we consider $P\big((\mathbf{0}, \mathbf{b}, R_X, R_Y, \lambda); \mu\mathbf{I}, \mathbf{A}\big)$, where $\lambda = \frac{L_A}{6}\sqrt{k}$, and the tuple $(\mathbf{A}, \mathbf{b}, R_X, R_Y)$ is the same as that in the above proof. Let ϕ be the primal objective of the new instance. Then by the same arguments as those in the proof of Lemma 4.1, we can show $\phi^* \leq \frac{\mu}{2}\|\mathbf{x}^*\|^2$, where $\mathbf{x}^*$ is given in (2.25). Furthermore, similar to (4.7), we can show that for any $\mathbf{x} \in \mathcal{K}_{k-1}$, it holds

$$\phi(\mathbf{x}) \geq \frac{\mu}{2}\|\mathbf{x}\|^2 + \frac{L_A R_Y\sqrt{k}}{3} \geq \frac{L_A R_Y\sqrt{k}}{3}.$$

Therefore, $\min_{\mathbf{x}\in\mathcal{K}_{k-1}} \phi(\mathbf{x}) - \phi^* \geq \frac{L_A R_Y\sqrt{k}}{3} - \frac{\mu}{2}\|\mathbf{x}^*\|^2$. Now applying Proposition 3.1, we obtain a rotated instance $P\big((\mathbf{0}, \mathbf{b}, R_X, R_Y, \lambda); \mu\mathbf{I}, \tilde{\mathbf{A}}\big)$ with primal objective $\tilde{\phi}$, and by (3.3), we have

$$\begin{aligned}\tilde{\phi}(\bar{\mathbf{x}}^{(t)}) - \tilde{\phi}^* &\geq \frac{L_A R_Y\sqrt{k}}{3} - \frac{\mu}{2}\|\mathbf{x}^*\|^2 \geq \frac{4\sqrt{2}}{3}\left(\frac{2}{\sqrt{15}} - \frac{1}{2}\right)\mu k\sqrt{(k+1)^3(2k+1)} \\ &= \Omega\big(\frac{L_A^2R_Y^2}{\mu k^2}\big).\end{aligned}$$

5 On the tightness of the established lower complexity bounds

In this section, we compare the established lower complexity bounds to the best known upper complexity bounds. It turns out that the lower complexity bounds developed in this paper are tight in terms of the order, and thus they can be used to justify the optimality of first-order methods in the literature.

5.1 Upper complexity bounds of first-order methods on affinely constrained problems

The work [37] proposes an accelerated linearized alternating direction method of multipliers (AL-ADMM). Applying to (1.6), i.e., setting one block to zero, we have from one convergence rate result in [37, eqn. (2.34)] that

$$f(\mathbf{x}^{(t)}) - f^* \leq \frac{2L_f D_X^2}{t(t+1)} + \frac{2\|\mathbf{A}\| D_X D_Y}{t+1},$$

where D_X and D_Y are the diameters of the primal and dual feasible sets. If the size of the optimal primal and dual solutions is assumed, then the above result coincides with that in (3.7a) up to the difference of a constant multiple.

For the strongly convex case, the result in (3.10) indicates that given any $\varepsilon > 0$, to have an iterate within $\sqrt{\varepsilon}$-neighborhood of $\mathbf{x}^*$, the iterate number is at least

$$t = \left\lceil \frac{\sqrt{5} L_A \|\mathbf{y}^*\|}{32\mu\sqrt{\varepsilon}} - \frac{5}{2} \right\rceil, \tag{5.1}$$

where $\lceil a \rceil$ denotes the smallest integer no less than $a \in \mathbb{R}$. In [42, proof of Thm.4], it is shown that

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) + \langle \mathbf{y}^*, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \rangle \leq \frac{\|\mathbf{y}^*\|^2}{2\rho_0} + \varepsilon_0, \tag{5.2}$$

where $(\mathbf{x}^*, \mathbf{y}^*)$ is a pair of primal-dual solution, and $\mathbf{x}^{(t)}$ is the output of Nesterov's optimal first-order method applied to a penalized problem after t iterations. In addition, with $\rho_0 = \frac{2\|\mathbf{y}^*\|^2}{\mu\varepsilon}$ and $\varepsilon_0 = \frac{\mu\varepsilon}{4}$ in (5.2), [42, eqn.(49)] shows that the iteration number t satisfies:

$$t \leq 2\left(\sqrt{\frac{L_f}{\mu}} + \frac{2L_A\|\mathbf{y}^*\|}{\mu\sqrt{\varepsilon}} \right) \left(O(1) + \log\frac{1}{\varepsilon} \right). \tag{5.3}$$

From the strong convexity of f, it follows that

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) + \langle \mathbf{y}^*, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \rangle \geq \frac{\mu}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2,$$

which together with (5.2) gives $\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2 \leq \varepsilon$. Hence, the dominant term in the upper bound (5.3) is the same as that in (5.1) except for a logarithmic term.

5.2 Upper complexity bounds of first-order methods on saddle-point problems

For optimization problems in the form of (1.3), a smoothing technique is proposed in [35]. It first approximates the nonsmooth objective function by a smooth one and then minimizes the smooth approximation function by an accelerated gradient method. In [35], it is shown that, if X and Y are compact with diameter D_X and D_Y respectively, and the total number of iterations is pre-specified to t, then the convergence rate of this smoothing scheme applied to (1.3) is given in (4.1). Comparing the upper bound in (4.1) and the lower bound in (4.9), we conclude that our lower complexity bound in Theorem 4.1 is tight in terms of the order, and that Nesterov's smooth scheme is an optimal method for computing approximate solutions to bilinear SPPs in the form of (1.1).

Note that Theorem 4.1 also confirms the optimality of several follow-up works of [35]. For example, when the algorithms in [7,8] are applied to solve (1.3), their convergence rates all coincide with the lower bound in (4.9) up to a constant multiple, and hence these methods are all optimal first-order methods for solving problems in the form of (1.3).

In the literature, there have also been several results on either the saddle point or the variational inequality formulation of (1.3) [6,17,27–29]. When applied to solve (1.3) with $f \equiv 0$ (and hence $L_f \leq L_A$), those results all imply

$$\phi(\mathbf{x}^{(t)}) - \phi^* = O\left(\frac{L_A D_X D_Y}{t}\right),$$

where D_X and D_Y are the diameters of X and Y. The above result indicates the tightness of the lower bound in (4.9).

6 Concluding remarks

On finding solutions to bilinear saddle-point problems, we have established lower complexity bounds of first-order methods that acquire problem information through a first-order oracle and are described by a sequence of updating rules. Through designing "hard" instances of convex quadratic programming, we first show the lower complexity bound results under a linear span assumption on solving affinely constrained problems. Then by a rotation invariance technique, we extend the results to general first-order methods that are still applied to affinely constrained problems. Finally, we establish the results for general first-order methods on solving bilinear saddle-point problems with compact primal and dual feasible regions. The established lower complexity bounds have been compared to several existing upper bound results. The comparison implies the tightness of our bounds and optimality of a few first-order methods in the literature.

We conclude the paper with a few more remarks. First, note that for affinely constrained problems, the feasibility residual in none of our results depends on the objective; see (2.36b) and (3.7b) for example. This is reasonable because we can choose not to use the objective gradient though the oracle (1.7) provides such infor-

mation. However, towards finding an optimal solution, the objective information must be used. All existing works (e.g., [8,24,41]) on primal-dual first-order methods have objective-dependent quantity in their upper bounds on the feasibility error. One interesting question is how to derive a lower complexity bound of the feasibility residual that depends on the constraint itself and also the objective. To achieve that, we would need to enforce a minimum portion of objective information to be used in the solution update. Second, a few existing works [22,23,25] have shown that if ∇f is much more expensive than matrix-vector multiplication $\mathbf{A}\mathbf{x}$ and $\mathbf{A}^\top\mathbf{y}$, it could be beneficial to skip computing ∇f at the cost of more $\mathbf{A}\mathbf{x}$ and/or $\mathbf{A}^\top\mathbf{y}$. This setting is different from what we have made. In (1.7), we assume that one inquiry of the first-order oracle will obtain gradient and matrix-vector multiplications simultaneously. In the future work, we will allow multiple oracles that can return separate pieces of information, and we will pursue the lower bound of each oracle inquiry to reach a solution with desired accuracy and also design optimal oracle-based algorithms. Thirdly, in all our established results, we do not pre-specify the size of X and Y but allow them to be determined in the designed instances. That is the key reason why we obtain a lower complexity bound that looks greater than existing upper bound, e.g., by comparing (4.1) and (4.9). It is interesting to design "hard" instances to establish similar lower complexity bound results, provided that L_f, L_A and the diameters of X, Y are all given. We leave this to the future work.

References

1. Arjevani, Y., Shamir, O.: Dimension-free iteration complexity of finite sum optimization problems. In: Advances in Neural Information Processing Systems, pp. 3540–3548. (2016)
2. Arjevani, Y., Shamir, O.: On the iteration complexity of oblivious first-order optimization algorithms. In: International Conference on Machine Learning, pp. 908–916. (2016)
3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
4. Carmon Y, Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points I. (2017) arXiv preprint arXiv:1710.11606
5. Carmon Y, Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points II: First-order methods. (2017) arXiv preprint arXiv:1711.00841
6. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
7. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal-dual methods for a class of saddle point problems. SIAM J. Optim. **24**(4), 1779–1814 (2014)
8. Chen, Y., Lan, G., Ouyang, Y.: Accelerated schemes for a class of variational inequalities. Math. Program. **165**(1), 113–149 (2017)
9. Condat, L.: A primal-dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**(2), 460–479 (2013)
10. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**(1–2), 37–75 (2014)
11. Esser, E., Zhang, X., Chan, T.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)
12. Gao, X., Xu, Y., Zhang, S.: Randomized primal-dual proximal block coordinate updates. J. Oper. Res. Soc. China **7**(2), 205–250 (2019)
13. Gao, X., Zhang, S.-Z.: First-order algorithms for convex optimization with nonseparable objective and coupled constraints. J. Oper. Res. Soc. China **5**(2), 131–159 (2017)

14. Goldstein, T., O'Donoghue, B., Setzer, S., Baraniuk, R.: Fast alternating direction optimization methods. SIAM J. Imaging Sci. **7**(3), 1588–1623 (2014)
15. Guzmán, C., Nemirovski, A.: On lower complexity bounds for large-scale smooth convex optimization. J. Complexity **31**(1), 1–14 (2015)
16. Hamedani, E.Y., Aybat, N.S.: A primal-dual algorithm for general convex-concave saddle point problems. (2018) arXiv preprint arXiv:1803.01401
17. He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the douglas-rachford alternating direction method. SIAM J. Numer. Anal. **50**(2), 700–709 (2012)
18. He, Y., Monteiro, R.D.: An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems. SIAM J. Optim. **26**(1), 29–56 (2016)
19. Jaggi, M.: Revisiting frank-wolfe: Projection-free sparse convex optimization. In: ICML, vol. 1, pp. 427–435. (2013)
20. Juditsky, A., Nesterov, Y.: Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. Stoch. Syst. **4**(1), 44–80 (2014)
21. Lan, G.: The complexity of large-scale convex programming under a linear optimization oracle. (2013) arXiv preprint arXiv:1309.5550
22. Lan, G.: Gradient sliding for composite optimization. Math. Program. **159**(1–2), 201–235 (2016)
23. Lan, G., Ouyang, Y.: Accelerated gradient sliding for structured convex optimization. (2016) arXiv preprint arXiv:1609.04905
24. Lan, G., Renato, D., Monteiro, C.: Iteration-complexity of first-order augmented lagrangian methods for convex programming. Math. Program. **155**(1–2), 511–547 (2016)
25. Lan, G., Zhou, Y.: Conditional gradient sliding for convex optimization. SIAM J. Optim. **26**(2), 1379–1409 (2016)
26. Lan, G., Zhou, Y.: An optimal randomized incremental gradient method. Math. Program. **171**(1–2), 167–215 (2018)
27. Monteiro, R.D., Svaiter, B.F.: Complexity of variants of Tseng's modified F-B splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. SIAM J. Optim. **21**(4), 1688–1720 (2011)
28. Monteiro, R.D., Svaiter, B.F.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim. **23**(1), 475–507 (2013)
29. 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 J. Optim. **15**(1), 229–251 (2004)
30. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
31. Nemirovski, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience Series in Discrete Mathematics, Wiley, New York (1983)
32. Nemirovski, A.S.: Information-based complexity of linear operator equations. J. Complexity **8**(2), 153–175 (1992)
33. Nemirovsky, A.: On optimality of krylov's information when solving linear operator equations. J. Complexity **7**(2), 121–130 (1991)
34. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publisher, Dordrecht (2004)
35. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
36. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
37. Ouyang, Y., Chen, Y., Lan, G., Pasiliao Jr., E.: An accelerated linearized alternating direction method of multipliers. SIAM J. Imaging Sci. **8**(1), 644–681 (2015)
38. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (2015)
39. Simchowitz, M.: On the randomized complexity of minimizing a convex quadratic function. (2018) arXiv preprint arXiv:1807.09386
40. Woodworth, B.E., Srebro, N.: Tight complexity bounds for optimizing composite objectives. In: Advances in Neural Information Processing Systems, pp. 3639–3647. (2016)
41. Xu, Y.: Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. SIAM J. Optim. **27**(3), 1459–1484 (2017)
42. Xu, Y.: Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. (2017) arXiv preprint arXiv:1711.05812

43. Xu, Y., Zhang, S.: Accelerated primal-dual proximal block coordinate updating methods for constrained convex optimization. Comput. Optim. Appl. **70**(1), 91–128 (2018)
44. Yan, M.: A new primal-dual algorithm for minimizing the sum of three functions with a linear operator. J. Sci. Comput. **76**(3), 1698–1717 (2018)

Affiliations

Yuyuan Ouyang[1] · Yangyang Xu[2]

Yuyuan Ouyang
yuyuano@clemson.edu

1 School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC, USA

2 Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA