Gradient Methods With Dynamic Inexact Oracles

Shuo Han, *Member, IEEE*

***Abstract*—We present a framework for generalizing the primal-dual gradient method, also known as the gradient descent ascent method, for solving convex-concave minimax problems. The framework is based on the observation that the primal-dual gradient method can be viewed as an inexact gradient method applied to the primal problem. Unlike the setting of traditional inexact gradient methods, the inexact gradient is computed by a *dynamic inexact oracle*, which is a discrete-time dynamical system whose output asymptotically approaches the exact gradient. For minimax problems, dynamic inexact oracles are capable of modeling a range of first-order methods for computing the gradient of the primal objective, which relies on solving the inner maximization problem. We provide a unified convergence analysis of gradient methods with dynamic inexact oracles and demonstrate its use in creating new accelerated primal-dual algorithms.**

***Index Terms*—Optimization algorithms, stability of nonlinear systems.**

I. Introduction

We consider algorithms for solving the unconstrained minimax problem

$$\min_{x\in\mathbb{R}^n} \quad \max_{y\in\mathbb{R}^m} L(x,y) := f(x) + y^T Ax - g(y). \tag{1}$$

We assume that f is smooth and convex (but not necessarily strongly convex), g is smooth and strongly convex, and $A \in \mathbb{R}^{m\times n}$ has full column rank. For convenience, we define $p(x) := \max_y L(x,y)$ and write problem (1) as

$$\min_x \quad p(x), \tag{2}$$

which we refer to as the *primal problem*. We also define $d(y) := \min_x L(x,y)$ and refer to the problem

$$\max_y \quad d(y) \tag{3}$$

as the *dual problem*. Under the given assumptions, it follows from standard results (see, [13, Ch. 10]) in convex analysis that both p and $-d$ are strictly convex (in fact, strongly convex). Therefore, the primal-dual optimal solution of problems (2) and (3) is unique, which we denote by $(x^\star, y^\star)$.

The minimax problem (1) has a number of applications. For example, when $f(x) = -b^T x$ for some $b \in \mathbb{R}^n$, the dual problem (3) becomes equivalent to the equality-constrained convex optimization problem given by

$$\max_y \quad -g(y), \qquad \text{s.t.} \quad A^T y = b.$$

Other applications include image processing [6] and empirical risk minimization [23]. More broadly, when the function L is a general convex-concave function, the minimax problem formulation also arises in game theory [17] and robust optimization [2].

One important algorithm for computing the primal-dual optimal solution $(x^\star, y^\star)$ is the *primal-dual gradient method (PDGM)*:

$$\begin{aligned} x^{k+1} &= x^k - \eta_1 \nabla_1 L(x^k, y^k) \\ y^{k+1} &= y^k + \eta_2 \nabla_2 L(x^k, y^k), \end{aligned} \tag{4}$$

where η_1 and η_2 are step sizes, and $\nabla_1 L(x^k, y^k) = \nabla f(x^k) + A^T y^k$ and $\nabla_2 L(x^k, y^k) = Ax^k - \nabla g(y^k)$ are the partial derivatives of L with respect to the first and second arguments, respectively. The PDGM is also known by various other names such as the Arrow–Hurwicz gradient method [1, p. 155] and the (simultaneous) gradient descent ascent method (see [8]). It has also been generalized to the case where L is nondifferentiable [18] and the case where the dynamics in (4) are in continuous time [7], [12], [20]. Convergence of the PDGM has been studied extensively in the literature. Under the assumption we made on f, g, and A, it has been shown [10] that the PDGM converges exponentially to the optimal solution $(x^\star, y^\star)$.

Because the update rule (4) of the PDGM performs gradient descent/ascent on the primal/dual variable, a natural question arises as to whether these gradient updates can be substituted by other first-order methods (e.g., Nesterov's accelerated gradient method) to create new primal-dual algorithms. This letter attempts to address this question based on an alternative view of the PDGM: We show that the PDGM is equivalent to applying an inexact gradient method to the primal problem (2), where the gradient ∇p is computed approximately by a *dynamic inexact oracle* (Definition 1), whose output approaches the exact gradient asymptotically. For the

Manuscript received July 2, 2020; accepted August 12, 2020. Date of publication August 24, 2020; date of current version September 11, 2020. Recommended by Senior Editor J.-F. Zhang.

The author is with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: hanshuo@uic.edu).

Digital Object Identifier 10.1109/LCSYS.2020.3019222

PDGM, the inexact oracle is realized by running one iteration of gradient descent with warm starts (see Section III-A).

While the notion of inexact oracles has long existed in the study of optimization algorithms, including approximating the gradient mapping (see [4, Ch. 3.3]) and the proximal operator [21], these inexact oracles are memoryless mappings and hence less general than our proposed notion of dynamic inexact oracles, which are permitted to have internal states necessary for modeling warm starts used in iterative algorithms. The introduction of dynamics also demands a new analysis for capturing the dynamical interaction between the gradient method and the inexact oracle. By modeling the dynamical interaction as a feedback interconnection of two dynamical systems and using the small-gain principle, we derive a unified convergence analysis (Theorem 2) that does not rely on the detailed realization of the oracle. The convergence analysis also enables us to build new primal-dual algorithms by simply changing the realization of the inexact oracle used in PDGM to other first-order methods in a "plug-and-play" manner.

II. Mathematical Preliminaries

For a vector x, we denote by $\|x\|$ its ℓ_2-norm and $\|x\|_P := (x^TPx)^{1/2}$ its P-quadratic norm, where P is a positive definite matrix (written as $P \succ 0$). For a bivariate function $f(\cdot,\cdot)$, we denote by $\nabla_i f$ $(i = 1, 2)$ the partial derivative of f with respect to the ith argument. Unless noted otherwise, we reserve the use of superscripts for indexing an infinite sequence $\{x^k\}_{k=0}^{\infty}$.

For a real-valued function f, we denote by f^* its convex conjugate, defined by $f^*(s) := \sup_x\{s^Tx - f(x)\}$. We denote by $\mathcal{S}(\mu,\beta)$ the set of μ-strongly convex and β-smooth functions. By convention, we use $\mathcal{S}(0,\beta)$ to denote the set of β-smooth and convex functions. The assumptions on f, g, and A in the beginning of Section I can be stated as follows.

Assumption 1: Let f, g, and A in the minimax problem (1) be such that $f \in \mathcal{S}(0,\beta_f)$, $g \in \mathcal{S}(\mu_g,\beta_g)$, and A has full column rank.

Functions in $\mathcal{S}(\mu,\beta)$ are known to have the following basic properties.

Proposition 1 (Basic Properties): If $f \in \mathcal{S}(\mu,\beta)$, then

1) $(x-y, \nabla f(x)-\nabla f(y)) \in \mathsf{sec}(\mu,\beta)$ for all x and y, where

$$\mathsf{sec}(\mu,\beta) := \left\{(v,w) : \begin{bmatrix} v \\ w \end{bmatrix}^T \begin{bmatrix} -2\mu\beta I & (\mu+\beta)I \\ (\mu+\beta)I & -2I \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} \right\} \geq 0$$

is called the *sector constraint.*

Furthermore, if $\mu > 0$, then

2) $f^* \in \mathcal{S}(1/\beta, 1/\mu)$;
3) ∇f is invertible and $(\nabla f)^{-1} = \nabla f^*$, where ∇f^* is the gradient of f^*.

A proof of item 1 can be found in [19, Th. 2.1.12]. Proofs of items 2 and 3 can be found in [13, Ch. 10]. Let $\sigma_{\max}$ and $\sigma_{\min}$ be the maximum and minimum singular values of A, respectively. From Proposition 1, we have $p \in \mathcal{S}(\mu_p,\beta_p)$, where $\mu_p = \sigma_{\min}^2/\beta_g$ and $\beta_p = \sigma_{\max}^2/\mu_g + \beta_f$.

III. Dynamic Inexact Oracles

We begin by considering another way to solve the primal problem (2) by directly applying the gradient method, which reveals that the PDGM can be viewed alternatively as an inexact gradient method applied to the primal problem. An abstraction of the inexact gradient computation leads to the definition of dynamic inexact oracles, the central topic of study in this letter.

A. The PDGM as Inexact Gradient Descent

Consider solving the primal problem (2) using the gradient method:

$$x_{\text{ex}}^{k+1} = x_{\text{ex}}^k - \eta_1 \nabla p(x_{\text{ex}}^k), \tag{5}$$

where η_1 is the step size. (The subscript "ex" stands for *exact*, to distinguish from the inexact gradient method to be presented shortly.) Using Danskin's theorem (see [5, p. 245]), we obtain

$$\nabla p(x_{\text{ex}}^k) = \nabla f(x_{\text{ex}}^k) + A^T y_{\text{ex}}^k,$$

where $y_{\text{ex}}^k = \arg\min_y\{g(y) - y^TAx_{\text{ex}}^k\}$ and is unique because g is strongly convex. Define

$$\tilde{g}(y,x) := g(y) - y^TAx.$$

The gradient method (5) can be rewritten as[1]

$$y_{\text{ex}}^k = \arg\min_y \tilde{g}(y, x_{\text{ex}}^k) \tag{6a}$$

$$x_{\text{ex}}^{k+1} = x_{\text{ex}}^k - \eta_1(\nabla f(x_{\text{ex}}^k) + A^T y_{\text{ex}}^k). \tag{6b}$$

The PDGM can be derived from (6) by allowing the minimization problem in (6a) to be solved approximately. To avoid confusion, we will use (x^k, y^k) in place of $(x_{\text{ex}}^k, y_{\text{ex}}^k)$ in (6) when approximation occurs. Let the approximate solution $\{y^k\}$ be generated from applying one iteration of gradient descent with step size η_2 to the minimization problem in (6a):

$$\begin{aligned} y^{k+1} &= y^k - \eta_2 \nabla_1 \tilde{g}(y^k, x^k) \\ &= y^k + \eta_2(Ax^k - \nabla g(y^k)). \end{aligned} \tag{7}$$

Note that the update rule (7) uses a warm start: It uses the approximate solution y^k at iteration k to initialize iteration $k+1$. Replacing y_{ex}^k in (6b) with the approximate solution y^k yields

$$x^{k+1} = x^k - \eta_1(\nabla f(x^k) + A^T y^k). \tag{8}$$

It can be seen that (7) and (8) recover the update rule (4) of PDGM. Because $\nabla f(x^k) + A^T y^k$ in (8) no longer equals the exact gradient $\nabla p(x^k)$, the PDGM can be viewed as an inexact gradient method applied to the primal problem.

B. Definition of Dynamic Inexact Oracles

It is conceivable that the gradient method (7) is not the only iterative algorithm for approximating the minimizer in (6a), which is needed for computing the gradient ∇p. To simplify analysis, we introduce the notion of dynamic inexact oracles as a high-level abstraction of the iterative algorithms that replace the exact minimization in (6a).

[1]Equivalent to the augmented Lagrangian method (see [4, p. 262]).

Definition 1 (Dynamic Inexact Oracles): A (discrete-time) dynamical system $\mathcal{G}$ is called a *dynamic inexact oracle for approximating a mapping* ϕ (called the *exact oracle*) if for any input sequence $u = \{u^k\}_{k=0}^{\infty}$ converging to $u^\star$, the output $\mathcal{G}u$ converges to $\phi(u^\star)$.

The condition in Definition 1 is rather mild: It only requires that a dynamic inexact oracle $\mathcal{G}$ produces the same output as the exact oracle ϕ in a steady state; no condition is imposed on the transient, where approximation errors may occur. In the context of (6a), the exact oracle corresponds to

$$\phi(\cdot) = \arg\min_y \tilde{g}(y, \cdot), \tag{9}$$

which provides gradient information due to the following lemma. A proof of Lemma 1 can be found in Appendix A.

Lemma 1: Let ϕ be given by (9). Then we have $\phi(x_{\text{ex}}^k) = \nabla g^*(Ax_{\text{ex}}^k)$ for any x_{ex}^k.

C. An Oracle Based on Gradient Descent

The following shows that the recursion (7) based on gradient descent, if viewed as a dynamical system $\mathcal{G}_{\text{gd}}$ with input x and output y, is indeed an inexact oracle for approximating ϕ in (9).

Proposition 2: Let $\mathcal{G}_{\text{gd}}$ be a dynamical system whose input $x = \{x^k\}$ and output $y = \{y^k\}$ are described by (7), where $\eta_2 \in (0, 2/(\mu_g+\beta_g)]$. Then $\mathcal{G}_{\text{gd}}$ is a dynamic inexact oracle for approximating ϕ in (9); namely, for any input $\{x^k\}$ converging to $x^\star$, the output $\{y^k\}$ of $\mathcal{G}_{\text{gd}}$ converges to $\phi(x^\star)$.

To prove Proposition 2, we need to make use of the following lemma, modified from a standard result (see [19, Th. 2.1.15]) in convex analysis for establishing the convergence of gradient descent. A proof of Lemma 2 can be found in Appendix B.

Lemma 2 (Contraction): Let μ and β be constants satisfying $0 < \mu \le \beta$, and $\alpha = \mu\beta/(\mu+\beta)$. Suppose $(\xi, w) \in \mathsf{sec}(\mu, \beta)$. Then for any $\eta \in (0, 2/(\mu+\beta)]$, we have $\|\xi - \eta w\| \le \rho\|\xi\|$, where $\rho = 1 - \alpha\eta \in [0, 1)$.

Proof (Proposition 2): Define the desired steady-state output of $\mathcal{G}_{\text{gd}}$ as $y^\star := \phi(x^\star) = \nabla g^*(Ax^\star)$. Rewrite (7) as

$$y^{k+1} - y^\star = (y^k - y^\star) - \eta_2(\nabla g(y^k) - Ax^\star) + \eta_2 A(x^k - x^\star).$$

Because $g \in \mathcal{S}(\mu_g, \beta_g)$ and $Ax^\star = \nabla g(\nabla g^*(Ax^\star)) = \nabla g(y^\star)$, we have $(y^k - y^\star, \nabla g(y^k) - Ax^\star) \in \mathsf{sec}(\mu_g, \beta_g)$. By Lemma 2, there exists $\rho \in [0, 1)$ such that

$$\|y^{k+1} - y^\star\| \le \rho\|y^k - y^\star\| + \eta_2\|A(x^k - x^\star)\|.$$

The result then follows as a consequence of input-to-state stability [14, p. 192]. ∎

D. Oracles Based on General First-Order Algorithms

Other than gradient descent, a dynamical inexact oracle $\mathcal{G}_{\text{io}}$ for approximating the minimizer in (6a) can be constructed from a range of first-order optimization algorithms that use $\nabla_1\tilde{g}$. Inspired by the work in [22], we consider $\mathcal{G}_{\text{io}}$ expressed in the following state-space form:

$$\begin{aligned} \xi^{k+1} &= A_{\text{io}}\xi^k + B_{\text{io}}\nabla_1\tilde{g}(v^k, x^k) \\ &= A_{\text{io}}\xi^k + B_{\text{io}}\left[\nabla g(v^k) - Ax^k\right] \\ v^k &= C_{\text{io}}\xi^k, \qquad y^k = E_{\text{io}}\xi^k, \end{aligned} \tag{10}$$

where A_{io}, B_{io}, C_{io}, and E_{io} are given by

$$A_{\text{io}} = \begin{bmatrix} (1+c_1)I & -c_1 I \\ I & 0 \end{bmatrix}, \quad B_{\text{io}} = \begin{bmatrix} -\eta_2 I \\ 0 \end{bmatrix}$$
$$C_{\text{io}} = \begin{bmatrix} (1+c_2)I & -c_2 I \end{bmatrix}, \quad E_{\text{io}} = \begin{bmatrix} (1+c_3)I & -c_3 I \end{bmatrix}.$$

Here, $\xi = (\xi_1, \xi_2)$ is the state, x is the input, y is the output, v is the feedback output, η_2 is the step size, and c_1, c_2, and c_3 are constants. When the system (10) is not a minimal realization, the matrices A_{io}, B_{io}, C_{io}, and E_{io} need to be further simplified by removing redundant states. The form (10) captures a number of important first-order optimization algorithms. For example, setting $c_1 = c_2 = c_3 = 0$ recovers the gradient method, and setting $c_1 = c_2 \neq 0$ and $c_3 = 0$ recovers Nesterov's accelerated gradient method. Interested readers can refer to [22, Table I] for more examples. By constructing the inexact oracle from different first-order algorithms, we can create new primal-dual first-order methods beyond the PDGM.

To ensure that $\mathcal{G}_{\text{io}}$ is a dynamic inexact oracle for approximating ϕ in (9), we shall make the following assumption on the coefficient matrices in (10), which generalizes the conditions in Lemma 2. The proof that $\mathcal{G}_{\text{io}}$ is a dynamic inexact oracle is similar to that of Proposition 2.

Assumption 2 (Generalized Contraction): Let μ and β be constants satisfying $0 < \mu \le \beta$. Then there exist $P \succ 0$, $\eta_2 > 0$, and $\rho_2 \in [0, 1)$ such that

$$\|A_{\text{io}}\xi + B_{\text{io}}w\|_P \le \rho_2\|\xi\|_P$$

for all w satisfying $(v, w) \in \mathsf{sec}(\mu, \beta)$, where $v = C_{\text{io}}\xi$.

Assumption 2 can often be verified for a given first-order algorithm. For instance, when $\mathcal{G}_{\text{io}}$ is realized by gradient descent, i.e., when $c_1 = c_2 = c_3 = 0$ in (10), the second component ξ_2 of ξ becomes irrelevant and can be dropped, from which we obtain (with an abuse of notion) $A_{\text{io}} = I$, $B_{\text{io}} = -\eta_2 I$, and $C_{\text{io}} = I$. In other words, the inexact oracle $\mathcal{G}_{\text{io}}$ becomes equivalent to $\mathcal{G}_{\text{gd}}$. Therefore, by Lemma 2, Assumption 2 is satisfied for $P = I$. For other first-order algorithms, verifying Assumption 2 is equivalent to checking absolute stability under a sector-bounded uncertainty. While it is generally difficult to obtain closed-form expressions of P, η_2, and ρ_2 that satisfy Assumption 2, a numerical method [16, Figs. 3 and 5] has been used to find P, η_2, and ρ_2 that certify that both Nesterov's accelerated gradient method and the heavy-ball method satisfy the assumption, at least when the condition number β/μ is small.

E. Other Inexact Oracles in the Literature

The notion of dynamic inexact oracles is more general than the inexact oracles studied in the existing literature, which are memoryless inexact oracles. In particular, the dynamic oracles considered herein should not be confused with the time-varying memoryless oracles studied in time-varying

optimization [3]. For a memoryless oracle, the output of the oracle only depends on the instantaneous input. Incorporating dynamics into inexact oracles is necessary because a memoryless oracle is not able to model iterative optimization algorithms with warm starts, in which the solution during the current iteration needs to be memorized to initialize the next iteration such as in (7).

One example of memoryless inexact oracles is approximate gradient mappings used in first-order methods, such as in the ϵ-(sub)gradient method (see [4, Ch. 3.3]). Other examples include approximate proximal operators used in the proximal point algorithm [21, p. 880] and in the Douglas–Rachford splitting method [11, Th. 8]. A general treatment of memoryless inexact oracles in first-order methods can be found in [9].

IV. Convergence Analysis

We observe that gradient methods with dynamic inexact oracles can be viewed as a feedback interconnection of two dynamical systems. By applying the small-gain principle, we develop a unified convergence analysis that is applicable to inexact oracles constructed from a range of first-order optimization algorithms.

A. The Oracle Based on Gradient Descent

For the purpose of illustration, we begin by analyzing the convergence of the gradient method (8) with the dynamic inexact oracle $\mathcal{G}_{\rm gd}$ constructed from gradient descent as given in (7). We define the error e between the inexact and the exact oracles as $e^k := y^k - \nabla g^*(Ax^k)$ and rewrite (7) and (8) as

$$x^{k+1} = x^k - \eta_1 \nabla p(x^k) - \eta_1 A^T e^k \tag{11a}$$

$$e^{k+1} = e^k - \eta_2 \nabla_1 \tilde{g}(y^k, x^k) - [\nabla g^*(Ax^{k+1}) - \nabla g^*(Ax^k)]. \tag{11b}$$

Although the gradient update (11a) converges when the error $e \equiv 0$, and the error dynamics in (11b) converge when $x \equiv x^\star$ (Proposition 2), the joint recursion (11) is not guaranteed to converge. Indeed, the joint recursion (11) can be viewed as a feedback interconnection of two dynamical systems (11a) and (11b) as illustrated in Fig. 1, and it is well known in control theory that a feedback connection of two internally stable systems may be unstable.

A powerful method for analyzing the stability of feedback interconnections of dynamical systems is the small-gain principle. The small-gain principle can take various forms depending on the specific setup. The following is what we will use in this letter. See Appendix C for a detailed proof.

Lemma 3 (Small-Gain): Let $\{s_1^k\}$ and $\{s_2^k\}$ be two nonnegative real-valued sequences satisfying

$$s_1^{k+1} \le \gamma_{11} s_1^k + \gamma_{12} s_2^k$$
$$s_2^{k+1} \le \gamma_{21} s_1^k + \gamma_{22} s_2^k$$

for some nonnegative constants γ_{ij} $(i, j = 1, 2)$. Then, both $\{s_1^k\}$ and $\{s_2^k\}$ converge exponentially to 0 if $\gamma_{11} < 1$, $\gamma_{22} < 1$, and $\gamma_{12}\gamma_{21} < (1 - \gamma_{11})(1 - \gamma_{22})$.

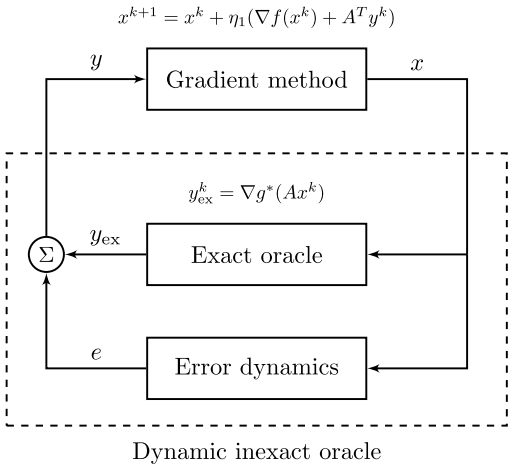

Fig. 1. The gradient method with a dynamic inexact oracle. The exact (gradient) oracle computes $\phi(x^k) := \arg\min_y \tilde{g}(y, x^k) = \nabla g^*(Ax^k)$. The difference between the inexact and the exact oracles is characterized by the additive error dynamics.

The small-gain lemma shows that, in order for the feedback interconnection of two (nonnegative) systems to be stable, aside from the stability of individual systems ($\gamma_{11} < 1$ and $\gamma_{22} < 1$), the coupling coefficients γ_{12} and γ_{21} must be small enough. We can apply the small-gain lemma to establish the convergence of the PDGM, viewed as an inexact gradient method (8) with the oracle $\mathcal{G}_{\rm gd}$ given by (7).

Theorem 1: Consider the gradient method given by (8), where $\{y^k\}$ is given by the dynamic inexact oracle $\mathcal{G}_{\rm gd}$ defined by (7) with $\eta_2 \in (0, 2/(\mu_g + \beta_g)]$. Suppose f, g, and A satisfy Assumption 1, and let $\beta_\phi = \sigma_{\max}/\mu_g$, $\alpha_p = \mu_p\beta_p/(\mu_p + \beta_p)$, and $\alpha_g = \mu_g\beta_g/(\mu_g + \beta_g)$. Then, for any η_1 satisfying

$$0 < \eta_1 < \min\left\{\frac{\alpha_p \alpha_g \eta_2}{\sigma_{\max}\beta_\phi(\alpha_p + \beta_p)}, \frac{2}{\mu_p + \beta_p}\right\}, \tag{12}$$

the sequences $\{x^k\}$ and $\{y^k\}$ converge exponentially to the primal and dual optimal solutions $x^\star$ and $y^\star$, respectively.

Proof: Denote by $e^\star$ the steady-state value of e. Then, we have $e^\star = y^\star - \nabla g^*(Ax^\star) = 0$. Define $\hat{x}^k := x^k - x^\star$ and $\hat{e}^k := e^k - e^\star$. We can rewrite (11) as

$$\hat{x}^{k+1} = \hat{x}^k - \eta_1 \nabla p(x^k) - \eta_1 A^T \hat{e}^k$$
$$\hat{e}^{k+1} = \hat{e}^k - \eta_2 \nabla_1 \tilde{g}(y^k, x^k) - [\nabla g^*(Ax^{k+1}) - \nabla g^*(Ax^k)].$$

Because $p \in \mathcal{S}(\mu_p, \beta_p)$ and $g \in \mathcal{S}(\mu_g, \beta_g)$, from Proposition 1, we have $(\hat{x}^k, \nabla p(x^k)) = (x^k - x^\star, \nabla p(x^k) - \nabla p(x^\star)) \in \mathsf{sec}(\mu_p, \beta_p)$ and $(\hat{e}^k, \nabla_1 \tilde{g}(y^k, x^k)) = (y^k - \nabla g^*(Ax^k), \nabla g(y^k) - Ax^k) \in \mathsf{sec}(\mu_g, \beta_g)$, where we have used the fact $Ax^k = \nabla g(\nabla g^*(Ax^k))$. Applying Lemma 2, since $0 < \eta_1 \le 2/(\mu_p + \beta_p)$, we have

$$\begin{aligned}\|\hat{x}^{k+1}\| &\le \|\hat{x}^k - \eta_1 \nabla p(x^k)\| + \|\eta_1 A^T \hat{e}^k\| \\ &\le \rho_1 \|\hat{x}^k\| + \eta_1 \sigma_{\max} \|\hat{e}^k\|,\end{aligned} \tag{13}$$

where $\rho_1 = 1 - \alpha_p \eta_1$; similarly, we also obtain

$$\begin{aligned}\|\hat{e}^{k+1}\| &\le \rho_2 \|\hat{e}^k\| + \beta_\phi \|x^{k+1} - x^k\| \\ &= \rho_2 \|\hat{e}^k\| + \beta_\phi \|-\eta_1 \nabla p(x^k) - \eta_1 A^T \hat{e}^k\| \\ &\le \eta_1 \beta_\phi \beta_p \|\hat{x}^k\| + (\rho_2 + \eta_1 \beta_\phi \sigma_{\max}) \|\hat{e}^k\|,\end{aligned} \tag{14}$$

where $\rho_2 = 1 - \alpha_g \eta_2$, and $\beta_\phi = \sigma_{\max}/\mu_g$ is the Lipschitz constant of the mapping $\phi\colon x^k \mapsto \nabla g^*(Ax^k)$ in (9). The relationship given by (13) and (14) allows us to apply the small-gain lemma (Lemma 3) and derive the condition (12) ensuring that both $\hat{x}$ and $\hat{e}$ converge exponentially to 0, i.e., $x^k \to x^\star$ and $y^k \to \nabla g^*(Ax^\star) = y^\star$ as required. ■

Although exponential convergence of the PDGM has already been established [10], the technique used in the proof of Theorem 1 is different from those in the existing literature. The proof reveals two attractive features of the small-gain principle in the analysis of the inexact gradient method. First, it is capable of incorporating existing convergence results, i.e., internal stability of the gradient dynamics (11a) and (11b) as manifested in Lemma 2. This simplifies the construction of a Lyapunov function for proving convergence, which is often otherwise nontrivial except for the simplest algorithms. Second, the small-gain analysis only relies on a coarse description of the input-output behavior of the error dynamics such as what is given in (14). Therefore, when the dynamic inexact oracle is realized by a different iterative algorithm, the small-gain analysis can be readily applied as long as a relationship between x and e similar to (14) can be derived for the error dynamics. The "plug-and-play" nature of this approach allows us to easily generalize the analysis to a wide range of dynamic inexact oracles, which we will discuss shortly in Section IV-B.

B. Oracles Based on General First-Order Algorithms

For a dynamic inexact oracle $\mathcal{G}_{\text{io}}$ constructed from more general first-order algorithms in (10), convergence of the inexact gradient method (8) can be established using a small-gain analysis similar to the proof of Theorem 1. A detailed proof can be found in Appendix D. The key is deriving (17) for the error dynamics, which is needed for applying the small-gain lemma.

Theorem 2: Consider the gradient method given by (8), where $\{y^k\}$ is given by a dynamic inexact oracle $\mathcal{G}_{\text{io}}$ of the form (10). Suppose f, g, and A satisfy Assumption 1, and A_{io}, B_{io}, and C_{io} satisfy Assumption 2. Then there exists η_1 such that $\{x^k\}$ and $\{y^k\}$ converge exponentially to the primal and dual optimal solutions $x^\star$ and $y^\star$, respectively.

As a concrete instance of Theorem 2, we give a convergence result for the case where $\mathcal{G}_{\text{io}}$ is realized by Nesterov's accelerated gradient method.

Corollary 1 (Accelerated Primal-Dual Method): Let $\gamma = (\sqrt{\beta_g} - \sqrt{\mu_g})/(\sqrt{\beta_g} + \sqrt{\mu_g})$ and $\eta_2 = 1/\beta_g$. Consider the gradient method given by (8), where $\{y^k\}$ is given by a dynamic inexact oracle realized by Nesterov's accelerated gradient method:

$$\begin{aligned} y^{k+1} &= v^k - \eta_2(\nabla g(v^k) - Ax^k) \\ v^{k+1} &= (1+\gamma)y^{k+1} - \gamma y^k. \end{aligned} \tag{15}$$

Suppose f, g, and A satisfy Assumption 1. Then there exists η_1 such that $\{x^k\}$ and $\{y^k\}$ converge exponentially to the primal and dual optimal solutions $x^\star$ and $y^\star$, respectively, when β_g/μ_g is small enough.

Proof: The recursion (15) can be derived from (10) by setting $c_1 = c_2 = \gamma$ and $c_3 = 0$ followed by eliminating ξ. Under the given choice of γ and η_2, it has been shown in [16, Fig. 3] that Assumption 2 holds when β_g/μ_g is small enough. The corollary then follows from Theorem 2. ■

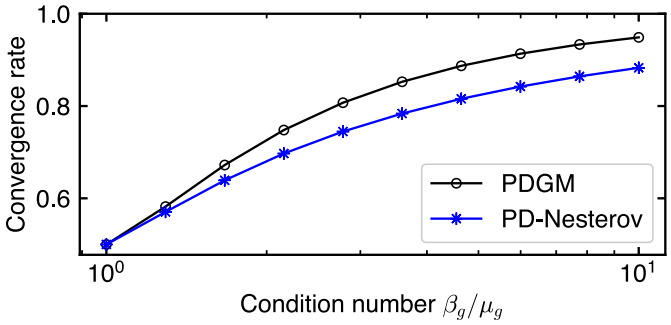

Fig. 2. Convergence rate of the gradient method in (8) with different inexact oracles: gradient descent (in black, which is equivalent to the PDGM) and Nesterov's accelerated method (in blue).

For a numerical comparison between the method in Corollary 1 and the PDGM, we considered a simple case where f is linear, and g is convex quadratic. Under the choice of f and g, the dynamics of both methods become linear, which implies that the convergence rate can be found by computing eigenvalues. For both methods, we fixed $\eta_2 = 1/\beta_g$ and chose η_1 via a grid search to achieve the best exponential convergence rate. Fig. 2 shows the convergence rate under different condition numbers β_g/μ_g. It can be seen that the method in Corollary 1 (referred to as "PD-Nesterov") not only ensures convergence but also achieves a faster convergence rate compared to the PDGM.

V. Conclusion

We have studied the convergence of inexact gradient methods in which the gradient is provided by a dynamic inexact oracle. Dynamic inexact oracles naturally arise when the gradient method is applied to minimax optimization problems, in which computing the gradient requires solving the inner optimization problem; a dynamic inexact oracle approximates the exact solution of the inner optimization problem by the output from an iterative optimization algorithm. Changing the realization of the dynamic inexact oracle leads to different algorithms for solving minimax problems. For instance, when the inexact oracle is realized by one step of gradient descent with warm starts, the corresponding inexact gradient method recovers the PDGM, a common algorithm for solving minimax problems. Using the small-gain principle, we have derived a unified convergence analysis applicable to a range of inexact oracles that can be used for creating new primal-dual algorithms.

Appendix A
Proof of Lemma 1

The optimality condition of the minimization problem in (6a) gives

$$0 = \nabla_1 \tilde{g}(y^k_{\text{ex}}, x^k_{\text{ex}}) = \nabla g(y^k_{\text{ex}}) - Ax^k_{\text{ex}}.$$

Because $g \in \mathcal{S}(\mu_g, \beta_g)$, using Proposition 1, we obtain

$$\phi(x^k_{\text{ex}}) = y^k_{\text{ex}} = (\nabla g)^{-1}(Ax^k_{\text{ex}}) = \nabla g^*(Ax^k_{\text{ex}}).$$

■

APPENDIX B
PROOF OF LEMMA 2

From [19, Th. 2.1.15], we have $\|\xi - \eta w\|^2 \leq (1-2\eta\alpha)\|\xi\|^2$. The result follows from the fact $(1-2\eta\alpha)^{1/2} \leq 1-\eta\alpha$. ■

APPENDIX C
PROOF OF LEMMA 3

Consider a single-input single-output linear system whose input u and output y are described by $y^{k+1} = ay^k + bu^k$, where $a \in [0, 1)$ and $b \geq 0$. It can be shown that the ℓ_2-gain of the system is given by $b/(1-a)$. The result then follows from the (usual) small-gain theorem for feedback interconnections (see [14, Th. 5.6]) and the (discrete-time) comparison lemma (see [15, Th. 1.9.1]). ■

APPENDIX D
PROOF THEOREM 2

Define $\bar{\xi}_i^k := \xi_i^k - \nabla g^*(Ax^k)$ $(i = 1, 2)$, $\bar{v}^k := v^k - \nabla g^*(Ax^k)$, and $e^k := y^k - \nabla g^*(Ax^k)$. In the new variables, the dynamics (8) can be rewritten as (11a), and the dynamics (10) of $\mathcal{G}_{\text{io}}$ can be rewritten as

$$\bar{\xi}^{k+1} = A_{\text{io}}\bar{\xi}^k + B_{\text{io}}\nabla_1\tilde{g}(v^k, x^k) + B_\phi\Big[\phi(x^{k+1}) - \phi(x^k)\Big]$$
$$\bar{v}^k = C_{\text{io}}\bar{\xi}^k, \qquad e^k = E_{\text{io}}\bar{\xi}^k,$$

where $B_\phi = -[I \;\; I]^T$. We have used the fact $\phi(x^k) = \nabla g^*(Ax^k)$ from Lemma 1.

Because $g \in \mathcal{S}(\mu_g, \beta_g)$, using Proposition 1, we have $(\bar{v}^k, \nabla_1\tilde{g}(v^k, x^k)) = (v^k - \nabla g^*(Ax^k), \nabla g(v^k) - Ax^k) \in \mathsf{sec}(\mu_g, \beta_g)$. Since Assumption 2 holds, we have

$$\|\bar{\xi}^{k+1}\|_P \leq \rho_2\|\bar{\xi}^k\|_P + c_\phi\|x^{k+1} - x^k\| \tag{16}$$

for some $P \succ 0$, $\rho_2 \in [0, 1)$, and $c_\phi > 0$. The existence of c_ϕ is ensured by the Lipschitz continuity of ϕ and the equivalence of norms in finite dimensions.

The second term on the right side of (16) can be further bounded by making use of (11a). Let $\hat{x}^k := x^k - x^\star$, we have

$$\begin{aligned}\|x^{k+1} - x^k\| &= \eta_1\|\nabla p(x^k) + A^T e^k\| \\ &= \eta_1\|\nabla p(x^k) + A^T E_{\text{io}}\bar{\xi}^k\| \\ &\leq \eta_1(\beta_p\|\hat{x}^k\| + c_\xi\|\bar{\xi}^k\|_P)\end{aligned}$$

for some $c_\xi > 0$, where we have used the equivalence of norms again. Substituting this into (16), we have

$$\|\bar{\xi}^{k+1}\|_P \leq \eta_1 c_\phi\beta_p\|\hat{x}^k\| + (\rho_2 + \eta_1 c_\xi c_\phi)\|\bar{\xi}^k\|_P. \tag{17}$$

In the meantime, because the x-update (11a) is given by the gradient method, when $\eta_1 \in (0, 2/(\mu_p + \beta_p)]$, similar to (13), we have

$$\|\hat{x}^{k+1}\| \leq \rho_1\|\hat{x}^k\| + \eta_1 c_\xi\|\bar{\xi}^k\|_P, \tag{18}$$

where $\rho_1 = 1 - \alpha_p\eta_1$ for α_p defined in Theorem 1.

Apply the small-gain lemma (Lemma 3) to (17) and (18). In order to ensure convergence, we need

$$\rho_1 = 1 - \alpha_p\eta_1 < 1, \qquad \rho_2 + \eta_1 c_\xi c_\phi < 1$$
$$\eta_1 c_\xi \cdot \eta_1 c_\phi\beta_p < (1-\rho_1)(1-\rho_2-\eta_1 c_\xi c_\phi). \tag{19}$$

A straightforward algebraic manipulation shows that the last condition in (19) is equivalent to $\eta_1 < \alpha_p(1-\rho_2)/(c_\xi c_\phi(\alpha_p + \beta_p))$. Therefore, when η_1 is small enough and strictly positive, all the conditions in (19) are satisfied, which implies that the joint recursion consisting of (8) and (10) converges exponentially. ■

REFERENCES

[1] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*. Stanford, CA, USA: Stanford Univ. Press, 1958.

[2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[3] A. Bernstein, E. Dall'Anese, and A. Simonetto, "Online primal-dual methods with measurement feedback for time-varying convex optimization," *IEEE Trans. Signal Process.*, vol. 67, no. 8, pp. 1978–1991, Apr. 2019.

[4] D. Bertsekas, *Convex Optimization Algorithms*. Nashua, NH, USA: Athena Sci., 2015.

[5] D. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA, USA: Athena Sci., 2003.

[6] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[7] A. Cherukuri, B. Gharesifard, and J. Cortés, "Saddle-point dynamics: Conditions for asymptotic stability of saddle points," *SIAM J. Control Optim.*, vol. 55, no. 1, pp. 486–511, 2017.

[8] C. Daskalakis and I. Panageas, "The limit points of (optimistic) gradient descent in min–max optimization," in *Advances in Neural Information Processing Systems*, vol. 31. Red Hook, NY, USA: Curran Assoc., Inc., 2018, pp. 9236–9246.

[9] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Math. Program.*, vol. 146, nos. 1–2, pp. 37–75, 2014.

[10] S. S. Du and W. Hu, "Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity," in *Proc. Mach. Learn. Res.*, vol. 89, pp. 196–205, 2019.

[11] J. Eckstein and D. P. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, nos. 1–3, pp. 293–318, 1992.

[12] D. Feijer and F. Paganini, "Stability of primal–dual gradient dynamics and applications to network optimization," *Automatica*, vol. 46, no. 12, pp. 1974–1981, 2010.

[13] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*. Berlin, Germany: Springer-Verlag, 1993.

[14] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[15] V. Lakshmikantham, S. Leela, and A. Martynyuk, *Stability Analysis of Nonlinear Systems*, 2nd ed. Cham, Switzerland: Springer, 2015.

[16] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM J. Optim.*, vol. 26, no. 1, pp. 57–95, 2016.

[17] R. Myerson, *Game Theory: Analysis of Conflict*. Cumberland, RI, USA: Harvard Univ. Press, 2013.

[18] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *J. Optim. Theory Appl.*, vol. 142, no. 1, pp. 205–228, 2009.

[19] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Boston, MA, USA: Kluwer Acad. Publ., 2004.

[20] G. Qu and N. Li, "On the exponential stability of primal-dual gradient dynamics," *IEEE Control Syst. Lett.*, vol. 3, no. 1, pp. 43–48, Jan. 2019.

[21] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM J. Control Optim.*, vol. 14, no. 5, pp. 877–898, 1976.

[22] B. Van Scoy, R. A. Freeman, and K. M. Lynch, "The fastest known globally convergent first-order method for minimizing strongly convex function," *IEEE Control Syst. Lett.*, vol. 2, no. 1, pp. 49–54, Jan. 2018.

[23] Y. Zhang and L. Xiao, "Stochastic primal-dual coordinate method for regularized empirical risk minimization," *J. Mach. Learn. Res.*, vol. 18, pp. 1–42, Jul. 2017.