Noname manuscript No.
(will be inserted by the editor)

An optimal randomized incremental gradient method

Guanghui Lan · Yi Zhou

the date of receipt and acceptance should be inserted later

Abstract In this paper, we consider a class of finite-sum convex optimization problems whose objective function is given by the summation of m ($\geq$ 1) smooth components together with some other relatively simple terms. We first introduce a deterministic primal-dual gradient (PDG) method that can achieve the optimal black-box iteration complexity for solving these composite optimization problems using a primal-dual termination criterion. Our major contribution is to develop a randomized primal-dual gradient (RPDG) method, which needs to compute the gradient of only one randomly selected smooth component at each iteration, but can possibly achieve better complexity than PDG in terms of the total number of gradient evaluations. More specifically, we show that the total number of gradient evaluations performed by RPDG can be $\mathcal{O}(\sqrt{m})$ times smaller, both in expectation and with high probability, than those performed by deterministic optimal first-order methods under favorable situations. We also show that the complexity of the RPDG method is not improvable by developing a new lower complexity bound for a general class of randomized methods for solving large-scale finite-sum convex optimization problems. Moreover, through the development of PDG and RPDG, we introduce a novel game-theoretic interpretation for these optimal methods for convex optimization.

Keywords: convex programming, complexity, incremental gradient, primal-dual gradient method, Nesterov's method, data analysis

AMS 2000 subject classification: 90C25, 90C06, 90C22, 49M37

1 Introduction

The basic problem of interest in this paper is the convex programming (CP) problem given by

$$\Psi^* := \min_{x\in X} \left\{\Psi(x) := \textstyle\sum_{i=1}^m f_i(x) + h(x) + \mu\,\omega(x)\right\}. \tag{1.1}$$

Here, $X \subseteq \mathbb{R}^n$ is a closed convex set, h is a relatively simple convex function, $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, are smooth convex functions with Lipschitz continuous gradient, i.e., $\exists L_i \geq 0$ such that

$$\|\nabla f_i(x_1) - \nabla f_i(x_2)\|_* \leq L_i\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^n, \tag{1.2}$$

$\omega : X \to \mathbb{R}$ is a strongly convex function with modulus 1 w.r.t. an arbitrary norm $\|\cdot\|$, i.e.,

$$\langle \omega'(x_1) - \omega'(x_2), x_1 - x_2\rangle \geq \tfrac{1}{2}\|x_1 - x_2\|^2, \quad \forall x_1, x_2 \in X, \tag{1.3}$$

The author of this paper was partially supported by NSF grant CMMI-1537414, DMS-1319050, ONR grant N00014-13-1-0036 and NSF CAREER Award CMMI-1254446.

Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611. (email: glan@ise.ufl.edu).

Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611. (email: yizhou@ufl.edu).

Address(es) of author(s) should be given

and $\mu \geq 0$ is a given constant. Hence, the objective function Ψ is strongly convex whenever $\mu > 0$. For notational convenience, we also denote $f(x) \equiv \sum_{i=1}^m f_i(x)$ and $L \equiv \sum_{i=1}^m L_i$. It is easy to see that for some $L_f \geq 0$,

$$\|\nabla f(x_1) - \nabla f(x_2)\|_* \leq L_f \|x_1 - x_2\| \leq L\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^n. \tag{1.4}$$

Throughout this paper, we assume subproblems of the form

$$\operatorname{argmin}_{x \in X} \langle g, x\rangle + h(x) + \mu\, \omega(x) \tag{1.5}$$

are easy to solve. CP given in the form of (1.1) has recently found a wide range of applications in machine learning, statistics, and image processing, and hence becomes the subject of intensive studies during the past few years.

Stochastic (sub)gradient descent (SGD) (a.k.a. stochastic approximation (SA)) type methods have been proven useful to solve problems given in the form of (1.1). SGD was originally designed to solve stochastic optimization problems given by

$$\min_{x \in X} \mathbb{E}_\xi[F(x, \xi)], \tag{1.6}$$

where ξ is a random variable with support $\Xi \subseteq \mathbb{R}^d$. Problem (1.1) can be viewed as a special case of (1.6) by setting ξ to be a discrete random variable supported on $\{1, \ldots, m\}$ with $\text{Prob}\{\xi = i\} = \nu_i$ and $F(x, i) = \nu_i^{-1} f_i(x) + h(x) + \mu\omega(x)$, $i = 1, \ldots, m$. Since each iteration of SGDs needs to compute the (sub)gradient of only one randomly selected f_i [1], their iteration cost is significantly smaller than that for deterministic first-order methods (FOM), which involves the computation of first-order information of f and thus all the m (sub)gradients of f_i's. Moreover, when f_i's are general nonsmooth convex functions, by properly specifying the probabilities ν_i, $i = 1, \ldots, m$ [2], it can be shown (see [25]) that the iteration complexities for both SGD and FOM are in the same order of magnitude. Consequently, the total number of subgradients required by SGDs can be m times smaller than those by FOMs.

Note however, that there is a significant gap on the complexity bounds between SGDs and deterministic FOMs if f_i's are smooth convex functions. For the sake of simplicity, let us focus on the strongly convex case when $\mu > 0$ and let x^* be the optimal solution of (1.1). In order to find a solution $\bar{x} \in X$ s.t. $\|\bar{x} - x^*\|^2 \leq \epsilon$, the total number of gradient evaluations for f_i's performed by optimal FOMs can be bounded by

$$\mathcal{O}\left\{m\sqrt{\tfrac{L}{\mu}} \log \tfrac{1}{\epsilon}\right\}, \tag{1.7}$$

which was first achieved by the well-known Nesterov's accelerated gradient method [27,28], see also relevant extensions in [31,4,35]. On the other hand, a direct application of optimal SGDs to the aforementioned stochastic optimization reformulation of (1.1) would yield an

$$\mathcal{O}\left\{\sqrt{\tfrac{L}{\mu}} \log \tfrac{1}{\epsilon} + \tfrac{\sigma^2}{\mu\epsilon}\right\} \tag{1.8}$$

iteration complexity bound on the number of gradient evaluations for f_i's, which was first achieved by the accelerated stochastic approximation method ([19,14,15]). Here $\sigma > 0$ denotes variance of the stochastic gradients. Clearly, the latter bound is significantly better than the one in (1.7) in terms of its dependence on m, but much worse in terms of its dependence on accuracy ϵ and a few other problem parameters (e.g., L and μ).

It should be noted that the optimality of (1.8) for general stochastic programming (1.6) does not preclude the existence of more efficient algorithms for solving (1.1), because (1.1) is a special case of (1.6) with finite support Ξ. Last few years have seen very active and fruitful research in this field (e.g., [32,17,12,34,36]). In particular, Schmidt, Roux and Bach [32] presented a stochastic average gradient (SAG) method, which recursively computes an estimator of ∇f by aggregating the gradient of a randomly selected f_i with some other previously computed gradient information. They proved that the complexity of SAG is bounded by $\mathcal{O}\left((m + L/\mu)\log \frac{1}{\epsilon}\right)$, see also Johnson and Zhang [17] and Defazio et al. [12] for similar complexity results for solving (1.1). In a related but different line of research, Shalev-Shwartz and Zhang [34] studied a special class of CP problems given in the form of (1.1) with $f_i(x)$ given by $\phi_i(a_i^T x)$, where a_i denotes an affine mapping. Under the assumption that $\omega(x) = \|x\|_2^2$, they presented an accelerated stochastic dual coordinate ascent (A-SDCA) method, obtained by properly restarting a stochastic coordinate ascent method in

[1] Observe that the subgradients of h and ω are not required due to the assumption in (1.5).

[2] Suppose that f_i are Lipschitz continuous with constants M_i and let us denote $M := \sum_{i=1}^m M_i$, we should set $\nu_i = M_i/M$ in order to get the optimal complexity for SGDs.

[33] applied to the dual of (1.1). Shalev-Shwartz and Zhang show that the iteration complexity of this method can be bounded by $\mathcal{O}\left\{\left(m+\sqrt{\frac{mL}{\mu}}\right)\log\frac{1}{\epsilon}\right\}$. However, each iteration of A-SDCA requires, instead of the computation of ∇f_i, the solution of a subproblem given in the form of

$$\operatorname{argmin}\{\langle g, y\rangle + \phi_i^*(y) + \|y\|_*^2\}, \tag{1.9}$$

where ϕ_i^* denotes the conjugate function of ϕ_i. Moreover, these methods were also designed for solving a more special class of problems than (1.1). More recently, Lin, Lu, and Xiao [23] proposed to apply the accelerated coordinate descent methods by Nesterov [30], and Fercoq and Richtáriks [13] to obtain similar results for solving these "regularized empirical loss functions" as in [34]. Zhang and Xiao [36] had also obtained similar results by using different stochastic primal-dual coordinate decomposition techniques.

In this paper, we focus on randomized incremental gradient methods that can access the first-order information of only one randomly selected smooth component f_i at each iteration (see Bertsekas [5] for an introduction to incremental gradient methods). It should be noted that while the algorithms in [32,17,12] belong to incremental gradient methods, generally speaking, the dual coordinate algorithms in [23,34,36] cannot be considered as incremental gradient methods because they require the solutions of a different subproblem rather than the computation of the gradient of f_i. The previous attempts to improve the complexity of the existing incremental gradient methods, e.g., based on the extrapolation idea in Nesterov [27], however, turned out to be tricky and unsuccessful, see Section 1.2 of Bertsekas [5] and Section 5 of Agarwal and Bottou [1] for more discussions. Another important yet unresolved issue is that there does not exist a valid lower complexity bound for randomized incremental gradient methods in the literature. Hence, it remains unknown what would be the best possible performance that one can expect for these types of methods. Regarding this question, Agarwal and Bottou [1] recently suggested a lower complexity bound for solving problems given in the form of (1.1). However, as pointed out by them in a recent ISMP talk in 2015, the lower complexity bound in [1] is deterministic by construction, and hence cannot be used to justify the optimality or suboptimality for the randomized incremental gradient methods in [32,17,12] or dual coordinate methods in [23,34,36].

Our contribution in this paper mainly lies on the following several aspects. Firstly, we present a new class of deterministic FOMs, referred to as the primal-dual gradient (PDG) methods, which can achieve the optimal black-box iteration complexity in (1.7) for solving (1.1). The novelty of these methods exists in: 1) a proper reformulation of (1.1) as a primal-dual saddle point problem and 2) the incorporation of a new non-differentiable prox-function (or Bregman distance) based on the conjugate functions of f_i in the dual space. As a consequence, we are able to show that the PDG method covers a variant of the well-known Nesterov's accelerated gradient method as a special case. In particular, the computation of the gradient at the extrapolation point of the accelerated gradient method is equivalent to a primal prediction step combined with a dual ascent step (employed with the aforementioned dual prox-function) in the PDG method. While it is often difficult to interpret Nesterov's method, the development of the PDG method allows us to view this method as a natural iterative buyer-supplier game. Such a game-theoretic view of the accelerated gradient method seems to be new in the literature. In fact, the obtained complexity results for the PDG method are slightly stronger than the one in (1.7) and those in [27,28] for Nesterov's accelerated gradient method, because a stronger primal-dual termination criterion has been used in our analysis.

Secondly, we develop a randomized primal-dual gradient (RPDG) method, which is an incremental gradient method using only one randomly selected component ∇f_i at each iteration. A variant of PDG, this algorithm incorporates an additional dual prediction step before performing the primal descent step (with a properly defined primal prox-function). We prove that the number of iterations (and hence the number of gradients) required by RPDG is bounded by

$$\mathcal{O}\left(\left(m+\sqrt{\frac{mL}{\mu}}\right)\log\frac{1}{\epsilon}\right), \tag{1.10}$$

both in expectation and with high probability. The complexity bounds of the RPDG method are established in terms of not only the distance from the iterate x^k to the optimal solution, but also the primal optimality gap based on the ergodic mean of the iterates. In comparison with the accelerated stochastic dual coordinate ascent method in [34], RPDG deals with a wider class of problems and can be applied to the cases when f_i's involve a more complicated composite structure (see examples in [5]) and/or a more general regularization term ω that is strongly convex with respect to an arbitrary norm (see open problems in Section 7 of [34]). Moreover, each iteration of RPDG only involves the computation ∇f_i, rather than the more complicated subproblem in (1.9), which sometimes may not have explicit solutions [34] (e.g., the logistics regression problem). The RPDG method also admits an interesting game theoretic interpretation, implying

that by properly incorporating randomization, the buyer and supplier can reach the equilibrium with possibly fewer price changes at the expense of more order transactions.

Thirdly, we show that the number of gradient evaluations required by any randomized incremental gradient methods to find an ϵ-solution of (1.1), i.e., a point $\bar{x} \in X$ s.t. $\mathbb{E}[\|\bar{x} - x^*\|_2^2] \leq \epsilon$, cannot be smaller than

$$\Omega\left(\left(m + \sqrt{\tfrac{mL}{\mu}}\right)\log\tfrac{1}{\epsilon}\right), \tag{1.11}$$

whenever the dimension n is sufficiently large. This bound is obtained by carefully constructing a special class of separable quadratic programming problems and tightly bounding the expected distance to the optimal solution for any arbitrary distribution used to choose f_i at each iteration. Comparing (1.10) with (1.11), we conclude that the complexity of the RPDG method is optimal if n is large enough. To the best of our knowledge, this is the first time that such a lower complexity bound has been presented for randomized incremental gradient methods in the literature. As a byproduct, we also derived a lower complexity bound for randomized block coordinate descent methods by utilizing the separable structure of the aforementioned worst-case instances. These methods have been intensively studied recently, but a valid lower complexity bound is still missing in the literature.

Finally, we generalize RPDG for problems which are not necessarily strongly convex (i.e., $\mu = 0$) and/or involve structured nonsmooth terms f_i. We show that for all these cases, the RPDG can save $\mathcal{O}(\sqrt{m})$ times gradient computations (up to certain logarithmic factors) in comparison with the corresponding optimal deterministic FOMs. In particular, we show that when both the primal and dual of (1.1) are not strongly convex, the total number of iterations performed by the RPDG method can be bounded by $\mathcal{O}(\sqrt{m}/\epsilon)$ (up to some logarithmic factors), which is $\mathcal{O}(\sqrt{m})$ times better, in terms of the total number of dual subproblems to be solved, than Nesterov's smoothing technique [29], Nemirovski's mirror-prox method [24], or Chambolle and Pock's primal-dual method [8]. It seems that this complexity result has not been obtained before in the literature.

It is worth mentioning a few relevant works to our development. The most two related ones are conducted independently by Dang and Lan [11], and Zhang and Xiao [36]. Both of these papers deal with randomized variants of the primal-dual method presented by Chambolle and Pock [8] (see also extensions in [10]) for solving saddle point problems. Zhang and Xiao's development [36] was based on a variant of the primal-dual method for solving strongly convex saddle point problems [8]. They were able to show that a block-wise randomized version of the algorithm can achieve similar complexity as the A-SDCA method in [34]. Since Zhang and Xiao's algorithm targets for solving a similar class of problems and requires the solutions of a similar subproblem to [34], it appears that the aforementioned possible advantages of RPDG over A-SDCA are also applicable to the stochastic primal-dual coordinate method in [36]. Moreover, the complexity bound of Zhang and Xiao's algorithm is only established in terms of the Euclidean distances of the iterate x^k, y^k to the optimal solution. They did not deal with the convergence of the ergodic mean of iterates. On the other hand, Dang and Lan's work was motivated by the observation in [9] that a direct extension of the alternating direction method of multiplier (ADMM) does not converge for multi-block problems. Their work in [11] then focuses on the non-strongly convex case and shows that a randomized primal-dual method, which is equivalent to a randomized pre-conditioned ADMM for linear constrained problems, does converge for multi-block problems. Without incorporating the aforementioned dual prediction step, the complexity obtained in [11] is $\mathcal{O}(\sqrt{m})$ times worse than Chambolle and Pock's method. Nevertheless, this is the first time that randomized algorithms for saddle point optimization with an $\mathcal{O}(1/\epsilon)$ complexity has been presented in the literature. More recently, close to the end of the preparation of this paper, we notice that Lin, Mairal, and Harchaoui [22] in a concurrent work presented a catalyst scheme that can be used to accelerate the SAG method in [32] and thus possibly achieve the complexity bound in (1.10) (under the Euclidean setting). While their approach is an indirect one obtained by properly restarting SAG (or other "non-accelerated" first-order methods), the proposed randomized primal-dual gradient method is a direct approach with a "built-in" acceleration. Also none of these works [11,36,22] discussed the lower complexity bound for randomized methods.

This paper is organized as follows. We first study the deterministic primal-dual method in Section 2. Section 3 is devoted to the design and analysis of the randomized primal-dual method for the strongly convex case, as well as the development of the lower complexity bound in (1.11). In Section 4, we generalize the RPDG method to different classes of CP problems that are not necessarily strongly convex. Important technical results and proofs of the main theorems in Sections 2 and 3 are provided in Section 5. Some brief concluding remarks are made in Section 6.

Notation and terminology. We use $\|\cdot\|$ to denote an arbitrary norm in $\mathbb{R}^n$, which is not necessarily associated with the inner product $\langle\cdot,\cdot\rangle$. We also use $\|\cdot\|_*$ to denote the conjugate norm of $\|\cdot\|$. For any convex function h,

$\partial h(x)$ is the set of subdifferential at x. Given any $X \subseteq \mathbb{R}^n$, we say a convex function $h : X \to \mathbb{R}$ is nonsmooth if $|h(x) - h(y)| \leq M_h\|x - y\|$ for any $x, y \in X$. We say that a convex function $f : X \to \mathbb{R}$ is smooth if it is Lipschitz continuously differentiable with Lipschitz constant $L > 0$, i.e., $\|\nabla f(y) - \nabla f(x)\|_* \leq L\|y - x\|$ for any $x, y \in X$. For any $p \geq 1$, $\|\cdot\|_p$ denotes the standard p-norm in $\mathbb{R}^n$, i.e.,

$$\|x\|_p^p = \sum_{i=1}^{n} |x_i|^p, \qquad \text{for any } x \in \mathbb{R}^n.$$

For any real number r, $\lceil r \rceil$ and $\lfloor r \rfloor$ denote the nearest integer to r from above and below, respectively. $\mathbb{R}_+$ and $\mathbb{R}_{++}$, respectively, denote the set of nonnegative and positive real numbers. $\mathcal{N}$ denotes the set of natural numbers $\{1, 2, \ldots\}$.

2 An optimal primal-dual gradient method

Our goal in this section is to present a novel primal-dual gradient (PDG) method for solving (1.1), which will also provide a basis for the development of the randomized primal-dual gradient methods in later sections. We establish the optimal convergence of this algorithm in terms of the primal-dual optimality gap under the assumption that the gradient of f is computed at each iteration. We show that PDG generalizes one variant of the well-known Nesterov's accelerated gradient method, and allows a natural game interpretation, and hence that the latter algorithm also admits a similar interpretation.

2.1 Preliminaries: primal and dual prox-functions

In this subsection, we discuss both primal and dual prox-functions (proximity control functions) in the primal and dual spaces, respectively.

Recall that the function $\omega : X \to \mathbb{R}$ in (1.1) is strongly convex with modulus 1 with respect to $\|\cdot\|$. We can define a primal *prox-function* associated with ω as

$$P(x^0, x) \equiv P_\omega(x^0, x) := \omega(x) - [\omega(x^0) + \langle \omega'(x^0), x - x^0\rangle], \tag{2.1}$$

where $\omega'(x^0) \in \partial\omega(x^0)$ is an arbitrary subgradient of ω at x^0. Clearly, by the strong convexity of ω, we have

$$P(x^0, x) \geq \tfrac{1}{2}\|x - x^0\|^2, \quad \forall x, x^0 \in X. \tag{2.2}$$

Note that the prox-function $P(\cdot, \cdot)$ described above generalizes the Bregman's distance in the sense that ω is not necessarily differentiable (see [6,2,3,18] and references therein). Throughout this paper, we assume that the prox-mapping associated with X, ω, and h, given by

$$\mathcal{M}_X(g, x^0, \eta) \equiv \mathcal{M}_{X,\omega,h}(g, x^0, \eta) := \arg\min_{x \in X} \left\{ \langle g, x\rangle + h(x) + \mu\,\omega(x) + \eta P(x^0, x) \right\}, \tag{2.3}$$

is easily computable for any $x^0 \in X, g \in \mathbb{R}^n$, $\mu \geq 0$, and $\eta > 0$. Clearly this is equivalent to the assumption that (1.5) is easy to solve. Whenever ω is non-differentiable, we need to specify a particular selection of the subgradient ω' before performing the prox-mapping. We assume throughout this paper that such a selection of ω' is defined recursively as follows. Denote $x^1 \equiv \mathcal{M}_X(g, x^0, \eta)$. By the optimality condition of (2.3), we have

$$g + h'(x^1) + (\mu + \eta)\omega'(x^1) - \eta\omega'(x^0) \in \mathcal{N}_X(x^1),$$

where $\mathcal{N}_X$ denotes the normal cone of X at x^1. Once such a $\omega'(x^1)$ satisfying the above relation is identified, we will use it as a subgradient when defining $P(x^1, x)$ in the next iteration.

Now let us consider the dual space $\mathcal{G}$, where the gradients of f reside, and equip it with the conjugate norm $\|\cdot\|_*$. Let $J_f : \mathcal{G} \to \mathbb{R}$ be the conjugate function of f such that

$$f(x) := \max_{g \in \mathcal{G}} \langle x, g\rangle - J_f(g). \tag{2.4}$$

It is clear that J_f is strongly convex with modulus $1/L_f$ w.r.t. $\|\cdot\|_*$. Therefore, we can define its associated dual prox-functions and dual prox-mappings as

$$D_f(g^0, g) := J_f(g) - [J_f(g^0) + \langle J_f'(g^0), g - g^0\rangle], \tag{2.5}$$

$$\mathcal{M}_{\mathcal{G}}(-\tilde{x}, g^0, \tau) := \arg\min_{g\in\mathcal{G}} \left\{\langle -\tilde{x}, g\rangle + J_f(g) + \tau D_f(g^0, g)\right\}, \tag{2.6}$$

for any $g^0, g \in \mathcal{G}$. Again, D_f may not be uniquely defined since J_f is not necessarily differentiable. Instead of choosing $J_f' \in \partial J_f$ similarly to ω', we can explicitly specify such selections as will be discussed later in this paper.

The following simple result shows that the computation of the dual prox-mapping associated with D_f is equivalent to the computation of ∇f.

Lemma 1 *Let $\tilde{x} \in X$ and $g^0 \in \mathcal{G}$ be given and $D_f(g^0, g)$ be defined in (2.5). For any $\tau > 0$, let us denote $z = [\tilde{x} + \tau J_f'(g^0)]/(1+\tau)$. Then we have $\nabla f(z) = \mathcal{M}_{\mathcal{G}}(-\tilde{x}, g^0, \tau)$.*

Proof. In view of the definition of D_f in (2.5), we have

$$\mathcal{M}_{\mathcal{G}}(-\tilde{x}, g^0, \tau) = \arg\min_{g\in\mathcal{G}} \left\{-\langle \tilde{x} + \tau J_f'(g^0), g\rangle + (1+\tau) J_f(g)\right\} = \arg\max_{g\in\mathcal{G}} \left\{\langle z, g\rangle - J_f(g)\right\} = \nabla f(z).$$

■

2.2 Primal-dual gradient method, Nesterov's method, and a game interpretation

By the definition of J_f in (2.4), problem (1.1) is equivalent to:

$$\Psi^* := \min_{x\in X} \left\{h(x) + \mu\,\omega(x) + \max_{g\in\mathcal{G}} \langle x, g\rangle - J_f(g)\right\}. \tag{2.7}$$

The primal-dual gradient method in Algorithm 1 can be viewed as a game iteratively performed by a primal player (buyer) and a dual player (supplier) for finding the optimal solution (order quantity and product price) of the saddle point problem in (2.7). In this game, both the buyer and supplier have access to their local cost $h(x) + \mu\omega(x)$ and $J_f(g)$, respectively, as well as their interactive cost (or revenue) represented by a bilinear function $\langle x, g\rangle$. Our goal is to design an algorithm such that the buyer and supplier can achieve a equilibrium as soon as possible. In the proposed algorithm, the supplier first applies (2.8) to predict the demand $\tilde{x}^t$ based on historical information, i.e., x^{t-1} and x^{t-2}. She then determines in (2.9) the price g^t in a way to maximize the predicted profit $\langle \tilde{x}^t, g\rangle - J_f(g)$, regularized by the dual prox-function $D_f(g^{t-1}, g)$ with a certain weight $\tau_t \geq 0$. Once after the supplier has made her decision, the buyer then determines his action according to (2.10) in order to minimize the cost $h(x) + \mu\omega(x) + \langle x, g\rangle$, regularized by the primal prox-function $P(x^{t-1}, x)$ with a certain weight $\eta_t \geq 0$.

Algorithm 1 The primal-dual gradient method

Let $x^0 = x^{-1} \in X$, and the nonnegative parameters $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ be given.
Set $g^0 = \nabla f(x^0)$.
for $t = 1, \ldots, k$ **do**
Update (x^t, g^t) according to

$$\tilde{x}^t = \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \tag{2.8}$$

$$g^t = \mathcal{M}_{\mathcal{G}}(-\tilde{x}^t, g^{t-1}, \tau_t). \tag{2.9}$$

$$x^t = \mathcal{M}_X(g^t, x^{t-1}, \eta_t). \tag{2.10}$$

end for

In order to implement the above primal-dual gradient method, it is more convenient to rewrite step (2.9) in a form involving the computation of gradient rather than the dual prox-mapping $\mathcal{M}_{\mathcal{G}}$. In order to do so, we shall specify explicitly the selection of the subgradient J_f' in (2.9). Denoting $\underline{x}^0 = x^0$, we can easily see from $g^0 = \nabla f(x^0)$ that

$\underline{x}^0 \in \partial J_f(g^0)$. Using this relation and letting $J_f'(g^{t-1}) = \underline{x}^{t-1}$ in $D_f(g^{t-1}, g)$ (see (2.5)), we then conclude from Lemma 1 that for any $t \geq 1$, (2.9) reduces to

$$\underline{x}^t = (\tilde{x}^t + \tau_t \underline{x}^{t-1})/(1+\tau_t) \quad \text{and} \quad g^t = \nabla f(\underline{x}^t).$$

With the above selection of the dual prox-function, we can specialize the primal-dual gradient method as follows.

Algorithm 2 A particular implementation of the primal-dual gradient method

Input: Let $x^0 = x^{-1} \in X$, and the nonnegative parameters $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ be given.
Set $\underline{x}^0 = x^0$.
for $t = 1, 2, \ldots, k$ **do**

$$\tilde{x}^t = \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \tag{2.11}$$
$$\underline{x}^t = \left(\tilde{x}^t + \tau_t \underline{x}^{t-1}\right)/(1+\tau_t). \tag{2.12}$$
$$g^t = \nabla f(\underline{x}^t). \tag{2.13}$$
$$x^t = \mathcal{M}_X(g^t, x^{t-1}, \eta_t). \tag{2.14}$$

end for

Observe that one potential problem associated with this scheme is that the search points $\underline{x}^t$ defined in (2.11) and (2.12), respectively, may fall outside X. As a result, we need to assume f to be differentiable over $\mathbb{R}^n$. However, it can be shown that by properly specifying α_t and τ_t, we can guarantee $\underline{x}^t \in X$ and thus relax such restrictions on the differentiability of f (see (2.31) and (2.32) below).

The above PDG method is related to the well-known Nesterov's accelerated gradient (AG) method. Let us focus on a simple variant of the AG method that has been extensively studied in the literature (e.g., [28,35,19,14–16]). Given $(x^{t-1}, \bar{x}^{t-1}) \in X \times X$, this AG algorithm updates $(x^t, \bar{x}^t)$ by

$$\underline{x}^t = (1-\lambda_t)\bar{x}^{t-1} + \lambda_t x^{t-1}, \tag{2.15}$$
$$x^t = \mathcal{M}_X(g^t, x^{t-1}, \eta_t), \tag{2.16}$$
$$\bar{x}^t = (1-\lambda_t)\bar{x}^{t-1} + \lambda_t x^t, \tag{2.17}$$

for some $\lambda_t \in [0,1]$. By (2.15) and (2.17), we have

$$\begin{aligned}
\underline{x}^t &= (1-\lambda_t)[(1-\lambda_{t-1})\bar{x}^{t-2} + \lambda_{t-1}x^{t-1}] + \lambda_t x^{t-1} \\
&= (1-\lambda_t)[\underline{x}^{t-1} - \lambda_{t-1}x^{t-2} + \lambda_{t-1}x^{t-1}] + \lambda_t x^{t-1} \\
&= (1-\lambda_t)\underline{x}^{t-1} + (1-\lambda_t)\lambda_{t-1}(x^{t-1} - x^{t-2}) + \lambda_t x^{t-1}.
\end{aligned}$$

Therefore, (2.15) is equivalent to (2.11) and (2.12) with $\tau_t = (1-\lambda_t)/\lambda_t$ and $\alpha_t = \lambda_{t-1}(1-\lambda_t)/\lambda_t$. Moreover, (2.16) is identical to (2.14)(and (2.10)), and (2.17) basically defines the output of the AG algorithm as an ergodic mean of the iterates x^t. We then conclude that the above variant of Nesterov's AG method is a special case of Algorithm 2 (and Algorithm 1). It should be noted, however, that Algorithm 1 provides more flexibility in the specification of parameters, which will be used later in the development of the RPDG method. Moreover, the presentation of the PDG method helps us to reveal a natural game interpretation out of the intertwined and somehow mysterious updating of the three search sequences in the AG method.

Algorithm 1 is also closely related to Chambolle and Pock's primal-dual method for solving saddle point problems [8], which explains the origin of its name. Two versions of primal-dual methods were discussed in [8]. One is designed for solving general saddle point problems without assuming the strong convexity of J_f and the other one is to deal with the case when J_f is strongly convex by incorporating an additional extrapolation step. As pointed out in Remark 3 of [8], the rate of convergence for the latter primal-dual method is only suboptimal for solving (1.1) as it uses a weaker termination criterion. On the other hand, the PDG method does not involve any additional extrapolation steps and so it shares a similar scheme to the basic version of the primal-dual method in [8]. Moreover, the original primal-dual methods in [8] do not employ general prox-functions, which, as shown in Lemma 1, is crucial to relate the dual step

(2.9) to the computation of the gradients. It should be noted that some recent extensions of the primal-dual method in [10,11,7] indeed consider the incorporation of prox-functions, but restricted to problems without strong convexity. Hence, none of these earlier primal-dual methods can be viewed as a generalized accelerated gradient method.

2.3 Convergence properties of the primal-dual gradient method

Our goal in this subsection is to show that Algorithm 1 exhibits an optimal rate of convergence for solving problem (1.1). It is worth mentioning that our analysis significantly differs from the previous studies on optimal gradient methods and those on primal-dual methods for saddle point problems.

Given a pair of feasible solutions $\bar{z} = (\bar{x}, \bar{g})$ and $z = (x, g)$ of (2.7), we define the primal-dual gap function $Q_f(\bar{z}, z)$ by

$$Q_f(\bar{z}, z) := \left[h(\bar{x}) + \mu\omega(\bar{x}) + \langle \bar{x}, g\rangle - J_f(g)\right] - \left[h(x) + \mu\omega(x) + \langle x, \bar{g}\rangle - J_f(\bar{g})\right]. \tag{2.18}$$

It can be easily seen that $\bar{z}$ (resp., z) is an optimal solution of (2.7) if and only if $Q_f(\bar{z}, z) \le 0$ for any $z \in X \times \mathcal{G}$ (resp., $Q_f(\bar{z}, z) \ge 0$ for any $\bar{z} \in X \times \mathcal{G}$). Therefore, one can assess the solution quality of $\bar{z}$ by the primal-dual optimality gap:

$$\mathrm{gap}(\bar{z}) := \max_{z \in X \times \mathcal{G}} Q_f(\bar{z}, z). \tag{2.19}$$

It should be noted that $\mathrm{gap}(\bar{z})$ may not be well-defined, for example, when X is unbounded and h is not strictly convex. In these cases, we can define a slightly modified primal-dual gap

$$\mathrm{gap}^*(\bar{z}) := \max\left\{Q_f(\bar{z}, z) : x = x^*, g \in \mathcal{G}\right\} \tag{2.20}$$

for an arbitrary optimal solution x^* of (1.1). Since J_f is strongly convex, gap^* is well-defined.

The following result establishes some relationship between the primal optimality gap $\Psi(\bar{z}) - \Psi^*$ and the above primal-dual optimality gaps.

Lemma 2 *Let $\bar{z} = (\bar{x}, \bar{g}) \in X \times \mathcal{G}$ be a given pair of feasible solutions of (2.7) and denote $\bar{g}^* = \nabla f(\bar{x})$. Also let $z^* = (x^*, g^*)$ be a pair of optimal solutions of (2.7). Then we have*

$$\Psi(\bar{x}) - \Psi(x^*) = Q_f((\bar{x}, g^*), (x^*, \bar{g}^*)) \le \mathrm{gap}^*(\bar{z}). \tag{2.21}$$

If in addition, X is bounded, then

$$\mathrm{gap}^*(\bar{z}) \le \mathrm{gap}(\bar{z}). \tag{2.22}$$

Proof. It follows from the definitions of $\bar{g}^*$, gap^* and the gap function Q_f that

$$\begin{aligned}
\Psi(\bar{x}) - \Psi(x^*) &= Q_f((\bar{x}, g^*), (x^*, \bar{g}^*)) \\
&= [h(\bar{x}) + \mu\omega(\bar{x}) + \max_{g \in \mathcal{G}} \langle \bar{x}, g\rangle - J_f(g)] - [h(x^*) + \mu\omega(x^*) + \langle x^*, g^*\rangle - J_f(g^*)] \\
&\le [h(\bar{x}) + \mu\omega(\bar{x}) + \max_{g \in \mathcal{G}} \langle \bar{x}, g\rangle - J_f(g)] - [h(x^*) + \mu\omega(x^*) + \langle x^*, \bar{g}\rangle - J_f(\bar{g})] \\
&= \mathrm{gap}^*(\bar{z}).
\end{aligned}$$

Relation (2.22) follows directly from the definitions of gap^* and gap. ∎

Theorem 1 below describes the main convergence properties of the PDG method. More specifically, we provide in Theorem 1.a) a constant stepsize policy which works for the strongly convex case where $\mu > 0$, and a different parameter setting that works for the non-strongly convex case with $\mu = 0$ in Theorem 1.b). Note that for the strongly convex case, we estimate the solution quality for the iterates $x^t, t = 1, \ldots, k$, as well as that for their ergodic mean

$$\bar{x}^k = (\textstyle\sum_{t=1}^k \theta_t)^{-1} \sum_{t=1}^k (\theta_t x^t) \tag{2.23}$$

for some $\theta_t \ge 0$, while only establishing the error bounds for $\bar{x}^k$ for the non-strongly convex case. We put the proof of Theorem 1 in Section 5 since it shares many basic elements with the convergence analysis of the RPDG method.

Theorem 1 *Let x^* be an optimal solution of (1.1), x^k and $\bar{x}^k$ be defined in (2.10) and (2.23), respectively.*

a) Suppose that $\mu > 0$ and that $\{\tau_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$ and $\{\theta_t\}$ are set to

$$\tau_t = \sqrt{\tfrac{2L_f}{\mu}}, \quad \eta_t = \sqrt{2L_f\mu}, \quad \alpha_t = \alpha \equiv \tfrac{\sqrt{2L_f/\mu}}{1+\sqrt{2L_f/\mu}}, \quad \textit{and} \quad \theta_t = \tfrac{1}{\alpha^t},\ \forall t = 1,\ldots,k. \tag{2.24}$$

Then,

$$P(x^k, x^*) \le \tfrac{\mu+L_f}{\mu}\alpha^k P(x^0, x^*), \tag{2.25}$$

$$\Psi(\bar{x}^k) - \Psi(x^*) \le \mathrm{gap}^*(\bar{z}^k) \le \mu(1-\alpha)^{-1}\left[1 + \tfrac{L_f}{\mu}(2 + \tfrac{L_f}{\mu})\right]\alpha^k P(x^0, x^*), \tag{2.26}$$

$$\Psi(\bar{x}^k) - \Psi(x^*) \le \mathrm{gap}(\bar{z}^k) \le \mu(1-\alpha)^{-1}\left[1 + \tfrac{L_f}{\mu}(2 + \tfrac{L_f}{\mu})\right]\alpha^k \max_{x\in X} P(x^0, x). \tag{2.27}$$

b) Suppose that $\{\tau_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$ and $\{\theta_t\}$ are set to

$$\tau_t = \frac{t-1}{2}, \quad \eta_t = \frac{4L_f}{t}, \quad \alpha_t = \frac{t-1}{t} \quad \textit{and} \quad \theta_t = t,\ \forall t = 1,\ldots,k. \tag{2.28}$$

Then,

$$\Psi(\bar{x}^k) - \Psi(x^*) \le \mathrm{gap}^*(\bar{z}^k) \le \frac{8L_f}{k(k+1)} P(x^0, x^*), \tag{2.29}$$

$$\Psi(\bar{x}^k) - \Psi(x^*) \le \mathrm{gap}(\bar{z}^k) \le \frac{8L_f}{k(k+1)} \max_{x\in X} P(x^0, x). \tag{2.30}$$

Observe that when the algorithmic parameters are set to (2.24), by using an inductive argument, we can easily show that

$$\underline{x}^k = (1-\alpha^2)x^{k-1} + (1-\alpha)\textstyle\sum_{t=1}^{k-2}(\alpha^{k-t}x^t) + \alpha^k x^0. \tag{2.31}$$

In other words, $\underline{x}^k$ can be written as a convex combination of $x^0, \ldots, x^{k-1}$ and hence $\underline{x}^k \in X$ for any $k \ge 1$. Similarly, when the algorithmic parameters are set to (2.28), we can show by using induction that

$$\underline{x}^k = \tfrac{2(2k-1)}{k(k+1)}x^{k-1} + \tfrac{2}{k(k+1)}\textstyle\sum_{t=1}^{k-2}(ix^i), \tag{2.32}$$

which implies $\underline{x}^k \in X$. Therefore, we only need to assume the differentiability of f over X rather than the whole $\mathbb{R}^n$.

In view of the results obtained in Theorem 1, the primal-dual gradient method is an optimal method for convex optimization. In fact, the rates of convergence in (2.26), (2.27), (2.29) and (2.30) associated with the ergodic mean $\bar{z}^k$ have employed the primal-dual optimality gaps $g^*(\bar{z}^k)$ and $g(\bar{z}^k)$, which are stronger than the primal optimality gap $\Psi(\bar{x}^k) - \Psi(x^*)$ used in the previous studies for accelerated gradient methods. Moreover, whenever X is bounded, the primal-dual optimality gap $g(\bar{z}^k)$ gives us a computable online accuracy certificates to check the quality of the solution $\bar{z}^k$ (see [21,14] for some related discussions). Also observe that each iteration of the PDG method requires the computation of ∇f, and hence all the m components ∇f_i. In the next section, we will develop a randomized PDG method that can possibly save the number of gradient evaluations for ∇f_i by utilizing the finite-sum structure of problem (1.1).

3 Randomized primal-dual gradient methods

In this section, we present a randomized primal-dual gradient (RPDG) method which needs to compute the gradient of only one randomly selected component function f_i at each iteration. We show that RPDG can possibly achieve a better complexity than PDG in terms of the total number of gradient evaluations.

3.1 Multi-dual-player reformulation and the RPDG algorithm

We start by introducing a different saddle point reformulation of (1.1) than (2.7). Let $J_i : \mathcal{Y}_i \to \mathbb{R}$ be the conjugate functions of f_i and $\mathcal{Y}_i$, $i = 1, \ldots, m$, denote the dual spaces where the gradients of f_i reside. For the sake of notational convenience, let us denote $J(y) := \sum_{i=1}^m J_i(y_i)$, $\mathcal{Y} := \mathcal{Y}_1 \times \mathcal{Y}_2 \times \ldots \times \mathcal{Y}_m$, and $y = (y_1; y_2; \ldots; y_m)$ for any $y_i \in \mathcal{Y}_i$, $i = 1, \ldots, m$. Clearly, we can reformulate problem (1.1) equivalently as a saddle point problem:

$$\Psi^* := \min_{x \in X} \left\{ h(x) + \mu\, \omega(x) + \max_{y \in \mathcal{Y}} \langle x, Uy \rangle - J(y) \right\}, \tag{3.1}$$

where $U \in \mathbb{R}^{n \times nm}$ is given by

$$U := [I, I, \ldots, I]. \tag{3.2}$$

Here I is the identity matrix in $\mathbb{R}^n$. Given a pair of feasible solutions $\bar{z} = (\bar{x}, \bar{y})$ and $z = (x, y)$ of (3.1), we define the primal-dual gap function $Q(\bar{z}, z)$ by

$$Q(\bar{z}, z) := [h(\bar{x}) + \mu\omega(\bar{x}) + \langle \bar{x}, Uy \rangle - J(y)] - [h(x) + \mu\omega(x) + \langle x, U\bar{y} \rangle - J(\bar{y})]. \tag{3.3}$$

It is well-known that $\bar{z} \in Z \equiv X \times \mathcal{Y}$ is an optimal solution of (3.1) if and only if $Q(\bar{z}, z) \le 0$ for any $z \in Z$.

Since $J_i, i = 1, \ldots, m$, are strongly convex with modulus $\sigma_i = 1/L_i$ w.r.t. $\|\cdot\|_*$, we can define their associated dual prox-functions and dual prox-mappings as

$$D_i(y_i^0, y_i) := J_i(y_i) - [J_i(y_i^0) + \langle J_i'(y_i^0), y_i - y_i^0 \rangle], \tag{3.4}$$

$$\mathcal{M}_{\mathcal{Y}_i}(-\tilde{x}, y_i^0, \tau) := \arg\min_{y_i \in \mathcal{Y}_i} \left\{ \langle -\tilde{x}, y \rangle + J_i(y_i) + \tau D_i(y_i^0, y_i) \right\}, \tag{3.5}$$

for any $y_i^0, y_i \in \mathcal{Y}_i$. Accordingly, we define

$$D(\tilde{y}, y) := \textstyle\sum_{i=1}^m D_i(\tilde{y}_i, y_i). \tag{3.6}$$

Again, D_i may not be uniquely defined since J_i are not necessarily differentiable. However, we will discuss how to specify the particular selection of $J_i' \in \partial J_i$ later in this subsection.

We are now ready to describe the randomized primal-dual method, which is obtained by properly modifying the primal-dual gradient method as follows. Firstly, in (3.8), we only compute a randomly selected dual prox-mapping $\mathcal{M}_{\mathcal{Y}_i}$ rather than the dual prox-mapping $\mathcal{M}_{\mathcal{G}}$ as in Algorithm 1. Secondly, in addition to the primal prediction step (3.7), we add a new dual prediction step (3.9), and then use the predicted dual variable $\tilde{y}^t$ for the computation of the new search point x^t in (3.10). It can be easily seen that the RPDG method reduces to the PDG method whenever this algorithm is directly applied to (2.7) (i.e., $m = 1$, $\mathcal{Y}_1 = \mathcal{G}$, and $J_1 = J_f$) .

Algorithm 3 A randomized primal-dual gradient (RPDG) method

Let $x^0 = x^{-1} \in X$, and the nonnegative parameters $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ be given.
Set $y_i^0 = \nabla f_i(x^0)$, $i = 1, \ldots, m$.
for $t = 1, \ldots, k$ **do**
Choose i_t according to $\mathrm{Prob}\{i_t = i\} = p_i$, $i = 1, \ldots, m$.
Update $z^t = (x^t, y^t)$ according to

$$\tilde{x}^t = \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \tag{3.7}$$

$$y_i^t = \begin{cases} \mathcal{M}_{\mathcal{Y}_i}(-\tilde{x}^t, y_i^{t-1}, \tau_t), & i = i_t, \\ y_i^{t-1}, & i \ne i_t. \end{cases} \tag{3.8}$$

$$\tilde{y}_i^t = \begin{cases} p_i^{-1}(y_i^t - y_i^{t-1}) + y_i^{t-1}, & i = i_t, \\ y_i^{t-1}, & i \ne i_t. \end{cases} \tag{3.9}$$

$$x^t = \mathcal{M}_X(\textstyle\sum_{i=1}^m \tilde{y}_i^t, x^{t-1}, \eta_t). \tag{3.10}$$

end for

Similarly to the PDG method, the RPDG method can be viewed as a game iteratively performed by a buyer and m suppliers for finding the solutions (order quantities and product prices) of the saddle point problem in (3.1). In this game, both the buyer and suppliers have access to their local cost $h(x)+\mu\omega(x)$ and $J_i(y_i)$, respectively, as well as their interactive cost (or revenue) represented by a bilinear function $\langle x, y_i\rangle$. Also, the buyer has to purchase the same amount of products from each supplier (e.g., for fairness). Although there are m suppliers, in each iteration only a randomly chosen supplier can make price changes according to (3.8) using the predicted demand $\tilde{x}^t$. In order to understand the buyer's decision in (3.10), let us first denote

$$\hat{y}_i^t := \mathcal{M}_{\mathcal{Y}_i}(-\tilde{x}^t, y_i^{t-1}, \tau_t), \quad i = 1, \ldots, m;\ t = 1, \ldots, k. \tag{3.11}$$

In other words, $\hat{y}_i^t$, $i = 1, \ldots, m$, denote the prices that all the suppliers can possibly set up at iteration t. Then we can see that

$$\mathbb{E}_t[\tilde{y}_i^t] = \hat{y}_i^t. \tag{3.12}$$

Indeed, we have

$$y_i^t = \begin{cases} \hat{y}_i^t, & i = i_t, \\ y_i^{t-1}, & i \neq i_t. \end{cases} \tag{3.13}$$

Hence $\mathbb{E}_t[y_i^t] = p_i\hat{y}_i^t + (1-p_i)y_i^{t-1}$, $i = 1, \ldots, m$. Using this identity in the definition of $\tilde{y}^t$ in (3.9), we obtain (3.12). Instead of using $\sum_{i=1}^m \hat{y}_i^t$ in determining his order in (3.10), the buyer notices that only one supplier has made a change on the price, and thus uses $\sum_{i=1}^m \tilde{y}_i^t$ to predict the case when all the dual players would modify the prices simultaneously.

In order to implement the above RPDG method, we shall explicitly specify the selection of the subgradient J'_{i_t} in the definition of the dual prox-mapping in (3.8). Denoting $\underline{x}_i^0 = x^0$, $i = 1, \ldots, m$, we can easily see from $y_i^0 = \nabla f_i(x^0)$ that $\underline{x}_i^0 \in \partial f_i^*(y_i^0)$, $i = 1, \ldots, m$. Using this relation and letting $J'_i(y_i^{t-1}) = \underline{x}_i^{t-1}$ in the definition of $D_i(y_i^{t-1}, y_i)$ in (3.8) (see (3.4)), we then conclude from Lemma 1 (with $J_f = J_{i_t}$ and $D_f = D_{i_t}$) and (3.8) that for any $t \geq 1$,

$$\begin{aligned} &\underline{x}_{i_t}^t = (\tilde{x}^t + \tau_t \underline{x}_{i_t}^{t-1})/(1+\tau_t), \quad \underline{x}_i^t = \underline{x}_i^{t-1},\ \forall i \neq i_t; \\ &y_{i_t}^t = \nabla f_{i_t}(\underline{x}_{i_t}^t), \quad y_i^t = y_i^{t-1},\ \forall i \neq i_t. \end{aligned}$$

Moreover, observe that the computation of x^t in (3.10) requires an involved computation of $\sum_{i=1}^m \tilde{y}_i^t$. In order to save computational time, we suggest to compute this quantity in a recursive manner as follows. Let us denote $g^t \equiv \sum_{i=1}^m y_i^t$. Clearly, in view of the fact that $y_i^t = y_i^{t-1}$, $\forall i \neq i_t$, we have

$$g^t = g^{t-1} + (y_{i_t}^t - y_{i_t}^{t-1}).$$

Also, by the definition of g^t and (3.9), we have

$$\begin{aligned} \textstyle\sum_{i=1}^m \tilde{y}_i^t &= \textstyle\sum_{i \neq i_t} y_i^{t-1} + p_{i_t}^{-1}(y_{i_t}^t - y_{i_t}^{t-1}) + y_{i_t}^{t-1} \\ &= \textstyle\sum_{i=1}^m y_i^{t-1} + p_{i_t}^{-1}(y_{i_t}^t - y_{i_t}^{t-1}) \\ &= g^{t-1} + p_{i_t}^{-1}(y_{i_t}^t - y_{i_t}^{t-1}). \end{aligned}$$

Incorporating these two ideas mentioned above, we present an efficient implementation of the RPDG method in Algorithm 4.

Clearly, the RPDG method is an incremental gradient type method since each iteration of this algorithm involves the computation of the gradient ∇f_{i_t} of only one component function. As shown in the following Subsection, such an randomization scheme can lead to significantly savings on the total number of gradient evaluations, at the expense of more primal prox-mappings.

It should also be noted that due to the randomness in the RPDG method, we can not guarantee that $\underline{x}_i^t \in X$ for all $i = 1, \ldots, m$, and $t \geq 1$ in general, even though we do have all the iterates $x^t \in X$. That is why we need to make the assumption that f_i's are differentiable over $\mathbb{R}^n$ for the RPDG method.

Algorithm 4 An efficient implementation of the RPDG method

Let $x^0 = x^{-1} \in X$, and nonnegative parameters $\{\alpha_t\}$, $\{\tau_t\}$, and $\{\eta_t\}$ be given.
Set $\underline{x}_i^0 = x^0$, $y_i^0 = \nabla f_i(x^0)$, $i = 1, \ldots, m$, and $g^0 = \sum_{i=1}^m y_i^0$.
for $t = 1, \ldots, k$ **do**
Choose i_t according to $\text{Prob}\{i_t = i\} = p_i$, $i = 1, \ldots, m$.
Update $z^t := (x^t, y^t)$ by

$$\tilde{x}^t = \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}. \tag{3.14}$$

$$\underline{x}_i^t = \begin{cases} (1+\tau_t)^{-1}\left(\tilde{x}^t + \tau_t \underline{x}_i^{t-1}\right), & i = i_t, \\ \underline{x}_i^{t-1}, & i \neq i_t. \end{cases} \tag{3.15}$$

$$y_i^t = \begin{cases} \nabla f_i(\underline{x}_i^t), & i = i_t, \\ y_i^{t-1}, & i \neq i_t. \end{cases} \tag{3.16}$$

$$x^t = \mathcal{M}_X(g^{t-1} + p_{i_t}^{-1}(y_{i_t}^t - y_{i_t}^{t-1}), x^{t-1}, \eta_t). \tag{3.17}$$

$$g^t = g^{t-1} + y_{i_t}^t - y_{i_t}^{t-1}. \tag{3.18}$$

end for

3.2 The convergence of the RPDG algorithm

Our goal in this subsection is to describe the convergence properties of the RPDG method for the strongly convex case when $\mu > 0$. Generalization of the RPDG method for the non-strongly convex case will be discussed in Section 4.

Theorem 2 below states some general convergence properties of RPDG. Similar to PDG method, we provide bounds on $\mathbb{E}[P(x^k, x^*)]$ and $\mathbb{E}[\Psi(\bar{x}^k) - \Psi(x^*)]$. However, we cannot provide a bound on the expected primal-dual gap $\mathbb{E}[\text{gap}(\bar{x}^k)]$ even though our analysis for the RPDG algorithm still relies on the primal-dual gap function Q in (3.3) (see [11] for some relevant disucssions).

Theorem 2 *Suppose that $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ in the RPDG method are set to*

$$\tau_t = \tau, \quad \eta_t = \eta, \quad \text{and} \quad \alpha_t = \alpha, \tag{3.19}$$

for any $t \geq 1$ such that

$$(1-\alpha)(1+\tau) \leq p_i, i = 1, \ldots, m, \tag{3.20}$$

$$\eta \leq \alpha(\mu + \eta), \tag{3.21}$$

$$\eta\tau p_i \geq 4L_i, i = 1, \ldots, m, \tag{3.22}$$

for some $\alpha \in (0, 1)$. Then, for any $k \geq 1$, we have

$$\mathbb{E}[P(x^k, x^*)] \leq \left(1 + \tfrac{L_f\alpha}{(1-\alpha)\eta}\right)\alpha^k P(x^0, x^*), \tag{3.23}$$

$$\mathbb{E}[\Psi(\bar{x}^k) - \Psi(x^*)] \leq \alpha^{k/2}\left(\alpha^{-1}\eta + \tfrac{3-2\alpha}{1-\alpha}L_f + \tfrac{2L_f^2\alpha}{(1-\alpha)\eta}\right)P(x^0, x^*), \tag{3.24}$$

where $\bar{x}^k = (\sum_{t=1}^k \theta_t)^{-1}\sum_{t=1}^k(\theta_t x^t)$ with $\{\theta_t\}$ defined as in (2.24), and x^* denotes the optimal solution of problem (1.1), and the expectation is taken w.r.t. $i_1, \ldots, i_k$.

We now provide a few specific selections of p_i, τ, η, and α satisfying (3.20)-(3.22) and establish the complexity of the RPDG method for computing a stochastic ϵ-solution of problem (1.1), i.e., a point $\bar{x} \in X$ s.t. $\mathbb{E}[P(\bar{x}, x^*)] \leq \epsilon$, as well as a stochastic (ϵ, λ)-solution of problem (1.1), i.e., a point $\bar{x} \in X$ s.t. $\text{Prob}\{P(\bar{x}, x^*) \leq \epsilon\} \geq 1 - \lambda$ for some $\lambda \in (0, 1)$. Moreover, in view of (3.24), similar complexity bounds of the RPDG method can be established in terms of the primal optimality gap, i.e. $\mathbb{E}[\Psi(\bar{x}) - \Psi^*]$.

The following corollary shows the convergence of RPDG under a non-uniform distribution for the random variables i_t, $t = 1, \ldots, k$.

Corollary 1 *Suppose that $\{i_t\}$ in the RPDG method are distributed over $\{1,\ldots,m\}$ according to*

$$p_i = \text{Prob}\{i_t = i\} = \tfrac{1}{2m} + \tfrac{L_i}{2L}, i = 1,\ldots,m. \tag{3.25}$$

Also assume that $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ are set to (3.19) with

$$\tau = \tfrac{\sqrt{(m-1)^2+4mC}-(m-1)}{2m}, \quad \eta = \tfrac{\mu\sqrt{(m-1)^2+4mC}+\mu(m-1)}{2}, \quad \textit{and} \quad \alpha = 1 - \tfrac{1}{(m+1)+\sqrt{(m-1)^2+4mC}}, \tag{3.26}$$

where

$$C = \tfrac{8L}{\mu}. \tag{3.27}$$

Then for any $k \geq 1$, we have

$$\mathbb{E}[P(x^k, x^*)] \leq (1 + \tfrac{3L_f}{\mu})\alpha^k P(x^0, x^*), \tag{3.28}$$

$$\mathbb{E}[\Psi(\bar{x}^k) - \Psi^*] \leq \alpha^{k/2}(1-\alpha)^{-1}\left[\mu + 2L_f + \tfrac{L_f^2}{\mu}\right] P(x^0, x^*). \tag{3.29}$$

As a consequence, the number of iterations performed by the RPDG method to find a stochastic ϵ-solution and a stochastic (ϵ,λ)-solution of (1.1), in terms of the distance to the optimal solution, i.e., $\mathbb{E}[P(x^k,x^*)]$, can be bounded by $K(\epsilon, C)$ and $K(\lambda\epsilon, C)$, respectively, where

$$K(\epsilon, C) := \left[(m+1) + \sqrt{(m-1)^2 + 4mC}\right] \log\left[(1 + \tfrac{3L_f}{\mu})\tfrac{P(x^0,x^*)}{\epsilon}\right]. \tag{3.30}$$

Similarly, the total number of iterations performed by the RPDG method to find a stochastic ϵ-solution and a stochastic (ϵ,λ)-solution of (1.1), in terms of the primal optimality gap, i.e., $\mathbb{E}[\Psi(\bar{x}^k) - \Psi^*]$, can be bounded by $\tilde{K}(\epsilon, C)$ and $\tilde{K}(\lambda\epsilon, C)$, respectively, where

$$\tilde{K}(\epsilon, C) := 2\left[(m+1) + \sqrt{(m-1)^2 + 4mC}\right] \log\left[2(\mu + 2L_f + \tfrac{L_f^2}{\mu})(m + \sqrt{mC})\tfrac{P(x^0,x^*)}{\epsilon}\right]. \tag{3.31}$$

Proof. It follows from (3.26) that

$$(1-\alpha)(1+\tau) = 1/(2m) \leq p_i, \quad (1-\alpha)\eta = (\alpha - 1/2)\mu \leq \alpha\mu, \quad \text{and} \quad \eta\tau p_i = \mu C p_i \geq 4L_i,$$

and hence that the conditions in (3.20)-(3.22) are satisfied. Notice that by the fact that $\alpha \geq 3/4, \ \forall m \geq 1$ and (3.26), we have

$$1 + \tfrac{L_f\alpha}{(1-\alpha)\eta} = 1 + L_f\tfrac{\alpha}{(\alpha-1/2)\mu} \leq 1 + \tfrac{3L_f}{\mu}.$$

Using the above bound in (3.23), we obtain (3.28). It follows from the facts $(1-\alpha)\eta \leq \alpha\mu$, $1/2 \leq \alpha \leq 1, \forall m \geq 1$, and $\eta \geq \mu\sqrt{C} > 2\mu$ that

$$\alpha^{-1}\eta + \tfrac{3-2\alpha}{1-\alpha}L_f + \tfrac{2L_f^2\alpha}{(1-\alpha)\eta} \leq (1-\alpha)^{-1}(\mu + 2L_f + \tfrac{L_f^2}{\mu}).$$

Using the above bound in (3.24), we obtain (3.29). Denoting $D \equiv (1 + \frac{3L_f}{\mu})P(x^0, x^*)$, we conclude from (3.28) and the fact that $\log x \leq x - 1$ for any $x \in (0,1)$ that

$$\mathbb{E}[P(x^{K(\epsilon,C)}, x^*)] \leq D\alpha^{\frac{\log(D/\epsilon)}{1-\alpha}} \leq D\alpha^{\frac{\log(D/\epsilon)}{-\log\alpha}} \leq D\alpha^{\frac{\log(\epsilon/D)}{\log\alpha}} = \epsilon.$$

Moreover, by Markov's inequality, (3.28) and the fact that $\log x \leq x - 1$ for any $x \in (0,1)$, we have

$$\text{Prob}\{P(x^{K(\lambda\epsilon,C)}, x^*) > \epsilon\} \leq \tfrac{1}{\epsilon}\mathbb{E}[P(x^{K(\lambda\epsilon,C)}, x^*)] \leq \tfrac{D}{\epsilon}\alpha^{\frac{\log(D/(\lambda\epsilon))}{1-\alpha}} \leq \tfrac{D}{\epsilon}\alpha^{\frac{\log(\lambda\epsilon/D)}{\log\alpha}} = \lambda.$$

The proofs for the complexity bounds in terms of the primal optimality gap is similar and hence the details are skipped. ∎

The non-uniform distribution in (3.25) requires the estimation of the Lipschitz constants L_i, $i = 1,\ldots,m$. In case such information is not available, we can use a uniform distribution for i_t, and as a result, the complexity bounds will depend on a larger condition number given by $m\max_{i=1,\ldots,m} L_i/\mu$. However, if we do have $L_1 = L_2 = \cdots = L_m$, then the results obtained by using a uniform distribution is slightly sharper than the one by using a non-uniform distribution in Corollary 1.

Corollary 2 *Suppose that $\{i_t\}$ in the RPDG method are uniformly distributed over $\{1,\ldots,m\}$ according to*

$$p_i = \text{Prob}\{i_t = i\} = \tfrac{1}{m}, i = 1,\ldots,m. \tag{3.32}$$

Also assume that $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ are set to (3.19) with

$$\tau = \tfrac{\sqrt{(m-1)^2+4m\bar{C}}-(m-1)}{2m}, \quad \eta = \tfrac{\mu\sqrt{(m-1)^2+4m\bar{C}}+\mu(m-1)}{2}, \quad \text{and} \quad \alpha = 1 - \tfrac{2}{(m+1)+\sqrt{(m-1)^2+4m\bar{C}}}, \tag{3.33}$$

where

$$\bar{C} := \tfrac{4m}{\mu} \max_{i=1,\ldots,m} L_i. \tag{3.34}$$

Then we have

$$\mathbb{E}[P(x^k, x^*)] \le (1 + \tfrac{L_f}{\mu})\alpha^k P(x^0, x^*), \tag{3.35}$$

$$\mathbb{E}[\Psi(\bar{x}^k) - \Psi^*] \le \alpha^{k/2}(1-\alpha)^{-1}\left(\mu + 2L_f + \tfrac{L_f^2}{\mu}\right) P(x^0, x^*). \tag{3.36}$$

for any $k \ge 1$. As a consequence, the number of iterations performed by the RPDG method to find a stochastic ϵ-solution and a stochastic (ϵ,λ)-solution of (1.1), in terms of the distance to the optimal solution, i.e., $\mathbb{E}[P(x^k,x^*)]$, can be bounded by $K_u(\epsilon,\bar{C})$ and $K_u(\lambda\epsilon,\bar{C})$, respectively, where

$$K_u(\epsilon,\bar{C}) := \tfrac{(m+1)+\sqrt{(m-1)^2+4m\bar{C}}}{2} \log\left[(1+\tfrac{L_f}{\mu})\tfrac{P(x^0,x^*)}{\epsilon}\right].$$

Similarly, the total number of iterations performed by the RPDG method to find a stochastic ϵ-solution and a stochastic (ϵ,λ)-solution of (1.1), in terms of the primal optimality gap, i.e., $\mathbb{E}[\Psi(\bar{x}^k) - \Psi^*]$, can be bounded by $\tilde{K}(\epsilon,\bar{C})/2$ and $\tilde{K}(\lambda\epsilon,\bar{C})/2$, respectively, where $\tilde{K}(\epsilon,\bar{C})$ is defined in (3.31).

Proof. It follows from (3.33) that

$$(1-\alpha)(1+\tau) = 1/m = p_i, \quad (1-\alpha)\eta - \alpha\mu = 0, \quad \text{and} \quad \eta\tau = \mu\bar{C} \ge 4mL_i,$$

and hence that the conditions in (3.20)-(3.22) are satisfied. By the identity $(1-\alpha)\eta = \alpha\mu$, we have

$$1 + \tfrac{L_f\alpha}{(1-\alpha)\eta} = 1 + \tfrac{L_f}{\mu}.$$

Using the above bound in (3.23), we obtain (3.35). Moreover, note that $\eta \ge \mu\sqrt{\bar{C}} \ge 2\mu$ and $2/3 \le \alpha \le 1, \forall m \ge 1$ we have

$$\alpha^{-1}\eta + \tfrac{3-2\alpha}{1-\alpha}L_f + \tfrac{2L_f^2\alpha}{(1-\alpha)\eta} \le (1-\alpha)^{-1}(\mu + 2L_f + \tfrac{L_f^2}{\mu}).$$

Using the above bound in (3.24), we obtain (3.36). The proofs for the complexity bounds are similar to those in Corollary 1 and hence the details are skipped. ∎

Comparing the complexity bounds obtained from Corollaries 1 and 2 with those of any optimal deterministic first-order method, they differ in a factor of $\mathcal{O}(\sqrt{mL_f/L})$, whenever $\sqrt{mC}\log(1/\epsilon)$ is dominating in (3.30). Clearly, when L_f and L are in the same order of magnitude, RPDG can save up to $\mathcal{O}(\sqrt{m})$ gradient evaluations for the component function f_i than the deterministic first-order methods. However, it should be pointed out that L_f can be much smaller than L. In particular, when $L_f = L_i, i = 1,\ldots,m$, $L_f = L/m$. In the next subsection, we will construct examples in such extreme cases to obtain the lower complexity bound for general randomized incremental gradient methods.

3.3 Lower complexity bound for randomized methods

Our goal in this subsection is to demonstrate that the complexity bounds obtained in Theorem 2, and Corollaries 1 and 2 for the RPDG method are essentially not improvable. Observe that although there exist rich lower complexity bounds in the literature for deterministic first-order methods (e.g. [26,28]), the study on lower complexity bounds for randomized methods are still quite limited. Recently Agarwal and Bottou [1] suggested a lower complexity bound for minimizing the finite-sum convex optimization problem given in the form of (1.1). However, their bounds are developed for deterministic algorithms and hence not applicable to randomized incremental gradient methods.

To derive the performance limit of the incremental gradient methods, we consider a special class of unconstrained and separable strongly convex optimization problems given in the form of

$$\min_{x_i\in\mathbb{R}^{\tilde{n}},i=1,\ldots,m} \left\{\Psi(x) := \textstyle\sum_{i=1}^m \left[f_i(x_i) + \frac{\mu}{2}\|x_i\|_2^2\right]\right\}. \tag{3.37}$$

Here $\tilde{n} \equiv n/m \in \{1,2,\ldots\}$ and $\|\cdot\|_2$ denotes standard Euclidean norm. To fix the notation, we also denote $x = (x_1,\ldots,x_m)$. Moreover, we assume that f_i's are quadratic functions given by

$$f_i(x_i) = \tfrac{\mu(\mathcal{Q}-1)}{4}\left[\tfrac{1}{2}\langle Ax_i, x_i\rangle - \langle e_1, x_i\rangle\right], \tag{3.38}$$

where $e_1 := (1,0,\ldots,0)$ and A is a symmetric matrix in $\mathbb{R}^{\tilde{n}\times\tilde{n}}$ given by

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & \kappa \end{pmatrix} \quad \text{with} \quad \kappa = \tfrac{\sqrt{\mathcal{Q}}+3}{\sqrt{\mathcal{Q}}+1}. \tag{3.39}$$

Compared with the classic worst-case example given in [28], the tridiagonal matrix A above consists of a different diagonal element κ (instead of 2). This modification allows us to study problems of finite dimension more conveniently. It can be easily checked that $A \succeq 0$ and its maximum eigenvalue does not exceeds 4. Indeed, for any $s \equiv (s_1,\ldots,s_{\tilde{n}}) \in \mathbb{R}^{\tilde{n}}$, we have

$$\begin{aligned} \langle As, s\rangle &= s_1^2 + \textstyle\sum_{i=1}^{\tilde{n}-1}(s_i - s_{i+1})^2 + (\kappa-1)s_{\tilde{n}}^2 \geq 0 \\ \langle As, s\rangle &\leq s_1^2 + \textstyle\sum_{i=1}^{\tilde{n}-1} 2(s_i^2 + s_{i+1}^2) + (\kappa-1)s_{\tilde{n}}^2 \\ &= 3s_1^2 + 4\textstyle\sum_{i=2}^{\tilde{n}-1} s_i^2 + (\kappa+1)s_{\tilde{n}}^2 \leq 4\|s\|_2^2, \end{aligned}$$

where the last inequality follows from the fact that $\kappa \leq 3$. Therefore, for any $\mathcal{Q} > 1$, the component functions f_i in (3.38) are convex and their gradients are Lipschitz continuous with constant bounded by $L_i = \mu(\mathcal{Q}-1)$, $i = 1,\ldots,m$.

We consider a general class of randomized incremental gradient methods which sequentially acquire the gradients of a randomly selected component function f_{i_t} at iteration t. More specifically, we assume that the independent random variables i_t, $t = 1,2,\ldots$, satisfy

$$\text{Prob}\{i_t = i\} = p_i \quad \text{and} \quad \textstyle\sum_{i=1}^m p_i = 1, \quad p_i \geq 0, i = 1,\ldots,m. \tag{3.40}$$

Similar to [28], we assume that these methods generate a sequence of test points $\{x^k\}$ such that

$$x^k \in x^0 + \text{Lin}\{\nabla f_{i_1}(x^0),\ldots,\nabla f_{i_k}(x^{k-1})\}, \tag{3.41}$$

where Lin denotes the linear span.

Theorem 3 below describes the performance limit of the above randomized incremental gradient methods for solving (3.37).

Theorem 3 *Let x^* be the optimal solution of problem (3.37) and denote*

$$q := \frac{\sqrt{\mathcal{Q}}-1}{\sqrt{\mathcal{Q}}+1}. \tag{3.42}$$

Then the iterates $\{x^k\}$ generated by any randomized incremental gradient method must satisfy

$$\frac{\mathbb{E}[\|x^k-x^*\|_2^2]}{\|x^0-x^*\|_2^2} \geq \frac{1}{2}\exp\left(-\frac{4k\sqrt{\mathcal{Q}}}{m(\sqrt{\mathcal{Q}}+1)^2-4\sqrt{\mathcal{Q}}}\right) \tag{3.43}$$

for any

$$n \geq \underline{n}(m,k) \equiv \frac{m\log\left[\left(1-(1-q^2)/m\right)^k/2\right]}{2\log q}. \tag{3.44}$$

As an immediate consequence of Theorem 3, we obtain a lower complexity bound for randomized incremental gradient methods.

Corollary 3 *The number of gradient evaluations performed by any randomized incremental gradient methods for finding a solution $\bar{x} \in X$ of problem (1.1) such that $\mathbb{E}[\|\bar{x}-x^*\|_2^2] \leq \epsilon$ cannot be smaller than*

$$\Omega\left\{\left(\sqrt{m\mathcal{C}}+m\right)\log\frac{\|x^0-x^*\|_2^2}{\epsilon}\right\}$$

if n is sufficiently large, where $\mathcal{C} = L/\mu$ and $L = \sum_{i=1}^m L_i$.

Proof. It follows from (3.43) that the number of iterations k required by any randomized incremental gradient methods to find an approximate solution $\bar{x}$ must satisfy

$$k \geq \left(\frac{m(\sqrt{\mathcal{Q}}+1)^2}{4\sqrt{\mathcal{Q}}}-1\right)\log\frac{\|x^0-x^*\|_2^2}{2\epsilon} \geq \left[\frac{m}{2}\left(\frac{\sqrt{\mathcal{Q}}}{2}+1\right)-1\right]\log\frac{\|x^0-x^*\|_2^2}{2\epsilon}. \tag{3.45}$$

Noting that for the worst-case instance in (3.37), we have $L_i = \mu(\mathcal{Q}-1)$, $i = 1,\ldots,m$, and hence that $L = \sum_{i=1}^m L_i = m\mu(Q-1)$. Using this relation, we conclude that

$$k \geq \left[\frac{1}{2}\left(\frac{\sqrt{m\mathcal{C}+m^2}}{2}+m\right)-1\right]\log\frac{\|x^0-x^*\|_2^2}{2\epsilon} =: \underline{k}.$$

The above bound holds when $n \geq \underline{n}(m,\underline{k})$. ∎

In view of Theorem 3, we can also derive a lower complexity bound for randomized block coordinate descent methods, which update one randomly selected block of variables at each iteration for $\min_{x\in X}\Psi(x)$. Here Ψ is smooth and strongly convex such that

$$\frac{\mu_\Psi}{2}\|x-y\|_2^2 \leq \Psi(x)-\Psi(y)-\langle\nabla\Psi(y),x-y\rangle \leq \frac{L_\Psi}{2}\|x-y\|_2^2, \forall x,y\in X.$$

Corollary 4 *The number of iterations performed by any randomized block coordinate descent methods for finding a solution $\bar{x} \in X$ of $\min_{x\in X}\Psi(x)$ such that $\mathbb{E}[\|\bar{x}-x^*\|_2^2] \leq \epsilon$ cannot be smaller than*

$$\Omega\left\{\left(m\sqrt{\mathcal{Q}_\Psi}\right)\log\frac{\|x^0-x^*\|_2^2}{\epsilon}\right\}$$

if n is sufficiently large, where $\mathcal{Q}_\Psi = L_\Psi/\mu_\Psi$ denotes the condition number of Ψ.

Proof. The worst-case instances in (3.37) have a block separable structure. Therefore, any randomized incremental gradient methods are equivalent to randomized block coordinate descent methods. The result then immediately follows from (3.45). ∎

4 Generalization of randomized primal-dual gradient methods

In this section, we generalize the RPDG method for solving a few different types of convex optimization problems which are not necessarily smooth and strongly convex.

4.1 Smooth problems with bounded feasible sets

Our goal in this subsection is to generalize RPDG for solving smooth problems without strong convexity (i.e., $\mu = 0$). Different from the deterministic PDG method, it is difficult to develop a simple stepsize policy for $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ which can guarantee the convergence of this method unless a weaker termination criterion is used (see [11]). In order to obtain stronger convergence results, we will discuss a different approach obtained by applying the RPDG method to a slightly perturbed problem of (1.1).

In order to apply this perturbation approach, we will assume that X is bounded (see Subsection 4.3 for possible extensions), i.e., given $x_0 \in X$, $\exists \Omega_X \geq 0$ s.t.

$$\max_{x \in X} P_\omega(x_0, x) \leq \Omega_X^2. \tag{4.1}$$

Now we define the perturbation problem as

$$\Psi_\delta^* := \min_{x \in X} \{\Psi_\delta(x) := f(x) + h(x) + \delta P_\omega(x_0, x)\}, \tag{4.2}$$

for some fixed $\delta > 0$. It is well-known that an approximate solution of (4.2) will also be an approximate solution of (1.1) if δ is sufficiently small. More specifically, it is easy to verify that

$$\Psi^* \leq \Psi_\delta^* \leq \Psi^* + \delta\Omega_X^2, \tag{4.3}$$

$$\Psi(x) \leq \Psi_\delta(x) \leq \Psi(x) + \delta\Omega_X^2, \quad \forall x \in X. \tag{4.4}$$

The following result describes the complexity associated with this perturbation approach for solving smooth problems without strong convexity (i.e., $\mu = 0$).

Proposition 1 *Let us apply the RPDG method with the parameter settings in Corollary 1 to the perturbation problem (4.2) with*

$$\delta = \tfrac{\epsilon}{2\Omega_X^2}, \tag{4.5}$$

for some $\epsilon > 0$. Then we can find a solution $\bar{x} \in X$ s.t. $\mathbb{E}[\Psi(\bar{x}) - \Psi^*] \leq \epsilon$ in at most

$$\mathcal{O}\left\{\left(m + \sqrt{\tfrac{mL\Omega_X^2}{\epsilon}}\right)\log\tfrac{mL_f\Omega_X}{\epsilon}\right\} \tag{4.6}$$

iterations. Moreover, we can find a solution $\bar{x} \in X$ s.t. $\mathrm{Prob}\{\Psi(\bar{x}) - \Psi^* > \epsilon\} \leq \lambda$ for any $\lambda \in (0,1)$ in at most

$$\mathcal{O}\left\{\left(m + \sqrt{\tfrac{mL\Omega_X^2}{\epsilon}}\right)\log\tfrac{mL_f\Omega_X}{\lambda\epsilon}\right\} \tag{4.7}$$

iterations.

Proof. Let x_δ^* be the optimal solution of (4.2). Denote $C := 16L\Omega_X^2/\epsilon$ and

$$K := 2\left[(m+1) + \sqrt{(m-1)^2 + 4mC}\right]\log\left[(m + \sqrt{mC})(\delta + 2L_f + \tfrac{L_f^2}{\delta})\tfrac{4\Omega_X^2}{\epsilon}\right].$$

It can be easily seen that

$$\Psi(\bar{x}^K) - \Psi^* \leq \Psi_\delta(\bar{x}^K) - \Psi_\delta^* + \delta\Omega_X^2 = \Psi_\delta(\bar{x}^K) - \Psi_\delta^* + \tfrac{\epsilon}{2}.$$

Note that problem (4.2) is given in the form of (1.1) with the strongly convex modulus $\mu = \delta$, and $h(x) = h(x) - \delta\langle\omega'(x_0), x\rangle$. Hence by applying Corollary 1, we have

$$\mathbb{E}[\Psi_\delta(\bar{x}^K) - \Psi_\delta^*] \leq \tfrac{\epsilon}{2}.$$

Combining these two inequalities, we have $\mathbb{E}[\Psi(\bar{x}^K) - \Psi^*] \le \epsilon$, which implies the bound in (4.6). The bound in (4.7) can be shown similarly and hence the details are skipped. ∎

Observe that if we apply a deterministic optimal first-order method (e.g., Nesterov's method or the PDG method), the total number of gradient evaluations for ∇f_i, $i = 1, \ldots, m$, would be given by

$$m\sqrt{\tfrac{L_f \Omega_X^2}{\epsilon}}.$$

Comparing this bound with (4.6), we can see that the number of gradient evaluations performed by the RPDG method can be $\mathcal{O}\left(\sqrt{m}\log^{-1}(mL_f\Omega_X/\epsilon)\right)$ times smaller than these deterministic methods when L and L_f are in the same order of magnitude.

4.2 Structured nonsmooth problems

In this subsection, we assume that the smooth components f_i are nonsmooth but can be approximated closely by smooth ones. More specifically, we assume that

$$f_i(x) := \max_{y_i \in Y_i} \langle A_i x, y_i\rangle - q_i(y_i). \tag{4.8}$$

Nesterov in an important work [29] shows that we can approximate $f_i(x)$ and f, respectively, by

$$\tilde{f}_i(x,\delta) := \max_{y_i \in Y_i} \langle A_i x, y_i\rangle - q_i(y_i) - \delta v_i(y_i) \quad \text{and} \quad \tilde{f}(x,\delta) = \textstyle\sum_{i=1}^m \tilde{f}_i(x,\delta), \tag{4.9}$$

where $v_i(y_i)$ is a strongly convex function with modulus 1 such that

$$0 \le v_i(y_i) \le \Omega_{Y_i}^2, \quad \forall y_i \in Y_i. \tag{4.10}$$

In particular, we can easily show that

$$\tilde{f}_i(x,\delta) \le f_i(x) \le \tilde{f}_i(x,\delta) + \delta\Omega_{Y_i}^2 \quad \text{and} \quad \tilde{f}(x,\delta) \le f(x) \le \tilde{f}(x,\delta) + \delta\Omega_Y^2, \tag{4.11}$$

for any $x \in X$, where $\Omega_Y^2 = \sum_{i=1}^m \Omega_{Y_i}^2$. Moreover, $f_i(\cdot,\delta)$ and $f(\cdot,\delta)$ are continuously differentiable and their gradients are Lipschitz continuous with constants given by

$$\tilde{L}_i = \frac{\|A_i\|^2}{\delta} \quad \text{and} \quad \tilde{L} = \frac{\sum_{i=1}^m \|A_i\|^2}{\delta} = \frac{\|A\|^2}{\delta}, \tag{4.12}$$

respectively. As a consequence, we can apply the RPDG method to solve the approximation problem

$$\tilde{\Psi}_\delta^* := \min_{x\in X} \left\{\tilde{\Psi}_\delta(x) := \tilde{f}(x,\delta) + h(x) + \mu\omega(x)\right\}. \tag{4.13}$$

The following result provides complexity bounds of the RPDG method for solving the above structured nonsmooth problems for the case when $\mu > 0$.

Proposition 2 *Let us apply the RPDG method with the parameter settings in Corollary 1 to the approximation problem (4.13) with*

$$\delta = \tfrac{\epsilon}{2\Omega_Y^2}, \tag{4.14}$$

for some $\epsilon > 0$. Then we can find a solution $\bar{x} \in X$ s.t. $\mathbb{E}[\Psi(\bar{x}) - \Psi^*] \le \epsilon$ in at most

$$\mathcal{O}\left\{\|A\|\Omega_Y\sqrt{\tfrac{m}{\mu\epsilon}}\log\tfrac{m\|A\|\Omega_X\Omega_Y}{\mu\epsilon}\right\} \tag{4.15}$$

iterations. Moreover, we can find a solution $\bar{x} \in X$ s.t. $\mathrm{Prob}\{\Psi(\bar{x}) - \Psi^* > \epsilon\} \le \lambda$ for any $\lambda \in (0,1)$ in at most

$$\mathcal{O}\left\{\|A\|\Omega_Y\sqrt{\tfrac{m}{\mu\epsilon}}\log\tfrac{m\|A\|\Omega_X\Omega_Y}{\lambda\mu\epsilon}\right\} \tag{4.16}$$

iterations.

Proof. It follows from (4.11) and (4.13) that

$$\Psi(\bar{x}^k) - \Psi^* \le \tilde{\Psi}_\delta(\bar{x}^k) - \tilde{\Psi}_\delta^* + \delta\Omega_Y^2 = \tilde{\Psi}_\delta(\bar{x}^k) - \tilde{\Psi}_\delta^* + \tfrac{\epsilon}{2}. \tag{4.17}$$

Using relation (4.12) and Corollaries 1, we conclude that a solution $\bar{x}^k \in X$ satisfying $\mathbb{E}[\tilde{\Psi}_\delta(\bar{x}^k) - \tilde{\Psi}_\delta^*] \le \epsilon/2$ can be found in

$$\mathcal{O}\left\{\|A\|\Omega_Y\sqrt{\tfrac{m}{\mu\epsilon}}\log\left[(m+\sqrt{\tfrac{m\tilde{L}}{\mu}})\left(\mu+2\tilde{L}+\tfrac{\tilde{L}^2}{\mu}\right)\tfrac{\Omega_X^2}{\epsilon}\right]\right\}$$

iterations. This observation together with (4.17) and the definition of $\tilde{L}$ in (4.12) then imply the bound in (4.15). The bound in (4.16) follows similarly from (4.17) and Corollaries 1, and hence the details are skipped. ∎

The following result holds for the RPDG method applied to the above structured nonsmooth problems when $\mu = 0$.

Proposition 3 *Let us apply the RPDG method with the parameter settings in Corollary 1 to the approximation problem (4.13) with δ in (4.14) for some $\epsilon > 0$. Then we can find a solution $\bar{x} \in X$ s.t. $\mathbb{E}[\Psi(\bar{x}) - \Psi^*] \le \epsilon$ in at most*

$$\mathcal{O}\left\{\tfrac{\sqrt{m}\|A\|\Omega_X\Omega_Y}{\epsilon}\log\tfrac{m\|A\|\Omega_X\Omega_Y}{\epsilon}\right\}$$

iterations. Moreover, we can find a solution $\bar{x} \in X$ s.t. $\mathrm{Prob}\{\Psi(\bar{x}) - \Psi^* > \epsilon\} \le \lambda$ for any $\lambda \in (0,1)$ in at most

$$\mathcal{O}\left\{\tfrac{\sqrt{m}\|A\|\Omega_X\Omega_Y}{\epsilon}\log\tfrac{m\|A\|\Omega_X\Omega_Y}{\lambda\epsilon}\right\}$$

iterations.

Proof. Similarly to the arguments used in the proof of Proposition 2, our results follow from (4.17), and an application of Proposition 1 to problem (4.13). ∎

By Propositions 2 and 3, the total number of gradient computations for $\tilde{f}(\cdot,\delta)$ performed by the RPDG method, after disregarding the logarithmic factors, can be $\mathcal{O}(\sqrt{m})$ times smaller than those required by deterministic first-order methods, such as Nesterov's smoothing technique [29].

4.3 Unconstrained smooth problems

In this subsection, we set $X = \mathbb{R}^n$, $h(x) = 0$, and $\mu = 0$ in (1.1) and consider the basic convex programming problem of

$$f^* := \min_{x\in\mathbb{R}^n}\left\{f(x) := \textstyle\sum_{i=1}^m f_i(x)\right\}. \tag{4.18}$$

We assume that the set of optimal solutions X^* of this problem is nonempty.

We will still use the perturbation-based approach as described in Subsection 4.1 by solving the perturbation problem given by

$$f_\delta^* := \min_{x\in\mathbb{R}^n}\left\{f_\delta(x) := f(x) + \tfrac{\delta}{2}\|x - x^0\|_2^2,\right\} \tag{4.19}$$

for some $x^0 \in X, \delta > 0$, where $\|\cdot\|_2$ denotes the Euclidean norm. Also let L_δ denote the Lipschitz constant for $f_\delta(x)$. Clearly, $L_\delta = L + \delta$. Since the problem is unconstrained and the information on the size of the optimal solution is unavailable, it is hard to estimate the total number of iterations by using the absolute accuracy in terms of $\mathbb{E}[f(\bar{x}) - f^*]$. Instead, we define the relative accuracy associated with a given $\bar{x} \in X$ by

$$R_{ac}(\bar{x}, x^0, f^*) := \tfrac{2[f(\bar{x})-f^*]}{L(1+\min_{u\in X^*}\|x^0-u\|_2^2)}. \tag{4.20}$$

We are now ready to establish the complexity of the RPDG method applied to (4.18) in terms of $R_{ac}(\bar{x}, x^0, f^*)$.

Proposition 4 *Let us apply the RPDG method with the parameter settings in Corollary 1 to the perturbation problem (4.19) with*

$$\delta = \tfrac{L\epsilon}{2}, \tag{4.21}$$

for some $\epsilon > 0$. Then we can find a solution $\bar{x} \in X$ s.t. $\mathbb{E}[R_{ac}(\bar{x}, x^0, f^*)] \le \epsilon$ in at most

$$\mathcal{O}\left\{\sqrt{\tfrac{m}{\epsilon}} \log \tfrac{m}{\epsilon}\right\} \tag{4.22}$$

iterations. Moreover, we can find a solution $\bar{x} \in X$ s.t. $\mathrm{Prob}\{R_{ac}(\bar{x}, x^0, f^*) > \epsilon\} \le \lambda$ for any $\lambda \in (0,1)$ in at most

$$\mathcal{O}\left\{\sqrt{\tfrac{m}{\epsilon}} \log \tfrac{m}{\lambda\epsilon}\right\} \tag{4.23}$$

iterations.

Proof. Let x_δ^* be the optimal solution of (4.19). Also let x^* be the optimal solution of (4.18) that is closest to x^0, i.e., $x^* = \mathrm{argmin}_{u \in X^*} \|x^0 - u\|_2$. It then follows from the strong convexity of f_δ that

$$\begin{aligned}
\tfrac{\delta}{2}\|x_\delta^* - x^*\|_2^2 &\le f_\delta(x^*) - f_\delta(x_\delta^*) \\
&= f(x^*) + \tfrac{\delta}{2}\|x^* - x^0\|_2^2 - f_\delta(x_\delta^*) \\
&\le \tfrac{\delta}{2}\|x^* - x^0\|_2^2,
\end{aligned}$$

which implies that

$$\|x_\delta^* - x^*\|_2 \le \|x^* - x^0\|_2. \tag{4.24}$$

Moreover, using the definition of f_δ and the fact that x^* is feasible to (4.19), we have

$$f^* \le f_\delta^* \le f^* + \tfrac{\delta}{2}\|x^* - x^0\|_2^2,$$

which implies that

$$\begin{aligned}
f(\bar{x}^K) - f^* &\le f_\delta(\bar{x}^K) - f_\delta^* + f_\delta^* - f^* \\
&\le f_\delta(\bar{x}^K) - f_\delta^* + \tfrac{\delta}{2}\|x^* - x^0\|_2^2.
\end{aligned}$$

Now suppose that we run the RPDG method applied to (4.19) for K iterations. Then by Corollary 1, we have

$$\begin{aligned}
\mathbb{E}[f_\delta(\bar{x}^K) - f_\delta^*] &\le \alpha^{K/2}(1-\alpha)^{-1}\left(\delta + 2L_\delta + \tfrac{L_\delta^2}{\delta}\right)\|x^0 - x_\delta^*\|_2^2 \\
&\le \alpha^{K/2}(1-\alpha)^{-1}\left(\delta + 2L_\delta + \tfrac{L_\delta^2}{\delta}\right)[\|x^0 - x^*\|_2^2 + \|x^* - x_\delta^*\|_2^2] \\
&= 2\alpha^{K/2}(1-\alpha)^{-1}\left(3\delta + 2L + \tfrac{(L+\delta)^2}{\delta}\right)\|x^0 - x^*\|_2^2,
\end{aligned}$$

where the last inequality follows from (4.24) and α is defined in (3.26) with $C = 8L_\delta/\delta = \frac{8(L+\delta)}{\delta} = 8(2/\epsilon + 1)$. Combining the above two relations, we have

$$\mathbb{E}[f(\bar{x}^K) - f^*] \le \left[2\alpha^{K/2}(1-\alpha)^{-1}\left(3\delta + 2L + \tfrac{(L+\delta)^2}{\delta}\right) + \tfrac{\delta}{2}\right][\|x^0 - x^*\|_2^2.$$

Dividing both sides of the above inequality by $L(1 + \|x^0 - x^*\|_2^2)/2$, we obtain

$$\begin{aligned}
\mathbb{E}[R_{ac}(\bar{x}^K, x^0, f^*)] &\le \tfrac{2}{L}\left[2\alpha^{K/2}(1-\alpha)^{-1}\left(3\delta + 2L + \tfrac{(L+\delta)^2}{\delta}\right) + \tfrac{\delta}{2}\right] \\
&\le 4\left(m + 2\sqrt{2m(\tfrac{2}{\epsilon} + 1)}\right)\left(3\epsilon + 4 + (2+\epsilon)(\tfrac{2}{\epsilon} + 1)\right)\alpha^{K/2} + \tfrac{\epsilon}{2},
\end{aligned}$$

which clearly implies the bound in (4.22). The bound in (4.23) also follows from the above inequality and the Markov's inequality. ∎

By Proposition 4, the total number of gradient evaluations for the component functions f_i required by the RPDG method can be $\mathcal{O}(\sqrt{m}\log^{-1}(m/\epsilon))$ times smaller than those performed by deterministic optimal first-order methods.

5 Complexity analysis

Our main goal in this section is to prove the main theorems in Sections 2 and 3. After introducing some basic tools and general results about PDG and RPDG methods in Subsection 5.1 and 5.2, respectively, we provide the proofs for Theorem 1 and Theorem 2, which describe the main convergence properties for the PDG and RPDG methods, in Subsection 5.3. Moreover, in Subsection 5.4, we provide the proof for the lower complexity bound in Theorem 3.

5.1 Some basic tools

The following result provides a few different bounds on the diameter of the dual feasible sets $\mathcal{G}$ and $\mathcal{Y}$ in (2.7) and (3.1).

Lemma 3 *Let $x^0 \in X$ be given, $y_i^0 = \nabla f_i(x^0)$, $i = 1, \ldots, m$, and $g^0 = \nabla f(x^0)$. Assume that $J_i'(y^0) = x^0$ and $J_f'(g^0) = x^0$ in the definition of $D(y^0, y)$ and $D_f(g^0, g)$ in (3.4) and (2.5), respectively.*

a) For any $x \in X$ and $y_i = \nabla f_i(x)$, $i = 1, \ldots, m$, we have

$$D(y^0, y) \le \frac{L_f}{2}\|x^0 - x\|^2 \le L_f P(x^0, x). \tag{5.1}$$

b) If $x^* \in X$ is an optimal solution of (1.1) and $y_i^* = \nabla f_i(x^*)$, $i = 1, \ldots, m$, then

$$D(y^0, y^*) \le \Psi(x^0) - \Psi(x^*). \tag{5.2}$$

c) For any $x \in X$ and $g = \nabla f(x)$, we have

$$D_f(g^0, g) \le \frac{L_f}{2}\|x^0 - x\|^2. \tag{5.3}$$

Proof. We first show part a). It follows from the definition of J_i, (3.4), and (3.6) that

$$\begin{aligned} D(y^0, y) &= J(y) - J(y^0) - \textstyle\sum_{i=1}^m \langle J_i'(y^0), y_i - y_i^0\rangle \\ &= \langle x, Uy\rangle - f(x) + f(x^0) - \langle x^0, Uy^0\rangle - \langle x^0, U(y - y^0)\rangle \\ &= f(x^0) - f(x) - \langle Uy, x^0 - x\rangle \\ &\le \frac{L_f}{2}\|x^0 - x\|^2 \le L_f P(x^0, x), \end{aligned}$$

where the last inequality follows from (2.2). We now show part b). By the above relation, the convexity of h and ω, and the optimality of (x^*, y^*), we have

$$\begin{aligned} D(y^0, y^*) &= f(x^0) - f(x^*) - \langle Uy^*, x^0 - x^*\rangle \\ &= f(x^0) - f(x^*) + \langle h'(x^*) + \mu\omega'(x^*), x^0 - x^*\rangle - \langle Uy^* + h'(x^*) + \mu\omega'(x^*), x^0 - x^*\rangle \\ &\le f(x^0) - f(x^*) + \langle h'(x^*) + \mu\omega'(x^*), x^0 - x^*\rangle \le \Psi(x^0) - \Psi(x^*). \end{aligned}$$

The proof of part c) is similar to part a) and hence the details are skipped. ∎

The following lemma gives an important bound for the primal optimality gap $\Psi(\bar{x}) - \Psi(x^*)$ for some $\bar{x} \in X$.

Lemma 4 *Let $(\bar{x}, \bar{y}) \in Z$ be a given pair of feasible solutions of (3.1), and $z^* = (x^*, y^*)$ be a pair of optimal solutions of (3.1). Then, we have*

$$\Psi(\bar{x}) - \Psi(x^*) \le Q((\bar{x}, \bar{y}), z^*) + \frac{L_f}{2}\|\bar{x} - x^*\|^2. \tag{5.4}$$

Proof. Let $\bar{y}_* = (\nabla f_1(\bar{x}); \nabla f_2(\bar{x}); \ldots; \nabla f_m(\bar{x}))$, and by the definition of $Q(\cdot,\cdot)$ in (3.3), we have

$$\begin{aligned}
Q((\bar{x},\bar{y}), z^*) &= \big[h(\bar{x}) + \mu\omega(\bar{x}) + \langle \bar{x}, Uy^*\rangle - J(y^*)\big] - \big[h(x^*) + \mu\omega(x^*) + \langle x^*, U\bar{y}\rangle - J(\bar{y})\big] \\
&\geq [h(\bar{x}) + \mu\omega(\bar{x}) + \langle \bar{x}, U\bar{y}_*\rangle - J(\bar{y}_*)] + \langle \bar{x}, U(y^* - \bar{y}_*)\rangle - J(y^*) + J(\bar{y}_*) \\
&\quad - \left[h(x^*) + \mu\omega(x^*) + \max_{y\in\mathcal{Y}} \big\{\langle x^*, Uy\rangle - J(y)\big\}\right] \\
&= \Psi(\bar{x}) - \Psi(x^*) + \langle \bar{x}, U(y^* - \bar{y}_*)\rangle - \langle x^*, Uy^*\rangle + f(x^*) + \langle \bar{x}, U\bar{y}_*\rangle - f(\bar{x}) \\
&= \Psi(\bar{x}) - \Psi(x^*) + f(x^*) - f(\bar{x}) + \langle \bar{x} - x^*, \nabla f(x^*)\rangle \geq \Psi(\bar{x}) - \Psi(x^*) - \frac{L_f}{2}\|\bar{x} - x^*\|^2,
\end{aligned}$$

where the second equality follows from the fact that $J_i, i = 1, \ldots, m$, are the conjugate functions of f_i. ∎

5.2 General results for both PDG and RPDG

We will establish some general convergence results in Proposition 5 which holds for both deterministic and randomized PDG methods by viewing PDG as a special case of RPDG with $m = 1$. Then both Theorems 1 and 2 follow as some immediate consequences of Proposition 5.

Before showing Proposition 5 we will develop a few technical results. Lemma 5 below characterizes the solutions of the prox-mapping in (2.3) and (3.5). This result generalizes some previous results (e.g., Lemma 6 of [20] and Lemma 2 of [14]).

Lemma 5 *Let U be a closed convex set and a point $\tilde{u} \in U$ be given. Also let $w : U \to \mathbb{R}$ be a convex function and*

$$W(\tilde{u}, u) = w(u) - w(\tilde{u}) - \langle w'(\tilde{u}), u - \tilde{u}\rangle, \tag{5.5}$$

for some $w'(\tilde{u}) \in \partial w(\tilde{u})$. Assume that the function $q : U \to \mathbb{R}$ satisfies

$$q(u_1) - q(u_2) - \langle q'(u_2), u_1 - u_2\rangle \geq \mu_0 W(u_2, u_1), \quad \forall u_1, u_2 \in U \tag{5.6}$$

for some $\mu_0 \geq 0$. Also assume that the scalars μ_1 and μ_2 are chosen such that $\mu_0 + \mu_1 + \mu_2 \geq 0$. If

$$u^* \in \mathrm{Argmin}\{q(u) + \mu_1 w(u) + \mu_2 W(\tilde{u}, u) : u \in U\}, \tag{5.7}$$

then for any $u \in U$, we have

$$q(u^*) + \mu_1 w(u^*) + \mu_2 W(\tilde{u}, u^*) + (\mu_0 + \mu_1 + \mu_2) W(u^*, u) \leq q(u) + \mu_1 w(u) + \mu_2 W(\tilde{u}, u).$$

Proof. Let $\phi(u) := q(u) + \mu_1 w(u) + \mu_2 W(\tilde{u}, u)$. It can be easily checked that for any $u_1, u_2 \in U$,

$$\begin{aligned}
W(\tilde{u}, u_1) &= W(\tilde{u}, u_2) + \langle W'(\tilde{u}, u_2), u_1 - u_2\rangle + W(u_2, u_1), \\
w(u_1) &= w(u_2) + \langle w'(u_2), u_1 - u_2\rangle + W(u_2, u_1).
\end{aligned}$$

Using these relations and (5.6), we conclude that

$$\phi(u_1) - \phi(u_2) - \langle \phi'(u_2), u_1 - u_2\rangle \geq (\mu_0 + \mu_1 + \mu_2) W(u_2, u_1) \tag{5.8}$$

for any $u_1, u_2 \in Y$, which together with the fact that $\mu_0 + \mu_1 + \mu_2 \geq 0$ then imply that ϕ is convex. Since u^* is an optimal solution of (5.7), we have $\langle \phi'(u^*), u - u^*\rangle \geq 0$. Combining this inequality with (5.8), we conclude that

$$\phi(u) - \phi(u^*) \geq (\mu_0 + \mu_1 + \mu_2) W(u^*, u),$$

from which the result immediately follows. ∎

The following simple result provides a few identities related to y^t and $\tilde{y}^t$ that will be useful for the analysis of the PDG algorithm.

Lemma 6 *Let y^t, $\tilde{y}^t$, and $\hat{y}^t$ be defined in (3.8), (3.9), and (3.11), respectively. Then we have, for any $i = 1, \ldots, m$ and $t = 1, \ldots, k$,*

$$\mathbb{E}_t[D_i(y_i^{t-1}, y_i^t)] = p_i D_i(y_i^{t-1}, \hat{y}_i^t), \tag{5.9}$$

$$\mathbb{E}_t[D_i(y_i^t, y_i)] = p_i D_i(\hat{y}_i^t, y_i) + (1 - p_i) D_i(y_i^{t-1}, y_i), \tag{5.10}$$

for any $y \in \mathcal{Y}$, where $\mathbb{E}_t$ denotes the conditional expectation w.r.t. i_t given $i_1, \ldots, i_{t-1}$.

Proof. (5.9) follows immediately from the facts that $\mathrm{Prob}_t\{y_i^t = \hat{y}_i^t\} = \mathrm{Prob}_t\{i_t = i\} = p_i$ and $\mathrm{Prob}_t\{y_i^t = y_i^{t-1}\} = 1 - p_i$. Here Prob_t denotes the conditional probability w.r.t. i_t given $i_1, \ldots, i_{t-1}$. Similarly, we can show (5.10). ∎

We now prove an important recursion about the RPDG method.

Lemma 7 *Let the gap function Q be defined in (3.3). Also let x^t and $\hat{y}^t$ be defined in (3.10) and (3.11), respectively. Then for any $t \geq 1$, we have*

$$\begin{aligned}\mathbb{E}[Q((x^t, \hat{y}^t), z)] \leq\ & \mathbb{E}\left[\eta_t P(x^{t-1}, x) - (\mu + \eta_t) P(x^t, x) - \eta_t P(x^{t-1}, x^t)\right] \\ & + \textstyle\sum_{i=1}^m \mathbb{E}\left[\left(p_i^{-1}(1 + \tau_t) - 1\right) D_i(y_i^{t-1}, y_i) - p_i^{-1}(1 + \tau_t) D_i(y_i^t, y_i)\right] \\ & + \mathbb{E}\left[\langle \tilde{x}^t - x^t, U(\tilde{y}^t - y)\rangle - \tau_t p_{i_t}^{-1} D_{i_t}(y_{i_t}^{t-1}, y_{i_t}^t)\right], \quad \forall z \in Z. \end{aligned} \tag{5.11}$$

Proof. It follows from Lemma 5 applied to (3.10) that $\forall x \in X$,

$$\langle x^t - x, U\tilde{y}^t\rangle + h(x^t) + \mu\omega(x^t) - h(x) - \mu\omega(x) \leq \eta_t P(x^{t-1}, x) - (\mu + \eta_t) P(x^t, x) - \eta_t P(x^{t-1}, x^t). \tag{5.12}$$

Moreover, by Lemma 5 applied to (3.11), we have, for any $i = 1, \ldots, m$ and $t = 1, \ldots, k$,

$$\langle -\tilde{x}^t, \hat{y}_i^t - y_i\rangle + J_i(\hat{y}_i^t) - J_i(y_i) \leq \tau_t D_i(y_i^{t-1}, y_i) - (1 + \tau_t) D_i(\hat{y}_i^t, y_i) - \tau_t D_i(y_i^{t-1}, \hat{y}_i^t).$$

Summing up these inequalities over $i = 1, \ldots, m$, we have, $\forall y \in \mathcal{Y}$,

$$\langle -\tilde{x}^t, U(\hat{y}^t - y)\rangle + J(\hat{y}^t) - J(y) \leq \textstyle\sum_{i=1}^m \left[\tau_t D_i(y_i^{t-1}, y_i) - (1 + \tau_t) D_i(\hat{y}_i^t, y_i) - \tau_t D_i(y_i^{t-1}, \hat{y}_i^t)\right]. \tag{5.13}$$

Using the definition of Q in (3.3), (5.12), and (5.13), we have

$$\begin{aligned} Q((x^t, \hat{y}^t), z) \leq\ & \eta_t P(x^{t-1}, x) - (\mu + \eta_t) P(x^t, x) - \eta_t P(x^{t-1}, x^t) \\ & + \textstyle\sum_{i=1}^m \left[\tau_t D_i(y_i^{t-1}, y_i) - (1 + \tau_t) D_i(\hat{y}_i^t, y_i) - \tau_t D_i(y_i^{t-1}, \hat{y}_i^t)\right] \\ & + \langle \tilde{x}^t, U(\hat{y}^t - y)\rangle - \langle x^t, U(\tilde{y}^t - y)\rangle + \langle x, U(\tilde{y}^t - \hat{y}^t)\rangle. \end{aligned} \tag{5.14}$$

Also observe that by (3.8), (3.12), (5.9), and (5.10),

$$\begin{aligned} D_i(y_i^{t-1}, \hat{y}_i^t) &= 0, \ \forall i \neq i_t, \\ \mathbb{E}[\langle x, U(\tilde{y}^t - \hat{y}^t)\rangle] &= 0, \\ \mathbb{E}[\langle \tilde{x}^t, U\hat{y}^t\rangle] &= \mathbb{E}[\langle \tilde{x}^t, U\tilde{y}^t\rangle], \\ \mathbb{E}[D_i(y_i^{t-1}, \hat{y}_i^t)] &= \mathbb{E}[p_i^{-1} D_i(y_i^{t-1}, y_i^t)] \\ \mathbb{E}[D_i(\hat{y}_i^t, y_i)] &= p_i^{-1}\mathbb{E}[D_i(y_i^t, y_i)] - (p_i^{-1} - 1)\mathbb{E}[D_i(y_i^{t-1}, y_i)], \end{aligned}$$

Taking expectation on both sides of (5.14) and using the above observations, we obtain (5.11). ∎

We are now ready to establish a general convergence result which holds for both PDG and RPDG.

Proposition 5 *Suppose that $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ in the RPDG method satisfy*

$$\theta_t\left(p_i^{-1}(1+\tau_t)-1\right) \le p_i^{-1}\theta_{t-1}(1+\tau_{t-1}), i=1,\ldots,m; t=2,\ldots,k, \tag{5.15}$$

$$\theta_t\eta_t \le \theta_{t-1}(\mu+\eta_{t-1}), t=2,\ldots,k, \tag{5.16}$$

$$\tfrac{\eta_k}{4} \ge \tfrac{L_i(1-p_i)^2}{\tau_k p_i}, i=1,\ldots,m, \tag{5.17}$$

$$\tfrac{\eta_{t-1}}{2} \ge \tfrac{L_i\alpha_t}{\tau_t p_i} + \tfrac{(1-p_j)^2L_j}{\tau_{t-1}p_j},\ i,j\in\{1,\ldots,m\}; t=2,\ldots,k, \tag{5.18}$$

$$\tfrac{\eta_k}{2} \ge \tfrac{\sum_{i=1}^m (p_iL_i)}{1+\tau_k}, \tag{5.19}$$

$$\alpha_t\theta_t = \theta_{t-1}, t=2,\ldots,k, \tag{5.20}$$

for some $\theta_t \ge 0$, $t=1,\ldots,k$. Then, for any $k\ge 1$ and any given $z\in Z$, we have

$$\begin{aligned}\textstyle\sum_{t=1}^k \theta_t\mathbb{E}[Q((x^t,\hat y^t),z)] \le{}& \eta_1\theta_1P(x^0,x) - (\mu+\eta_k)\theta_k\mathbb{E}[P(x^k,x)]\\ &+ \textstyle\sum_{i=1}^m \theta_1\left(p_i^{-1}(1+\tau_1)-1\right)D_i(y_i^0,y_i).\end{aligned} \tag{5.21}$$

Proof. Multiplying both sides of (5.11) by θ_t and summing the resulting inequalities, we have

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=1}^k\theta_tQ((x^t,\hat y^t),z)] \le{}& \mathbb{E}\left[\textstyle\sum_{t=1}^k\theta_t\left(\eta_tP(x^{t-1},x)-(\mu+\eta_t)P(x^t,x)-\eta_tP(x^{t-1},x^t)\right)\right]\\ &+\textstyle\sum_{i=1}^m\mathbb{E}\left\{\sum_{t=1}^k\theta_t\left[\left(p_i^{-1}(1+\tau_t)-1\right)D_i(y_i^{t-1},y_i)-p_i^{-1}(1+\tau_t)D_i(y_i^t,y_i)\right]\right\}\\ &+\mathbb{E}\left[\textstyle\sum_{t=1}^k\theta_t\left(\langle\tilde x^t-x^t,U(\tilde y^t-y)\rangle-\tau_tp_{i_t}^{-1}D_{i_t}(y_{i_t}^{t-1},y_{i_t}^t)\right)\right],\end{aligned}$$

which, in view of the assumptions in (5.16) and (5.15), then implies that

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=1}^k\theta_tQ((x^t,\hat y^t),z)] \le{}& \eta_1\theta_1P(x^0,x)-(\mu+\eta_k)\theta_k\mathbb{E}[P(x^k,x)]\\ &+\textstyle\sum_{i=1}^m\left[\theta_1\left(p_i^{-1}(1+\tau_1)-1\right)D_i(y_i^0,y_i)-p_i^{-1}\theta_k(1+\tau_k)D_i(y_i^k,y_i)\right]\\ &-\mathbb{E}\left[\textstyle\sum_{t=1}^k\theta_t\Delta_t\right],\end{aligned} \tag{5.22}$$

where

$$\Delta_t := \eta_tP(x^{t-1},x^t)-\langle\tilde x^t-x^t,U(\tilde y^t-y)\rangle+\tau_tp_{i_t}^{-1}D_{i_t}(y_{i_t}^{t-1},y_{i_t}^t). \tag{5.23}$$

We now provide a bound on $\sum_{t=1}^k\theta_t\Delta_t$ in (5.22). Note that by (3.7), we have

$$\begin{aligned}\langle\tilde x^t-x^t,U(\tilde y^t-y)\rangle &= \langle x^{t-1}-x^t,U(\tilde y^t-y)\rangle-\alpha_t\langle x^{t-2}-x^{t-1},U(\tilde y^t-y)\rangle\\ &= \langle x^{t-1}-x^t,U(\tilde y^t-y)\rangle-\alpha_t\langle x^{t-2}-x^{t-1},U(\tilde y^{t-1}-y)\rangle\\ &\quad-\alpha_t\langle x^{t-2}-x^{t-1},U(\tilde y^t-\tilde y^{t-1})\rangle\\ &= \langle x^{t-1}-x^t,U(\tilde y^t-y)\rangle-\alpha_t\langle x^{t-2}-x^{t-1},U(\tilde y^{t-1}-y)\rangle\\ &\quad-\alpha_tp_{i_t}^{-1}\langle x^{t-2}-x^{t-1},y_{i_t}^t-y_{i_t}^{t-1}\rangle\\ &\quad-\alpha_t(p_{i_{t-1}}^{-1}-1)\langle x^{t-2}-x^{t-1},y_{i_{t-1}}^{t-2}-y_{i_{t-1}}^{t-1}\rangle,\end{aligned} \tag{5.24}$$

where the last identity follows from the observation that by (3.8) and (3.9),

$$\begin{aligned}U(\tilde y^t-\tilde y^{t-1}) &= \textstyle\sum_{i=1}^m\left\{\left[p_i^{-1}(y_i^t-y_i^{t-1})+y_i^{t-1}\right]-\left[p_i^{-1}(y_i^{t-1}-y_i^{t-2})+y_i^{t-2}\right]\right\}\\ &= \textstyle\sum_{i=1}^m\left\{\left[p_i^{-1}y_i^t-(p_i^{-1}-1)y_i^{t-1}\right]-\left[p_i^{-1}y_i^{t-1}-(p_i^{-1}-1)y_i^{t-2}\right]\right\}\\ &= \textstyle\sum_{i=1}^m\left[p_i^{-1}(y_i^t-y_i^{t-1})+(p_i^{-1}-1)(y_i^{t-2}-y_i^{t-1})\right]\\ &= p_{i_t}^{-1}(y_{i_t}^t-y_{i_t}^{t-1})+(p_{i_{t-1}}^{-1}-1)(y_{i_{t-1}}^{t-2}-y_{i_{t-1}}^{t-1}).\end{aligned}$$

Using relation (5.24) in the definition of Δ_t in (5.23), we have

$$\begin{aligned}\textstyle\sum_{t=1}^k \theta_t \Delta_t = \textstyle\sum_{t=1}^k \theta_t \big[&\eta_t P(x^{t-1}, x^t) \\ &- \langle x^{t-1} - x^t, U(\tilde{y}^t - y)\rangle + \alpha_t \langle x^{t-2} - x^{t-1}, U(\tilde{y}^{t-1} - y)\rangle \\ &+ \alpha_t p_{i_t}^{-1} \langle x^{t-2} - x^{t-1}, y_{i_t}^t - y_{i_t}^{t-1}\rangle + \alpha_t (p_{i_{t-1}}^{-1} - 1)\langle x^{t-2} - x^{t-1}, y_{i_{t-1}}^{t-2} - y_{i_{t-1}}^{t-1}\rangle \\ &+ p_{i_t}^{-1} \tau_t D_{i_t}(y_{i_t}^{t-1}, y_{i_t}^t)\Big]. \end{aligned} \tag{5.25}$$

Observe that by (5.20) and the fact that $x^{-1} = x^0$,

$$\begin{aligned}&\textstyle\sum_{t=1}^k \theta_t \left[\langle x^{t-1} - x^t, U(\tilde{y}^t - y)\rangle - \alpha_t \langle x^{t-2} - x^{t-1}, U(\tilde{y}^{t-1} - y)\rangle\right] \\ &= \theta_k \langle x^{k-1} - x^k, U(\tilde{y}^k - y)\rangle \\ &= \theta_k \langle x^{k-1} - x^k, U(y^k - y)\rangle + \theta_k \langle x^{k-1} - x^k, U(\tilde{y}^k - y^k)\rangle \\ &= \theta_k \langle x^{k-1} - x^k, U(y^k - y)\rangle + \theta_k (p_{i_k}^{-1} - 1)\langle x^{k-1} - x^k, y_{i_k}^k - y_{i_k}^{k-1}\rangle,\end{aligned}$$

where the last identity follows from the definitions of y^k and $\tilde{y}^k$ in (3.8) and (3.9), respectively. Also, by the strong convexity of P and D_i, we have

$$P(x^{t-1}, x^t) \geq \tfrac{1}{2}\|x^{t-1} - x^t\|^2 \quad \text{and} \quad D_{i_t}(y_{i_t}^{t-1}, y_{i_t}^t) \geq \tfrac{1}{2L_{i_t}}\|y_{i_t}^{t-1} - y_{i_t}^t\|^2.$$

Using the previous three relations in (5.25), we have

$$\begin{aligned}\textstyle\sum_{t=1}^k \theta_t \Delta_t \geq \textstyle\sum_{t=1}^k \theta_t \big[&\tfrac{\eta_t}{2}\|x^{t-1} - x^t\|^2 + \alpha_t p_{i_t}^{-1}\langle x^{t-2} - x^{t-1}, y_{i_t}^t - y_{i_t}^{t-1}\rangle \\ &+ \alpha_t (p_{i_{t-1}}^{-1} - 1)\langle x^{t-2} - x^{t-1}, y_{i_{t-1}}^{t-2} - y_{i_{t-1}}^{t-1}\rangle + \tfrac{\tau_t}{2L_{i_t} p_{i_t}}\|y_{i_t}^{t-1} - y_{i_t}^t\|^2\big] \\ &- \theta_k \langle x^{k-1} - x^k, U(y^k - y)\rangle - \theta_k (p_{i_k}^{-1} - 1)\langle x^{k-1} - x^k, y_{i_k}^k - y_{i_k}^{k-1}\rangle.\end{aligned}$$

Regrouping the terms in the above relation, and the fact that $x^{-1} = x^0$, we obtain

$$\begin{aligned}\textstyle\sum_{t=1}^k \theta_t \Delta_t &\geq \theta_k \left[\tfrac{\eta_k}{4}\|x^{k-1} - x^k\|^2 - \langle x^{k-1} - x^k, U(y^k - y)\rangle\right] \\ &\quad + \theta_k \left[\tfrac{\eta_k}{4}\|x^{k-1} - x^k\|^2 - (p_{i_k}^{-1} - 1)\langle x^{k-1} - x^k, y_{i_k}^k - y_{i_k}^{k-1}\rangle + \tfrac{\tau_k}{4L_{i_k} p_{i_k}}\|y_{i_k}^{k-1} - y_{i_k}^k\|^2\right] \\ &\quad + \textstyle\sum_{t=2}^k \theta_t \left[\tfrac{\alpha_t}{p_{i_t}}\langle x^{t-2} - x^{t-1}, y_{i_t}^t - y_{i_t}^{t-1}\rangle + \tfrac{\tau_t}{4L_{i_t} p_{i_t}}\|y_{i_t}^{t-1} - y_{i_t}^t\|^2\right] \\ &\quad + \textstyle\sum_{t=2}^k \left[\alpha_t \theta_t (p_{i_{t-1}}^{-1} - 1)\langle x^{t-2} - x^{t-1}, y_{i_{t-1}}^{t-2} - y_{i_{t-1}}^{t-1}\rangle + \tfrac{\tau_{t-1}\theta_{t-1}}{4L_{i_{t-1}} p_{i_{t-1}}}\|y_{i_{t-1}}^{t-2} - y_{i_{t-1}}^{t-1}\|^2\right] \\ &\quad + \textstyle\sum_{t=2}^k \tfrac{\theta_{t-1}\eta_{t-1}}{2}\|x^{t-2} - x^{t-1}\|^2 \\ &\geq \theta_k \left[\tfrac{\eta_k}{4}\|x^{k-1} - x^k\|^2 - \langle x^{k-1} - x^k, U(y^k - y)\rangle\right] \\ &\quad + \theta_k \left(\tfrac{\eta_k}{4} - \tfrac{L_{i_k}(1 - p_{i_k})^2}{\tau_k p_{i_k}}\right)\|x^{k-1} - x^k\|^2 \\ &\quad + \textstyle\sum_{t=2}^k \left[\tfrac{\theta_{t-1}\eta_{t-1}}{2} - \tfrac{L_{i_t}\alpha_t^2\theta_t}{\tau_t p_{i_t}} - \tfrac{\alpha_t^2\theta_t^2(1 - p_{i_{t-1}})^2 L_{i_{t-1}}}{\tau_{t-1}\theta_{t-1} p_{i_{t-1}}}\right]\|x^{t-2} - x^{t-1}\|^2 \\ &= \theta_k \left[\tfrac{\eta_k}{4}\|x^{k-1} - x^k\|^2 - \langle x^{k-1} - x^k, U(y^k - y)\rangle\right] \\ &\quad + \theta_k \left(\tfrac{\eta_k}{4} - \tfrac{L_{i_k}(1 - p_{i_k})^2}{\tau_k p_{i_k}}\right)\|x^{k-1} - x^k\|^2 \\ &\quad + \textstyle\sum_{t=2}^k \theta_{t-1}\left(\tfrac{\eta_{t-1}}{2} - \tfrac{L_{i_t}\alpha_t}{\tau_t p_{i_t}} - \tfrac{(1 - p_{i_{t-1}})^2 L_{i_{t-1}}}{\tau_{t-1} p_{i_{t-1}}}\right)\|x^{t-2} - x^{t-1}\|^2 \\ &\geq \theta_k \left[\tfrac{\eta_k}{4}\|x^{k-1} - x^k\|^2 - \langle x^{k-1} - x^k, U(y^k - y)\rangle\right],\end{aligned} \tag{5.26}$$

where the second inequality follows from the simple relation that

$$b\langle u,v\rangle + a\|v\|^2/2 \geq -b^2\|u\|^2/(2a), \forall a>0, \tag{5.27}$$

and the last inequality follows from (5.17) and (5.18). Plugging the bound (5.26) into (5.22), we have

$$\begin{aligned}\textstyle\sum_{t=1}^k \theta_t \mathbb{E}[Q((x^t,\hat{y}^t),z)] \leq \theta_1\eta_1 P(x^0,x) - \theta_k(\mu+\eta_k)\mathbb{E}[P(x^k,x)] + \textstyle\sum_{i=1}^m \theta_1\left(p_i^{-1}(1+\tau_1)-1\right)D_i(y_i^0,y_i)\\ - \theta_k\mathbb{E}\left[\tfrac{\eta_k}{4}\|x^{k-1}-x^k\|^2 - \langle x^{k-1}-x^k, U(y^k-y)\rangle + \textstyle\sum_{i=1}^m p_i^{-1}(1+\tau_k)D_i(y_i^k,y_i)\right].\end{aligned}$$

Also observe that by (5.19) and (5.27),

$$\begin{aligned}&\tfrac{\eta_k}{4}\|x^{k-1}-x^k\|^2 - \langle x^{k-1}-x^k, U(y^k-y)\rangle + \textstyle\sum_{i=1}^m p_i^{-1}(1+\tau_k)D_i(y_i^k,y_i)\\ &\geq \tfrac{\eta_k}{4}\|x^{k-1}-x^k\|^2 + \textstyle\sum_{i=1}^m\left[-\langle x^{k-1}-x^k, y_i^k-y_i\rangle + \tfrac{1+\tau_k}{2L_ip_i}\|y_i^k-y_i\|^2\right]\\ &\geq \left(\tfrac{\eta_k}{4} - \tfrac{\sum_{i=1}^m (p_iL_i)}{2(1+\tau_k)}\right)\|x^{k-1}-x^k\|^2 \geq 0,\end{aligned}$$

The result then immediately follows by combining the above two conclusion. ∎

5.3 Proof of main convergence results

We now provide a proof for Theorem 1 which describes the main convergence properties of the deterministic PDG method.

We first specialize Proposition 5 for the PDG method applied to (2.7).

Proposition 6 *Suppose that $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ in the PDG method satisfy*

$$\theta_t\tau_t \leq \theta_{t-1}(1+\tau_{t-1}), t=2,\ldots,k, \tag{5.28}$$

$$\theta_t\eta_t \leq \theta_{t-1}(\mu+\eta_{t-1}), t=2,\ldots,k, \tag{5.29}$$

$$\eta_{t-1}\tau_t \geq 2L_f\alpha_t, t=2,\ldots,k, \tag{5.30}$$

$$\eta_k(1+\tau_k) \geq 2L_f, \tag{5.31}$$

$$\alpha_t = \theta_{t-1}/\theta_t, t=2,\ldots,k, \tag{5.32}$$

for some $\theta_t \geq 0$, $t=1,\ldots,k$. Also let us denote $z^t=(x^t,g^t)$, and

$$\bar{z}^k := \left(\textstyle\sum_{t=1}^k\theta_t\right)^{-1}\sum_{t=1}^k\theta_t z^t. \tag{5.33}$$

Then, for any $k\geq 1$ and any given $(x,g)\in X\times\mathcal{G}$, we have

$$\left(\textstyle\sum_{t=1}^k\theta_t\right)Q_f(\bar{z}^k,z) + \theta_k(\mu+\eta_k)P(x^k,x) \leq \theta_1\eta_1P(x^0,x) + \theta_1\tau_1D_f(g^0,g). \tag{5.34}$$

Proof. Notice that in the deterministic PDG method, we have $m=1$, $p_i=1$, and $\hat{y}^t=g^t$. It can be easily seen that the assumptions in (5.15)-(5.20) are implied by those in (5.28)-(5.32). It then follows from (5.21) that

$$\textstyle\sum_{t=1}^k\theta_tQ_f(z^t,z) \leq \theta_1\eta_1P(x^0,x) - \theta_k(\mu+\eta_k)P(x^k,x) + \theta_1\tau_1D_f(g^0,g).$$

Dividing both sides of the above inequality by $\sum_{t=1}^k\theta_t$ and using the convexity of $Q(\bar{z},z)$ w.r.t. $\bar{z}$, we have

$$\left(\textstyle\sum_{t=1}^k\theta_t\right)Q_f(\bar{z}^k,z) \leq \textstyle\sum_{t=1}^k\theta_tQ_f(z^t,z) \leq \theta_1\eta_1P(x^0,x) - \theta_k(\mu+\eta_k)P(x^k,x) + \theta_1\tau_1D_f(g^0,g).$$

Rearranging the terms in the above relation, we obtain (5.34). ∎

We are now ready to show Theorem 1.

Proof of Theorem 1 We first show part a). It can be easily checked that (5.28)-(5.32) are satisfied with the selection of $\{\tau_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$, and $\{\theta_t\}$ in (2.24). Using (5.34) (with $x = x^*$ and $y = y^*$), (5.3), and the fact that $Q_f(\bar{z}, z^*) \geq 0$, we have

$$\theta_k(\mu + \eta_k)P(x^k, x^*) \leq \theta_1(\eta_1 + L_f\tau_1)P(x^0, x^*), \quad \forall k \geq 1.$$

Using the parameter settings in (2.24), we conclude that

$$P(x^k, x^*) \leq \tfrac{\theta_1(\eta_1+L_f\tau_1)}{\theta_k(\mu+\eta_k)}P(x^0, x^*) = \tfrac{(\sqrt{2L_f\mu}+L_f\sqrt{2L_f/\mu})}{\alpha(\mu+\sqrt{2L_f\mu})}\alpha^k P(x^0, x^*) = \tfrac{\mu+L_f}{\mu}\alpha^k P(x^0, x^*).$$

Also using (5.34) and the fact that $P(x^k, x) \geq 0$, we have

$$\left(\textstyle\sum_{t=1}^k \theta_t\right) Q_f(\bar{z}^k, z) \leq \theta_1\eta_1 P(x^0, x) + \theta_1\tau_1 D_f(g^0, g), \quad \forall z \in Z. \tag{5.35}$$

Denoting $\bar{g}_*^k := (\nabla f_1(\bar{x}^k); \ldots; \nabla f_m(\bar{x}^k))$, we conclude from (5.3) that

$$\begin{aligned} D_f(g^0, \bar{g}_*^k) &\leq \tfrac{L_f}{2}\|\bar{x}^k - x^0\|^2 \leq \tfrac{L_f}{2}[\textstyle\sum_{t=1}^k\theta_t]^{-1}\sum_{t=1}^k\theta_t\|x^t - x^0\|^2 \\ &\leq \tfrac{L_f}{2}[\textstyle\sum_{t=1}^k\theta_t]^{-1}\sum_{t=1}^k\theta_t(\|x^t - x^*\|^2 + \|x^0 - x^*\|^2) \\ &\leq \tfrac{L_f}{2}\left[\tfrac{2(\mu+L_f)}{\mu}P(x^0, x^*) + \|x^0 - x^*\|^2\right] \leq L_f\left(\tfrac{2\mu+L_f}{\mu}\right)P(x^0, x^*), \end{aligned}$$

where the second inequality follows from the convexity of $\|\cdot\|^2$, the third inequality follows from the triangular inequality, the fourth inequality follows from $\|x^t - x^*\|^2 \leq 2P(x^t, x^*)$ and (2.25), and the last inequality follows from $\|x^0 - x^*\|^2 \leq 2P(x^0, x^*)$. Also note that by the definition of θ_t, we have

$$\textstyle\sum_{t=1}^k\theta_t = \sum_{t=1}^k\alpha^{-t} = \frac{1-\alpha^k}{(1-\alpha)\alpha^k} \geq \frac{1}{\alpha^k}, \tag{5.36}$$

where the last inequality follows from the fact that $\alpha \leq 1$ due to (2.24). Fixing $g = \bar{g}_*^k$ in (5.35) and using the above two relations, we obtain

$$\begin{aligned} Q_f(\bar{z}^k, (x, \bar{g}_*^k)) &\leq \alpha^k\left[\theta_1\eta_1 P(x^0, x) + L_f\theta_1\tau_1\left(\tfrac{2\mu+L_f}{\mu}\right)P(x^0, x^*)\right] \\ &\leq (\mu + \sqrt{2L\mu})\alpha^k\left[P(x^0, x) + \tfrac{L_f}{\mu}(2 + \tfrac{L_f}{\mu})P(x^0, x^*)\right] \\ &= \tfrac{\mu\alpha^k}{1-\alpha}\left[P(x^0, x) + \tfrac{L_f}{\mu}(2 + \tfrac{L_f}{\mu})P(x^0, x^*)\right]. \end{aligned}$$

The result in (2.26) then directly follows from the above relation and (2.21). If X is bounded, the result in (2.27) then follows from the above relation, (2.21), and (2.22).

We now show part b). It is trivial to check that the conditions in (5.28)-(5.32) hold by using our selection of $\{\tau_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$, and $\{\theta_t\}$. Using (5.34) and the facts $\tau_1 = 0$ and $P(x^k, x) \geq 0$, we have

$$\left(\textstyle\sum_{t=1}^k \theta_t\right) Q_f(\bar{z}^k, z) \leq \theta_1\eta_1 P(x^0, x) = 4L_f P(x^0, x).$$

which, in view of (2.20) and (2.21) and the fact that $\sum_{t=1}^k\theta_t = k(k+1)/2$, clearly implies (2.29). In case X is bounded, the result in (2.30) immediately follows from (2.21), (2.22), and the above inequality. ∎

We are now ready to provide a proof for Theorem 2, which describes the main convergence properties of the RPDG method applied to strongly convex problems with $\mu > 0$.

Proof of Theorem 2. It can be easily checked that the conditions in (5.15)-(5.20) are satisfied with our requirements (3.19)-(3.22) of $\{\tau_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$, and $\{\theta_t\}$. Using the fact that $Q((x^t, \bar{y}^t), z^*) \geq 0$, we then conclude from (5.21) (with $x = x^*$ and $y = y^*$) that, for any $k \geq 1$,

$$\mathbb{E}[P(x^k, x^*)] \leq \tfrac{1}{\theta_k(\mu+\eta)}\left[\theta_1\eta P(x^0, x^*) + \tfrac{\theta_1\alpha}{1-\alpha}D(y^0, y^*)\right] \leq \left(1 + \tfrac{L_f\alpha}{(1-\alpha)\eta}\right)\alpha^k P(x^0, x^*),$$

where the first inequality follows from (3.19) and (3.20), and the second inequality follows from (3.21) and (5.1).

Let us denote $\bar{y}^k \equiv (\sum_{t=1}^k \theta_t)^{-1} \sum_{t=1}^k (\theta_t \hat{y}^t)$, $\bar{z}^k = (\bar{x}^k, \bar{y}^k)$. In view of (5.4), the convexity of $\|\cdot\|$, and (2.2), we have

$$\begin{aligned}\mathbb{E}[\Psi(\bar{x}^k) - \Psi(x^*)] &\le \mathbb{E}[Q(\bar{z}^k, z^*)] + \tfrac{L_f}{2}(\textstyle\sum_{t=1}^k \theta_t)^{-1}\mathbb{E}[\sum_{t=1}^k \theta_t \|x^t - x^*\|^2] \\ &\le \mathbb{E}[Q(\bar{z}^k, z^*)] + L_f(\textstyle\sum_{t=1}^k \theta_t)^{-1}\mathbb{E}[\sum_{t=1}^k \theta_t P(x^t, x^*)]. \end{aligned} \tag{5.37}$$

Using (5.21) (with $x = x^*$ and $y = y^*$), the fact that $P(x^k, x) \ge 0$, and (5.36), we obtain

$$\mathbb{E}[Q(\bar{z}^k, z^*)] \le \left(\textstyle\sum_{t=1}^k \theta_t\right)^{-1} \sum_{t=1}^k \theta_t \mathbb{E}[Q((x^t, \hat{y}^t), z^*)] \le \alpha^k \left(\alpha^{-1}\eta + \tfrac{L_f}{1-\alpha}\right) P(x^0, x^*).$$

We conclude from (3.23) and the definition of $\{\theta_t\}$ that

$$\begin{aligned}(\textstyle\sum_{t=1}^k \theta_t)^{-1}\mathbb{E}[\sum_{t=1}^k \theta_t P(x^t, x^*)] &= (\textstyle\sum_{t=1}^k \alpha^{-t})^{-1}\sum_{t=1}^k \alpha^{-t}(1 + \tfrac{L_f\alpha}{(1-\alpha)\eta})\alpha^t P(x^0, x^*) \\ &\le \tfrac{1-\alpha}{\alpha^{-k}-1}\textstyle\sum_{t=1}^k \tfrac{\alpha^t}{\alpha^{3t/2}}(1 + \tfrac{L_f\alpha}{(1-\alpha)\eta})P(x^0, x^*) \\ &= \tfrac{1-\alpha}{\alpha^{-k}-1}\tfrac{\alpha^{-k/2}-1}{1-\alpha^{1/2}}(1 + \tfrac{L_f\alpha}{(1-\alpha)\eta})P(x^0, x^*) \\ &= \tfrac{1+\alpha^{1/2}}{1+\alpha^{-k/2}}(1 + \tfrac{L_f\alpha}{(1-\alpha)\eta})P(x^0, x^*) \le 2\alpha^{k/2}(1 + \tfrac{L_f\alpha}{(1-\alpha)\eta})P(x^0, x^*).\end{aligned}$$

Using the above two relations, and (5.37), we obtain

$$\begin{aligned}\mathbb{E}[\Psi(\bar{x}^k) - \Psi(x^*)] &\le \alpha^k \left(\alpha^{-1}\eta + \tfrac{L_f}{1-\alpha}\right) P(x^0, x^*) + L_f 2\alpha^{k/2}\left(1 + \tfrac{L_f\alpha}{(1-\alpha)\eta}\right) P(x^0, x^*) \\ &\le \alpha^{k/2}\left(\alpha^{-1}\eta + \tfrac{3-2\alpha}{1-\alpha}L_f + \tfrac{2L_f^2\alpha}{(1-\alpha)\eta}\right) P(x^0, x^*).\end{aligned}$$

■

5.4 Proof of the lower complexity bound

This subsection is devoted to the proof of Theorem 3, which describes the performance limit for randomized incremental gradient methods.

The following result provides an explicit expression for the optimal solution of (3.37).

Lemma 8 *Let q be defined in (3.42), $x_{i,j}^*$ is the j-th element of x_i, and define*

$$x_{i,j}^* = q^j, i = 1, \ldots, m; j = 1, \ldots, \tilde{n}. \tag{5.38}$$

Then x^* is the unique optimal solution of (3.37).

Proof. It can be easily seen that q is the smallest root of the equation

$$q^2 - 2\tfrac{\mathcal{Q}+1}{\mathcal{Q}-1}q + 1 = 0. \tag{5.39}$$

Note that x^* satisfies the optimality condition of (3.37), i.e.,

$$\left(A + \tfrac{4}{\mathcal{Q}-1}I\right) x_i^* = e_1, \quad i = 1, \ldots, m. \tag{5.40}$$

Indeed, we can write the coordinate form of (5.40) as

$$2\tfrac{\mathcal{Q}+1}{\mathcal{Q}-1}x_{i,1}^* - x_{i,2}^* = 1, \tag{5.41}$$

$$x_{i,j+1}^* - 2\tfrac{\mathcal{Q}+1}{\mathcal{Q}-1}x_{i,j}^* + x_{i,j-1}^* = 0,\ j = 2, 3, \ldots, \tilde{n} - 1, \tag{5.42}$$

$$-(\kappa + \tfrac{4}{\mathcal{Q}-1})x_{i,\tilde{n}}^* + x_{i,\tilde{n}-1}^* = 0, \tag{5.43}$$

where the first two equations follow directly from the definition of x^* and relation (5.39), and the last equation is implied by the definitions of κ and x^* in (3.39) and (5.38), respectively. ■

We also need a few technical results to establish the lower complexity bounds.

Lemma 9 *a) For any $x > 1$, we have*

$$\log(1 - \tfrac{1}{x}) \geq -\tfrac{1}{x-1}. \tag{5.44}$$

b) Let $\rho, q, \bar{q} \in (0,1)$ be given. If we have

$$\tilde{n} \geq \tfrac{t \log \bar{q} + \log(1-\rho)}{2 \log q},$$

for any $t \geq 0$, then

$$\bar{q}^t - q^{2\tilde{n}} \geq \rho \bar{q}^t (1 - q^{2\tilde{n}}).$$

Proof. We first show part a). Denote $\phi(x) = \log(1 - \frac{1}{x}) + \frac{1}{x-1}$. It can be easily seen that $\lim_{x \to +\infty} \phi(x) = 0$. Moreover, for any $x > 1$, we have

$$\phi'(x) = \tfrac{1}{x(x-1)} - \tfrac{1}{(x-1)^2} = \tfrac{1}{x-1}\left(\tfrac{1}{x} - \tfrac{1}{x-1}\right) < 0,$$

which implies that ϕ is a strictly decreasing function for $x > 1$. Hence, we must have $\phi(x) > 0$ for any $x > 1$. Part b) follows from the following simple calculation.

$$\bar{q}^t - q^{2\tilde{n}} - \rho \bar{q}^t (1 - q^{2\tilde{n}}) = (1-\rho)\bar{q}^t - q^{2\tilde{n}} + \rho \bar{q}^t q^{2\tilde{n}} \geq (1-\rho)\bar{q}^t - q^{2\tilde{n}} \geq 0.$$

■

We are now ready to prove Theorem 3.

Proof of Theorem 3 Without loss of generality, we may assume that the initial point $x_i^0 = 0$, $i = 1, \ldots, m$. Indeed, the incremental gradient methods described in Subsection 3.3 are invariant with respect to a simultaneous shift of the decision variables. In other words, the sequence of iterates $\{x^k\}$, which is generated by such a method for minimizing the function $\Psi(x)$ starting from x^0, is just a shift of the sequence generated for minimizing $\bar{\Psi}(x) = \Psi(x + x^0)$ starting from the origin.

Now let k_i, $i = 1, \ldots, m$, denote the number of times that the gradients of the component function f_i are computed from iteration 1 to k. Clearly k_i's are binomial random variables supported on $\{0, 1, \ldots, k\}$ such that $\sum_{i=1}^m k_i = k$. Also observe that we must have $x_{i,j}^k = 0$ for any $k \geq 0$ and $k_j + 1 \leq j \leq \tilde{n}$, because each time the gradient ∇f_i is computed, the incremental gradient methods add at most one more nonzero entry to the i-th component of x^k due to the structure of the gradient ∇f_i. Therefore, we have

$$\tfrac{\|x^k - x^*\|_2^2}{\|x^0 - x^*\|_2^2} = \tfrac{\sum_{i=1}^m \|x_i^k - x_i^*\|_2^2}{\sum_{i=1}^m \|x_i^*\|^2} \geq \tfrac{\sum_{i=1}^m \sum_{j=k_i+1}^{\tilde{n}} (x_{i,j}^*)^2}{\sum_{i=1}^m \sum_{j=1}^{\tilde{n}} (x_{i,j}^*)^2} = \tfrac{\sum_{i=1}^m (q^{2k_i} - q^{2\tilde{n}})}{m(1-q^{2\tilde{n}})}. \tag{5.45}$$

Observing that for any $i = 1, \ldots, m$,

$$\mathbb{E}[q^{2k_i}] = \textstyle\sum_{t=0}^k \left[q^{2t} \binom{k}{t} p_i^t (1-p_i)^{k-t}\right] = [1 - (1-q^2)p_i]^k,$$

we then conclude from (5.45) that

$$\tfrac{\mathbb{E}[\|x^k - x^*\|_2^2]}{\|x^0 - x^*\|_2^2} \geq \tfrac{\sum_{i=1}^m [1 - (1-q^2)p_i]^k - m q^{2\tilde{n}}}{m(1-q^{2\tilde{n}})}.$$

Noting that $[1 - (1-q^2)p_i]^k$ is convex w.r.t. p_i for any $p_i \in [0,1]$ and $k \geq 1$, by minimizing the RHS of the above bound w.r.t. p_i, $i = 1, \ldots, m$, subject to $\sum_{i=1}^m p_i = 1$ and $p_i \geq 0$, we conclude that

$$\tfrac{\mathbb{E}[\|x^k - x^*\|_2^2]}{\|x^0 - x^*\|_2^2} \geq \tfrac{[1-(1-q^2)/m]^k - q^{2\tilde{n}}}{1-q^{2\tilde{n}}} \geq \tfrac{1}{2}[1 - (1-q^2)/m]^k, \tag{5.46}$$

for any $n \geq \underline{n}(m,k)$ (see (3.44)) and possible selection of p_i, $i = 1, \ldots, m$ satisfying (3.40), where the last inequality follows from Lemma 9.b). Noting that

$$\begin{aligned} 1 - (1-q^2)/m &= 1 - \left[1 - \left(\tfrac{\sqrt{\mathcal{Q}}-1}{\sqrt{\mathcal{Q}}+1}\right)^2\right] \tfrac{1}{m} = 1 - \tfrac{1}{m} + \tfrac{1}{m}\left(1 - \tfrac{2}{\sqrt{\mathcal{Q}}+1}\right)^2 \\ &= 1 - \tfrac{4}{m(\sqrt{\mathcal{Q}}+1)} + \tfrac{4}{m(\sqrt{\mathcal{Q}}+1)^2} = 1 - \tfrac{4\sqrt{\mathcal{Q}}}{m(\sqrt{\mathcal{Q}}+1)^2}, \end{aligned}$$

we then conclude from (5.46) and Lemma 9.a) that

$$\begin{aligned}\frac{\mathbb{E}[\|x^k-x^*\|_2^2]}{\|x^0-x^*\|_2^2} &\geq \frac{1}{2}\left[1-\frac{4\sqrt{\mathcal{Q}}}{m(\sqrt{\mathcal{Q}}+1)^2}\right]^k = \frac{1}{2}\exp\left(k\log\left(1-\frac{4\sqrt{\mathcal{Q}}}{m(\sqrt{\mathcal{Q}}+1)^2}\right)\right)\\ &\geq \frac{1}{2}\exp\left(-\frac{4k\sqrt{\mathcal{Q}}}{m(\sqrt{\mathcal{Q}}+1)^2-4\sqrt{\mathcal{Q}}}\right).\end{aligned}$$

■

6 Concluding remarks

In this paper, we present a new class of optimal first-order methods, referred to as primal-dual gradient methods, for solving the finite-sum composite convex optimization problems given in the form of (1.1). The optimal convergence of this algorithm has been established based on the primal-dual optimality gap for the ergodic mean of iterates, i.e., $\bar{z}^k$, and the distance from the iterate x^k to the optimal solution x^*. We also develop a randomized primal-dual gradient method which needs to compute the gradient of only one randomly selected component f_i. The complexity bounds of the randomized primal-dual gradient method have been established in terms of the distance from the iterate x^k to the optimal solution, and the primal optimality gap based on the ergodic mean of iterates, i.e., $\mathbb{E}[\Psi(\bar{x}^k)-\Psi^*]$. We show that these bounds are not improvable when the dimension n is large enough by developing new lower complexity bounds for randomized incremental gradient methods. Extensions of the randomized primal-dual gradient method to non-strongly convex, nonsmooth, and unbounded problems are also discussed in this paper. It should be noted that in this paper we focus on the theoretic convergence properties of these primal-dual gradient methods, and the algorithmic parameters were chosen in a conservative manner and were dependent on a few problem parameters, e.g., L and μ. In the future, it will be interesting to develop more adaptive versions of these algorithms which do not require the explicit estimation about L and μ.

References

1. A. Agarwal and L. Bottou. A Lower Bound for the Optimization of Finite Sums. *ArXiv e-prints*, Oct 2014.
2. A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16:697–725, 2006.
3. H.H. Bauschke, J.M. Borwein, and P.L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Controal and Optimization*, 42:596–636, 2003.
4. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2:183–202, 2009.
5. D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In S. Nowozin S. Sra and S. J. Wright, editors, *Optimization for Machine Learning*, pages 85–119. MIT Press, 2012.
6. L.M. Bregman. The relaxation method of finding the common point convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Phys.*, 7:200–217, 1967.
7. A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal-dual algorithm. Oct. 30, 2014.
8. A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40:120–145, 2011.
9. Caihua Chen, Bingsheng He, Yinyu Ye, and Xiaoming Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Optimization Online*, 2013.
10. Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.
11. C. Dang and G. Lan. Randomized first-order methods for saddle point optimization. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, September 2014.
12. A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances of Neural Information Processing Systems (NIPS)*, 27, 2014.
13. O. Fercoq and P. Richtárik. Smooth minimization of nonsmooth functions with parallel coordinate descent methods. *ArXiv e-prints*, Sep 2013.
14. S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. *SIAM Journal on Optimization*, 22:1469–1492, 2012.
15. S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23:2061–2089, 2013.
16. S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic optimization. Technical report, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, June 2013.

17. R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances of Neural Information Processing Systems (NIPS)*, 26:315–323, 2013.
18. K.C. Kiwiel. Proximal minimization methods with generalized bregman functions. *SIAM Journal on Controal and Optimization*, 35:1142–1168, 1997.
19. G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.
20. G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126:1–29, 2011.
21. G. Lan, A. S. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming*, 134:425–458, 2012.
22. H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. Technical report, 2015. hal-01160728.
23. Q. Lin, Z. Lu, and Lin Xiao. An accelerated proximal coordinate gradient method and its application to regularized empirical risk minimization. Technical report, 2014. no. MSR-TR-2014-94.
24. A. S. Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15:229–251, 2005.
25. A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.
26. A. S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.
27. Y. E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.
28. Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.
29. Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.
30. Y. E. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Feburary 2010.
31. Y. E. Nesterov. Gradient methods for minimizing composite objective functions. *Mathematical Programming., Series B*, 140:125–161, 2013.
32. M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. Technical report, September 2013.
33. S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567599, 2013.
34. S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 2015. to appear.
35. P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, University of Washington, Seattle, May 2008.
36. Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. Manuscript, September 2014.