SIAM J. Optim.
Vol. 25, No. 4, pp. 2244–2273

AN ACCELERATED RANDOMIZED PROXIMAL COORDINATE GRADIENT METHOD AND ITS APPLICATION TO REGULARIZED EMPIRICAL RISK MINIMIZATION[*]

QIHANG LIN[†], ZHAOSONG LU[‡], AND LIN XIAO[§]

Abstract. We consider the problem of minimizing the sum of two convex functions: one is smooth and given by a gradient oracle, and the other is separable over blocks of coordinates and has a simple known structure over each block. We develop an accelerated randomized proximal coordinate gradient (APCG) method for minimizing such convex composite functions. For strongly convex functions, our method achieves faster linear convergence rates than existing randomized proximal coordinate gradient methods. Without strong convexity, our method enjoys accelerated sublinear convergence rates. We show how to apply the APCG method to solve the regularized empirical risk minimization (ERM) problem and devise efficient implementations that avoid full-dimensional vector operations. For ill-conditioned ERM problems, our method obtains improved convergence rates than the state-of-the-art stochastic dual coordinate ascent method.

Key words. convex optimization, coordinate descent method, randomized algorithm, accelerated proximal gradient method, empirical risk minimization

AMS subject classifications. 65C60, 65Y20, 90C25, 68W20, 90C06

DOI. 10.1137/141000270

1. Introduction.

Coordinate descent methods have received extensive attention in recent years due to their potential for solving large-scale optimization problems arising from machine learning and other applications (e.g., [29, 10, 47, 17, 45, 30]). In this paper, we develop an accelerated proximal coordinate gradient (APCG) method for solving problems of the following form:

$$\underset{x\in\mathbb{R}^N}{\text{minimize}} \quad \left\{F(x) \overset{\text{def}}{=} f(x) + \Psi(x)\right\}, \tag{1.1}$$

where f and Ψ are proper and lower semicontinuous convex functions [34, section 7]. We assume that f is differentiable on $\mathbb{R}^N$ and Ψ has a block separable structure, i.e.,

$$\Psi(x) = \sum_{i=1}^{n} \Psi_i(x_i), \tag{1.2}$$

where each x_i denotes a subvector of x with cardinality N_i, and the collection $\{x_i : i = 1, \ldots, n\}$ forms a partition of the components of x. In addition to the capability of modeling nonsmooth terms such as $\Psi(x) = \lambda\|x\|_1$, this model includes optimization problems with block separable constraints. More specifically, each block constraint $x_i \in C_i$, where C_i is a closed convex set, can be modeled by an indicator function defined as $\Psi_i(x_i) = 0$ if $x_i \in C_i$ and ∞ otherwise.

[*] Received by the editors December 16, 2014; accepted for publication (in revised form) September 24, 2015; published electronically November 10, 2015. An extended abstract of this paper appeared in *Advances in Neural Information Processing Systems* 27, Montreal, Canada, 2014.
http://www.siam.org/journals/siopt/25-4/100027.html

[†] Tippie College of Business, University of Iowa, Iowa City, IA 52242 (qihang-lin@uiowa.edu).

[‡] Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada (zhaosong@sfu.ca).

[§] Machine Learning Groups, Microsoft Research, Redmond, WA 98052 (lin.xiao@microsoft.com).

At each iteration, coordinate descent methods choose one block of coordinates x_i to sufficiently reduce the objective value while keeping other blocks fixed. In order to exploit the known structure of each Ψ_i, a *proximal* coordinate gradient step can be taken [33]. To be more specific, given the current iterate $x^{(k)}$, we pick a block $i_k \in \{1, \ldots, n\}$ and solve a blockwise proximal subproblem in the form of

$$h_{i_k}^{(k)} = \operatorname*{arg\,min}_{h \in \Re^{N_{i_k}}} \left\{ \langle \nabla_{i_k} f(x^{(k)}), h \rangle + \frac{L_{i_k}}{2} \|h\|^2 + \Psi_{i_k}(x_{i_k}^{(k)} + h) \right\} \tag{1.3}$$

and then set the next iterate as

$$x_i^{(k+1)} = \begin{cases} x_{i_k}^{(k)} + h_{i_k}^{(k)} & \text{if } i = i_k, \\ x_i^{(k)} & \text{if } i \neq i_k, \end{cases} \qquad i = 1, \ldots, n. \tag{1.4}$$

Here $\nabla_i f(x)$ denotes the *partial gradient* of f with respect to x_i, and L_i is the Lipschitz constant of the partial gradient (which will be defined precisely later).

One common approach for choosing such a block is the *cyclic* scheme. The global and local convergence properties of the cyclic coordinate descent method have been studied in, e.g., [41, 22, 36, 2, 9]. Recently, *randomized* strategies for choosing the block to update became more popular [38, 15, 26, 33]. In addition to its theoretical benefits (randomized schemes are in general easier to analyze than the cyclic scheme), numerous experiments have demonstrated that randomized coordinate descent methods are very powerful for solving large-scale machine learning problems [6, 10, 38, 40]. Their efficiency can be further improved with parallel and distributed implementations [5, 31, 32, 23, 19]. Randomized block coordinate descent methods have also been proposed and analyzed for solving problems with coupled linear constraints [43, 24] and a class of structured nonconvex optimization problems (e.g., [20, 28]). Coordinate descent methods with more general schemes of choosing the block to update have also been studied; see, e.g., [3, 44, 46].

Inspired by the success of accelerated full gradient (AFG) methods [25, 1, 42, 27], several recent works extended Nesterov's acceleration technique to speed up randomized coordinate descent methods. In particular, Nesterov [26] developed an accelerated randomized coordinate gradient method for minimizing unconstrained smooth functions, which corresponds to the case of $\Psi(x) \equiv 0$ in (1.1). Lu and Xiao [21] gave a sharper convergence analysis of Nesterov's method using a randomized estimate sequence framework, and Lee and Sidford [14] developed extensions using weighted random sampling schemes. Accelerated coordinate gradient methods have also been used to speed up the solution of linear systems [14, 18]. But these work are all restricted to the case of unconstrained smooth minimization.

Extending accelerated coordinate gradient methods to the more general composite minimization problem in (1.1) appeared to be more challenging than extending the nonaccelerated versions as done in [33]. The key difficulty lies in handling the nonsmooth terms $\Psi_i(x_i)$ *coordinatewise* in an accelerated framework. More recently, Fercoq and Richtárik [8] made important progress by proposing an APPROX (accelerated, parallel, and proximal) coordinate descent method for solving the more general composite minimization problem (1.1) and obtained an accelerated sublinear convergence rate. But their method cannot exploit the strong convexity of the objective function to obtain accelerated linear rates in the composite case.

In this paper, we propose an APCG method that achieves accelerated linear convergence rates when the composite objective function is strongly convex. Without the strong convexity assumption, our method recovers a special case of the APPROX method [8]. Moreover, we show how to apply the APCG method to solve the regularized empirical risk minimization (ERM) problem and devise efficient implementations that avoid full-dimensional vector operations. For ill-conditioned ERM problems, our method obtains improved convergence rates over the state-of-the-art stochastic dual coordinate ascent (SDCA) method [40].

1.1. Outline of paper. This paper is organized as follows. The rest of this section introduces some notation and states our main assumptions. In section 2, we present the general APCG method and our main theorem on its convergence rate. We also give two simplified versions of APCG depending on whether the function f is strongly convex and explain how to exploit strong convexity in Ψ. Section 3 is devoted to the convergence analysis that proves our main theorem. In section 4, we derive equivalent implementations of the APCG method that can avoid full-dimensional vector operations.

In section 5, we apply the APCG method to solve the dual of the regularized ERM problem and give the corresponding complexity results. We also explain how to recover primal solutions to guarantee the same rate of convergence for the primal-dual gap. In addition, we present numerical experiments to demonstrate the performance of the APCG method.

1.2. Notation and assumptions. For any partition of $x \in \mathbb{R}^N$ into $\{x_i \in \mathbb{R}^{N_i} : i = 1, \ldots, n\}$ with $\sum_{i=1}^n N_i = N$, there is an $N \times N$ permutation matrix U partitioned as $U = [U_1 \cdots U_n]$, where $U_i \in \mathbb{R}^{N \times N_i}$, such that

$$x = \sum_{i=1}^{n} U_i x_i \quad \text{and} \quad x_i = U_i^T x, \quad i = 1, \ldots, n.$$

For any $x \in \mathbb{R}^N$, the *partial gradient* of f with respect to x_i is defined as

$$\nabla_i f(x) = U_i^T \nabla f(x), \quad i = 1, \ldots, n.$$

We associate each subspace $\mathbb{R}^{N_i}$, for $i = 1, \ldots, n$, with the standard Euclidean norm, denoted $\|\cdot\|_2$. We make the following assumptions, which are standard in the literature on coordinate descent methods (e.g., [26, 33]).

Assumption 1. The gradient of the function f is blockwise Lipschitz continuous with constants L_i, i.e.,

$$\|\nabla_i f(x + U_i h_i) - \nabla_i f(x)\|_2 \leq L_i \|h_i\|_2 \quad \forall\, h_i \in \mathbb{R}^{N_i}, \quad i = 1, \ldots, n, \quad x \in \mathbb{R}^N.$$

An immediate consequence of Assumption 1 is (see, e.g., [25, Lemma 1.2.3])

$$f(x + U_i h_i) \leq f(x) + \langle \nabla_i f(x), h_i \rangle + \frac{L_i}{2} \|h_i\|_2^2 \quad \forall\, x \in \mathbb{R}^N, \quad h_i \in \mathbb{R}^{N_i}, \tag{1.5}$$

for $i = 1, \ldots, n$. For convenience, we define a weighted norm in the whole space $\mathbb{R}^N$:

$$\|x\|_L = \left(\sum_{i=1}^{n} L_i \|x_i\|_2^2 \right)^{1/2} \quad \forall\, x \in \mathbb{R}^N. \tag{1.6}$$

Assumption 2. There exists $\mu \geq 0$ such that for all $y \in \mathbb{R}^N$ and $x \in \operatorname{dom}(\Psi)$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_L^2.$$

The *convexity parameter* of f with respect to the norm $\|\cdot\|_L$ is the largest μ such that the above inequality holds. Every convex function satisfies Assumption 2 with $\mu = 0$. If $\mu > 0$, then the function f is called *strongly* convex.

Remark. Together with (1.5) and the definition of $\|\cdot\|_L$ in (1.6), Assumption 2 implies $\mu \le 1$.

2. The APCG method.

In this section we describe the general APCG method (Algorithm 1) and its two simplified versions under different assumptions (whether or not the objective function is strongly convex). This algorithm can be viewed as a generalization of Nesterov's accelerated gradient method [25] which simultaneously covers the cases of block coordinate descent and composite minimization. In particular, if $n = 1$ (full gradient) and $\Psi(x) \equiv 0$, then it can be shown that Algorithm 1 is equivalent to Algorithm (2.2.8) in [25]. However, there are important differences that are not obvious to derive in the generalization; for example, here the proximal mapping appears in the update of $z^{(k)}$, instead of $x^{(k)}$ as done in Algorithm (2.2.19) of [25]. We derived this method using the framework of *randomized estimate sequence* developed in [21]. The convergence analysis given in section 3 is the result of further simplification, which does not rely on randomized estimate sequence.

We first explain the notation used in Algorithm 1. The algorithm proceeds in iterations, with k being the iteration counter. Lowercase letters x, y, z represent vectors in the full space $\mathbb{R}^N$, and $x^{(k)}$, $y^{(k)}$, and $z^{(k)}$ are their values at the kth iteration. Each block coordinate is indicated with a subscript, for example, $x_i^{(k)}$ represent the value of the ith block of the vector $x^{(k)}$. The Greek letters α, β, γ

```
ALGORITHM 1. THE APCG METHOD.
input: $x^{(0)} \in \mathrm{dom}\,(\Psi)$ and convexity parameter $\mu \ge 0$.
initialize: set $z^{(0)} = x^{(0)}$ and choose $0 < \gamma_0 \in [\mu, 1]$.
iterate: repeat for $k = 0, 1, 2, \ldots$
  1. Compute $\alpha_k \in (0, \frac{1}{n}]$ from the equation
       (2.1)   $n^2\alpha_k^2 = (1-\alpha_k)\,\gamma_k + \alpha_k\mu,$
     and set
       (2.2)   $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu, \qquad \beta_k = \frac{\alpha_k\mu}{\gamma_{k+1}}.$
  2. Compute $y^{(k)}$ as
       (2.3)   $y^{(k)} = \frac{1}{\alpha_k\gamma_k + \gamma_{k+1}}\Big(\alpha_k\gamma_k z^{(k)} + \gamma_{k+1}x^{(k)}\Big).$
  3. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random and compute
       $z^{(k+1)} = \operatorname*{arg\,min}_{x\in\mathbb{R}^N}\Big\{\frac{n\alpha_k}{2}\big\|x-(1-\beta_k)z^{(k)}-\beta_k y^{(k)}\big\|_L^2 + \langle\nabla_{i_k}f(y^{(k)}), x_{i_k}\rangle + \Psi_{i_k}(x_{i_k})\Big\}.$
  4. Set
       (2.4)   $x^{(k+1)} = y^{(k)} + n\alpha_k(z^{(k+1)} - z^{(k)}) + \frac{\mu}{n}(z^{(k)} - y^{(k)}).$
```

are scalars, and α_k, β_k, and γ_k represent their values at iteration k. For scalars, a superscript represents the power exponent; for example, n^2, α_k^2 denotes the squares of n and α_k, respectively.

At each iteration k, the APCG method picks a random coordinate $i_k \in \{1, \ldots, n\}$ and generates $y^{(k)}$, $x^{(k+1)}$, and $z^{(k+1)}$. One can observe that $x^{(k+1)}$, and $z^{(k+1)}$ depend on the realization of the random variable

$$\xi_k = \{i_0, i_1, \ldots, i_k\},$$

while $y^{(k)}$ is independent of i_k and only depends on ξ_{k-1}.

To better understand this method, we make the following observations. For convenience, let

$$\tilde{z}^{(k+1)} = \arg\min_{x\in\mathbb{R}^N}\Big\{\frac{n\alpha_k}{2}\big\|x-(1-\beta_k)z^{(k)}-\beta_k y^{(k)}\big\|_L^2+\langle\nabla f(y^{(k)}), x-y^{(k)}\rangle+\Psi(x)\Big\}, \tag{2.5}$$

which is a full-dimensional update version of step 3. One can observe that $z^{(k+1)}$ is updated as

$$z_i^{(k+1)} = \begin{cases} \tilde{z}_i^{(k+1)} & \text{if } i = i_k, \\ (1-\beta_k)z_i^{(k)} + \beta_k y_i^{(k)} & \text{if } i \neq i_k. \end{cases} \tag{2.6}$$

Notice that from (2.1), (2.2), (2.3), and (2.4) we have

$$x^{(k+1)} = y^{(k)} + n\alpha_k\left(z^{(k+1)} - (1-\beta_k)z^{(k)} - \beta_k y^{(k)}\right),$$

which together with (2.6) yields

$$x_i^{(k+1)} = \begin{cases} y_i^{(k)} + n\alpha_k\left(z_i^{(k+1)} - z_i^{(k)}\right) + \frac{\mu}{n}\left(z_i^{(k)} - y_i^{(k)}\right) & \text{if } i = i_k, \\ y_i^{(k)} & \text{if } i \neq i_k. \end{cases} \tag{2.7}$$

That is, in step 4, we only need to update the block coordinates $x_{i_k}^{(k+1)}$ as in (2.7) and set the rest to be $y_i^{(k)}$.

We now state our main result on the convergence rate of the APCG method, concerning the expected values of the optimality gap. The proof of the following theorem is given in section 3.

THEOREM 2.1. *Suppose Assumptions* 1 *and* 2 *hold. Let $F^\star$ be the optimal value of problem* (1.1)*, and let $\{x^{(k)}\}$ be the sequence generated by the APCG method. Then, for any $k \geq 0$, there holds*

$$\mathbf{E}_{\xi_{k-1}}[F(x^{(k)})] - F^\star \leq \min\left\{\left(1-\frac{\sqrt{\mu}}{n}\right)^k, \left(\frac{2n}{2n+k\sqrt{\gamma_0}}\right)^2\right\}\left(F(x^{(0)}) - F^\star + \frac{\gamma_0}{2}R_0^2\right),$$

where

$$R_0 \overset{\text{def}}{=} \min_{x^\star\in X^\star} \|x^{(0)} - x^\star\|_L, \tag{2.8}$$

and $X^\star$ is the set of optimal solutions of problem (1.1).

For $n = 1$, our results in Theorem 2.1 match exactly the convergence rates of the AFG method in [25, section 2.2]. For $n > 1$, our results improve upon the convergence rates of the randomized proximal coordinate gradient method described in (1.3) and (1.4). More specifically, if the block index $i_k \in \{1, \ldots, n\}$ is chosen uniformly at random, then the analysis in [33, 21] states that the convergence rate of (1.3) and (1.4) is on the order of

$$O\left(\min\left\{\left(1-\frac{\mu}{n}\right)^k, \ \frac{n}{n+k}\right\}\right).$$

Thus we obtain both the accelerated linear rate for strongly convex functions ($\mu > 0$) and the accelerated sublinear rate for non-strongly convex functions ($\mu = 0$). To the best of our knowledge, this is the first time that such an accelerated linear convergence rate is obtained for solving the general class of problems (1.1) using the coordinate descent type of methods.

2.1. Two special cases. For the strongly convex case with $\mu > 0$, we can initialize Algorithm 1 with the parameter $\gamma_0 = \mu$, which implies $\gamma_k = \mu$ and $\alpha_k = \beta_k = \sqrt{\mu}/n$ for all $k \geq 0$. This results in Algorithm 2. As a direct corollary of Theorem 2.1, Algorithm 2 enjoys an accelerated linear convergence rate:

$$\mathbf{E}_{\xi_{k-1}}[F(x^{(k)})] - F^\star \ \leq \ \left(1 - \frac{\sqrt{\mu}}{n}\right)^k \left(F(x^{(0)}) - F^\star + \frac{\mu}{2}\|x^{(0)} - x^\star\|_L^2\right),$$

where $x^\star$ is the unique solution of (1.1) under the strong convexity assumption.

Algorithm 3 shows the simplified version for $\mu = 0$, which can be applied to problems without strong convexity, or if the convexity parameter μ is unknown. According to Theorem 2.1, Algorithm 3 has an accelerated sublinear convergence rate, that is,

$$\mathbf{E}_{\xi_{k-1}}[F(x^{(k)})] - F^\star \ \leq \ \left(\frac{2n}{2n + kn\alpha_{-1}}\right)^2 \left(F(x^{(0)}) - F^\star + \frac{(n\alpha_{-1})^2}{2}R_0^2\right).$$

With the choice of $\alpha_{-1} = 1/\sqrt{n^2-1}$, which implies $\alpha_0 = 1/n$, Algorithm 3 reduces to the APPROX method [8] with single block update at each iteration (i.e., $\tau = 1$ in their Algorithm 1).

2.2. Exploiting strong convexity in Ψ. In this section we consider problem (1.1) with strongly convex Ψ. We assume that f and Ψ have convexity parameters $\mu_f \geq 0$ and $\mu_\Psi > 0$, both with respect to the standard Euclidean norm, denoted $\|\cdot\|_2$.

Algorithm 2. APCG with $\gamma_0 = \mu > 0$.

```
input: $x^{(0)} \in \mathrm{dom}\,(\Psi)$ and convexity parameter $\mu > 0$.
initialize: set $z^{(0)} = x^{(0)}$ and $\alpha = \frac{\sqrt{\mu}}{n}$.
iterate: repeat for $k = 0, 1, 2, \ldots$ and repeat for $k = 0, 1, 2, \ldots$
  1. Compute $y^{(k)} = \frac{x^{(k)} + \alpha z^{(k)}}{1+\alpha}$.
  2. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random and compute
       $z^{(k+1)} = \arg\min_{x \in \mathbb{R}^N} \Big\{ \frac{n\alpha}{2} \big\| x - (1-\alpha) z^{(k)} - \alpha y^{(k)} \big\|_L^2 + \langle \nabla_{i_k} f(y^{(k)}), x_{i_k} \rangle + \Psi_{i_k}(x_{i_k}) \Big\}$.
  3. Set $x^{(k+1)} = y^{(k)} + n\alpha(z^{(k+1)} - z^{(k)}) + n\alpha^2(z^{(k)} - y^{(k)})$.
```

Let $x^{(0)} \in \operatorname{dom}(\Psi)$ and $s^{(0)} \in \partial\Psi(x^{(0)})$ be arbitrarily chosen, and define two functions:

$$\tilde{f}(x) \overset{\text{def}}{=} f(x) + \Psi(x^{(0)}) + \langle s^{(0)}, x - x^{(0)}\rangle + \frac{\mu_\Psi}{2}\|x - x^{(0)}\|_2^2,$$
$$\tilde{\Psi}(x) \overset{\text{def}}{=} \Psi(x) - \Psi(x^{(0)}) - \langle s^{(0)}, x - x^{(0)}\rangle - \frac{\mu_\Psi}{2}\|x - x^{(0)}\|_2^2.$$

One can observe that the gradient of the function $\tilde{f}$ is blockwise Lipschitz continuous with constants $\tilde{L}_i = L_i + \mu_\Psi$ with respect to the norm $\|\cdot\|_2$. The convexity parameter of $\tilde{f}$ with respect to the norm $\|\cdot\|_{\tilde{L}}$ defined in (1.6) is

$$\mu := \frac{\mu_f + \mu_\Psi}{\max\limits_{1\le i\le n}\{L_i + \mu_\Psi\}}. \tag{2.9}$$

In addition, $\tilde{\Psi}$ is a block separable convex function which can be expressed as $\tilde{\Psi}(x) = \sum_{i=1}^n \tilde{\Psi}_i(x_i)$, where

$$\tilde{\Psi}_i(x_i) \;=\; \Psi_i(x_i) - \Psi_i(x_i^0) - \langle s_i^0, x_i - x_i^0\rangle - \frac{\mu_\Psi}{2}\|x_i - x_i^0\|_2^2, \qquad i = 1, \ldots, n.$$

As a result of the above definitions, we see that problem (1.1) is equivalent to

$$\operatorname*{minimize}_{x\in\Re^N} \left\{\tilde{f}(x) + \tilde{\Psi}(x)\right\}, \tag{2.10}$$

which can be suitably solved by the APCG method proposed in section 2 with f, Ψ_i, and L_i replaced by $\tilde{f}$, $\tilde{\Psi}_i$, and $L_i + \mu_\Psi$, respectively. The rate of convergence of APCG applied to problem (2.10) directly follows from Theorem 2.1, with μ given in (2.9) and the norm $\|\cdot\|_L$ in (2.8) replaced by $\|\cdot\|_{\tilde{L}}$.

3. Convergence analysis.

In this section, we prove Theorem 2.1. First we establish some useful properties of the sequences $\{\alpha_k\}_{k=0}^\infty$ and $\{\gamma_k\}_{k=0}^\infty$ generated in Algorithm 1. Then in section 3.1, we construct a sequence $\{\hat{\Psi}_k\}_{k=1}^\infty$ to bound the values of $\Psi(x^{(k)})$ and prove a useful property of the sequence. Finally we finish the proof of Theorem 2.1 in section 3.2.

LEMMA 3.1. *Suppose $\gamma_0 > 0$ and $\gamma_0 \in [\mu, 1]$ and $\{\alpha_k\}_{k=0}^\infty$ and $\{\gamma_k\}_{k=0}^\infty$ are generated in Algorithm 1. Then the following hold:*

```
ALGORITHM 3. APCG WITH $\mu = 0$.
input: $x^{(0)} \in \operatorname{dom}(\Psi)$.
initialize: set $z^{(0)} = x^{(0)}$ and choose $\alpha_{-1} \in (0, \frac{1}{n}]$.
iterate: repeat for $k = 0, 1, 2, \ldots$
  1. Compute $\alpha_k = \frac{1}{2}\left(\sqrt{\alpha_{k-1}^4 + 4\alpha_{k-1}^2} - \alpha_{k-1}^2\right)$.
  2. Compute $y^{(k)} = (1-\alpha_k)x^{(k)} + \alpha_k z^{(k)}$.
  3. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random and compute
       $z_{i_k}^{(k+1)} = \arg\min_{x\in\mathbb{R}^{N_{i_k}}}\Big\{\frac{n\alpha_k L_{i_k}}{2}\|x - z_{i_k}^{(k)}\|_2^2 + \langle\nabla_{i_k} f(y^{(k)}), x - y_{i_k}^{(k)}\rangle + \Psi_{i_k}(x)\Big\}$.
     and set $z_i^{(k+1)} = z_i^{(k)}$ for all $i \neq i_k$.
  4. Set $x^{(k+1)} = y^{(k)} + n\alpha_k(z^{(k+1)} - z^{(k)})$.
```

(i) *$\{\alpha_k\}_{k=0}^\infty$ and $\{\gamma_k\}_{k=0}^\infty$ are well-defined positive sequences.*
(ii) *$\sqrt{\mu}/n \le \alpha_k \le 1/n$ and $\mu \le \gamma_k \le 1$ for all $k \ge 0$.*
(iii) *$\{\alpha_k\}_{k=0}^\infty$ and $\{\gamma_k\}_{k=0}^\infty$ are nonincreasing.*
(iv) *$\gamma_k = n^2\alpha_{k-1}^2$ for all $k \ge 1$.*
(v) *With the definition of*

$$\lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i), \tag{3.1}$$

we have for all $k \ge 0$,

$$\lambda_k \;\le\; \min\left\{\left(1-\frac{\sqrt{\mu}}{n}\right)^k,\ \left(\frac{2n}{2n+k\sqrt{\gamma_0}}\right)^2\right\}.$$

Proof. Due to (2.1) and (2.2), statement (iv) always holds provided that $\{\alpha_k\}_{k=0}^\infty$ and $\{\gamma_k\}_{k=0}^\infty$ are well-defined. We now prove statements (i) and (ii) by induction. For convenience, let

$$g_\gamma(t) = n^2t^2 - \gamma(1-t) - \mu t.$$

Since $\mu \le 1$ and $\gamma_0 \in (0,1]$, one can observe that $g_{\gamma_0}(0) = -\gamma_0 < 0$ and

$$g_{\gamma_0}\left(\frac{1}{n}\right) = 1 - \gamma_0\left(1-\frac{1}{n}\right) - \frac{\mu}{n} \ge 1-\gamma_0 \ge 0.$$

These together with continuity of g_{γ_0} imply that there exists $\alpha_0 \in (0, 1/n]$ such that $g_{\gamma_0}(\alpha_0) = 0$, that is, α_0 satisfies (2.1) and is thus well-defined. In addition, by statement (iv) and $\gamma_0 \ge \mu$, one can see $\alpha_0 \ge \sqrt{\mu}/n$. Therefore, statements (i) and (ii) hold for $k = 0$.

Suppose that statements (i) and (ii) hold for some $k \ge 0$, that is, $\alpha_k, \gamma_k > 0$, $\sqrt{\mu}/n \le \alpha_k \le 1/n$, and $\mu \le \gamma_k \le 1$. Using these relations and (2.2), one can see that γ_{k+1} is well-defined and moreover $\mu \le \gamma_{k+1} \le 1$. In addition, we have $\gamma_{k+1} > 0$ due to statement (iv) and $\alpha_k > 0$. Using the fact $\mu \le 1$ (see the remark after Assumption 2), $\gamma_0 \in (0,1]$, and a similar argument as above, we obtain $g_{\gamma_k}(0) < 0$ and $g_{\gamma_k}(1/n) \ge 0$, which along with continuity of g_{γ_k} imply that there exists $\alpha_{k+1} \in (0, 1/n]$ such that $g_{\gamma_k}(\alpha_{k+1}) = 0$, that is, α_{k+1} satisfies (2.1) and is thus well-defined. By statement (iv) and $\gamma_{k+1} \ge \mu$, one can see that $\alpha_{k+1} \ge \sqrt{\mu}/n$. This completes the induction and hence statements (i) and (ii) hold.

Next, we show statement (iii) holds. Indeed, it follows from (2.2) that

$$\gamma_{k+1} - \gamma_k = \alpha_k(\mu - \gamma_k),$$

which together with $\gamma_k \ge \mu$ and $\alpha_k > 0$ implies that $\gamma_{k+1} \le \gamma_k$ and hence $\{\gamma_k\}_{k=0}^\infty$ is nonincreasing. Notice from statement (iv) and $\alpha_k > 0$ that $\alpha_k = \sqrt{\gamma_{k+1}}/n$. It follows that $\{\alpha_k\}_{k=0}^\infty$ is also nonincreasing.

Statement (v) can be proved by using the same arguments in the proof of [25, Lemma 2.2.4], and the details can be found in [21, section 4.2]. ▯

3.1. Construction and properties of $\hat{\Psi}_k$.

Motivated by [8], we give an explicit expression of $x^{(k)}$ as a convex combination of the vectors $z^{(0)}, \dots, z^{(k)}$ and use the coefficients to construct a sequence $\{\hat{\Psi}_k\}_{k=1}^\infty$ to bound $\Psi(x^{(k)})$.

LEMMA 3.2. *Let the sequences $\{\alpha_k\}_{k=0}^\infty$, $\{\gamma_k\}_{k=0}^\infty$, $\{x^{(k)}\}_{k=0}^\infty$, and $\{z^{(k)}\}_{k=0}^\infty$ be generated by Algorithm 1. Then each $x^{(k)}$ is a convex combination of $z^{(0)}, \dots, z^{(k)}$.*

More specifically, for all $k \geq 0$,

$$x^{(k)} = \sum_{l=0}^{k} \theta_l^{(k)} z^{(l)}, \tag{3.2}$$

where the constants $\theta_0^{(k)}, \ldots, \theta_k^{(k)}$ are nonnegative and sum to 1. Moreover, these constants can be obtained recursively by setting $\theta_0^{(0)} = 1$, $\theta_0^{(1)} = 1 - n\alpha_0$, $\theta_1^{(1)} = n\alpha_0$ and for $k \geq 1$,

$$\theta_l^{(k+1)} = \begin{cases} n\alpha_k, & l = k+1, \\ \left(1 - \frac{\mu}{n}\right) \frac{\alpha_k\gamma_k + n\alpha_{k-1}\gamma_{k+1}}{\alpha_k\gamma_k + \gamma_{k+1}} - \frac{(1-\alpha_k)\gamma_k}{n\alpha_k}, & l = k, \\ \left(1 - \frac{\mu}{n}\right) \frac{\gamma_{k+1}}{\alpha_k\gamma_k + \gamma_{k+1}} \theta_l^{(k)}, & l = 0, \ldots, k-1. \end{cases} \tag{3.3}$$

Proof. We prove the statements by induction. First, notice that $x^{(0)} = z^{(0)} = \theta_0^{(0)} z^{(0)}$. Using this relation and (2.3), we see that $y^{(0)} = z^{(0)}$. From (2.4) and $y^{(0)} = z^{(0)}$, we obtain

$$\begin{aligned} x^{(1)} &= y^{(0)} + n\alpha_0 \left(z^{(1)} - z^{(0)}\right) + \frac{\mu}{n}\left(z^{(0)} - y^{(0)}\right) \\ &= z^{(0)} + n\alpha_0 \left(z^{(1)} - z^{(0)}\right) \\ &= (1 - n\alpha_0) z^{(0)} + n\alpha_0 z^{(1)}. \end{aligned} \tag{3.4}$$

Since $\alpha_0 \in (0, 1/n]$ (Lemma 3.1(ii)), the vector $x^{(1)}$ is a convex combination of $z^{(0)}$ and $z^{(1)}$ with the coefficients $\theta_0^{(1)} = 1 - n\alpha_0$, $\theta_1^{(1)} = n\alpha_0$. For $k = 1$, substituting (2.3) into (2.4) yields

$$\begin{aligned} x^{(2)} &= y^{(1)} + n\alpha_1 \left(z^{(2)} - z^{(1)}\right) + \frac{\mu}{n}\left(z^{(1)} - y^{(1)}\right) \\ &= \left(1 - \frac{\mu}{n}\right) \frac{\gamma_2}{\alpha_1\gamma_1 + \gamma_2} x^{(1)} + \left[\left(1 - \frac{\mu}{n}\right) \frac{\alpha_1\gamma_1}{\alpha_1\gamma_1 + \gamma_2} - \frac{n^2\alpha_1^2 - \alpha_1\mu}{n\alpha_1}\right] z^{(1)} + n\alpha_1 z^{(2)}. \end{aligned}$$

Substituting (3.4) into the above equality, and using $(1 - \alpha_1)\gamma_1 = n^2\alpha_1^2 - \alpha_1\mu$ from (2.1), we get

$$x^{(2)} = \underbrace{\left(1 - \frac{\mu}{n}\right) \frac{\gamma_2(1 - n\alpha_0)}{\alpha_1\gamma_1 + \gamma_2}}_{\theta_0^{(2)}} z^{(0)} + \underbrace{\left[\left(1 - \frac{\mu}{n}\right) \frac{\alpha_1\gamma_1 + n\alpha_0\gamma_2}{\alpha_1\gamma_1 + \gamma_2} - \frac{(1 - \alpha_1)\gamma_1}{n\alpha_1}\right]}_{\theta_1^{(2)}} z^{(1)} + \underbrace{n\alpha_1}_{\theta_2^{(2)}} z^{(2)}. \tag{3.5}$$

From the definition of $\theta_1^{(2)}$ in the above equation, we observe that

$$\begin{aligned} \theta_1^{(2)} &= \left(1 - \frac{\mu}{n}\right) \frac{\alpha_1\gamma_1 + n\alpha_0\gamma_2}{\alpha_1\gamma_1 + \gamma_2} - \frac{(1 - \alpha_1)\gamma_1}{n\alpha_1} \\ &= \left(1 - \frac{\mu}{n}\right) \frac{\alpha_1\gamma_1(1 - n\alpha_0) + n\alpha_0(\alpha_1\gamma_1 + \gamma_2)}{\alpha_1\gamma_1 + \gamma_2} - \frac{n^2\alpha_1^2 - \alpha_1\mu}{n\alpha_1} \\ &= \frac{\alpha_1\gamma_1}{\alpha_1\gamma_1 + \gamma_2} \left(1 - \frac{\mu}{n}\right)(1 - n\alpha_0) + \left(1 - \frac{\mu}{n}\right) n\alpha_0 - n\alpha_1 + \frac{\mu}{n} \\ &= \frac{\alpha_1\gamma_1}{\alpha_1\gamma_1 + \gamma_2} \left(1 - \frac{\mu}{n}\right)(1 - n\alpha_0) + \left(1 - \frac{\mu}{n}\right) n(\alpha_0 - \alpha_1) + \frac{\mu}{n}(1 - n\alpha_1). \end{aligned}$$

From the above expression, and using the facts $\mu \le 1$, $\alpha_0 \ge \alpha_1$, $\gamma_k \ge 0$, and $0 \le \alpha_k \le 1/n$ (Lemma 3.1), we conclude that $\theta_1^{(2)} \ge 0$. Also considering the definitions of $\theta_0^{(2)}$ and $\theta_2^{(2)}$ in (3.5), we conclude that $\theta_l^{(2)} \ge 0$ for $0 \le l \le 2$. In addition, one can observe from (2.3), (2.4), and (3.4) that $x^{(1)}$ is an affine combination of $z^{(0)}$ and $z^{(1)}$, $y^{(1)}$ is an affine combination of $z^{(1)}$ and $x^{(1)}$, and $x^{(2)}$ is an affine combination of $y^{(1)}$, $z^{(1)}$, and $z^{(2)}$. It is known that substituting one affine combination into another yields a new affine combination. Hence, the combination given in (3.5) must be affine, which together with $\theta_l^{(2)} \ge 0$ for $0 \le l \le 2$ implies that it is also a convex combination.

Now suppose the recursion (3.3) holds for some $k \ge 1$. Substituting (2.3) into (2.4), we obtain that

$$x^{(k+1)} = \left(1-\frac{\mu}{n}\right)\frac{\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}}x^{(k)} + \left[\left(1-\frac{\mu}{n}\right)\frac{\alpha_k\gamma_k}{\alpha_k\gamma_k+\gamma_{k+1}} - \frac{(1-\alpha_k)\gamma_k}{n\alpha_k}\right]z^{(k)} + n\alpha_k z^{(k+1)}.$$

Further, substituting $x^{(k)} = n\alpha_{k-1}z^{(k)} + \sum_{l=0}^{k-1}\theta_l^{(k)}z^{(l)}$ (the induction hypothesis) into the above equation gives

$$\begin{aligned} x^{(k+1)} &= \sum_{l=0}^{k-1} \underbrace{\left(1-\frac{\mu}{n}\right)\frac{\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}}\theta_l^{(k)}}_{\theta_l^{(k+1)}} z^{(l)} \\ &\quad + \underbrace{\left[\left(1-\frac{\mu}{n}\right)\frac{\alpha_k\gamma_k+n\alpha_{k-1}\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}} - \frac{(1-\alpha_k)\gamma_k}{n\alpha_k}\right]}_{\theta_k^{(k+1)}} z^{(k)} + \underbrace{n\alpha_k}_{\theta_{k+1}^{(k+1)}} z^{(k+1)}. \end{aligned} \tag{3.6}$$

This gives the form of (3.2) and (3.3). In addition, by the induction hypothesis, $x^{(k)}$ is an affine combination of $z^{(0)}, \dots, z^{(k)}$. Also, notice from (2.3) and (2.4) that $y^{(k)}$ is an affine combination of $z^{(k)}$ and $x^{(k)}$, and $x^{(k+1)}$ is an affine combination of $y^{(k)}$, $z^{(k)}$, and $z^{(k+1)}$. Using these facts and a similar argument as for $x^{(2)}$, it follows that the combination (3.6) must be affine.

Finally, we claim $\theta_l^{(k+1)} \ge 0$ for all l. Indeed, we know from Lemma 3.1 that $\mu \le 1$, $\alpha_k \ge 0$, $\gamma_k \ge 0$. Also, $\theta_l^{(k)} \ge 0$ due to the induction hypothesis. It follows that $\theta_l^{(k+1)} \ge 0$ for all $l \ne k$. It remains to show that $\theta_k^{(k+1)} \ge 0$. To this end, we again use (2.1) to obtain $(1-\alpha_k)\gamma_k = n^2\alpha_k^2 - \alpha_k\mu$ and use (3.3) and a similar argument as for $\theta_1^{(2)}$ to rewrite $\theta_k^{(k+1)}$ as

$$\theta_k^{(k+1)} = \frac{\alpha_k\gamma_k}{\alpha_k\gamma_k+\gamma_{k+1}}\left(1-\frac{\mu}{n}\right)(1-n\alpha_{k-1}) + \left(1-\frac{\mu}{n}\right)n(\alpha_{k-1}-\alpha_k) + \frac{\mu}{n}(1-n\alpha_k).$$

Together with $\mu \le 1$, $0 \le \alpha_k \le 1/n$, $\gamma_k \ge 0$, and $\alpha_{k-1} \ge \alpha_k$, this implies that $\theta_k^{(k+1)} \ge 0$. Therefore, $x^{(k+1)}$ is a convex combination of $z^{(0)}, \dots, z^{(k+1)}$ with the coefficients given in (3.3). □

In the following lemma, we construct the sequence $\{\hat{\Psi}_k\}_{k=0}^{\infty}$ and prove a recursive inequality.

LEMMA 3.3. *Let $\hat{\Psi}_k$ denote the convex combination of $\Psi(z^{(0)}), \dots, \Psi(z^{(k)})$ using the same coefficients given in Lemma 3.2, i.e.,*

$$\hat{\Psi}_k = \sum_{l=0}^{k}\theta_l^{(k)}\Psi(z^{(l)}).$$

Then for all $k \geq 0$, we have $\Psi(x^{(k)}) \leq \hat{\Psi}_k$ and

$$\mathbf{E}_{i_k}[\hat{\Psi}_{k+1}] \;\leq\; \alpha_k \Psi(\tilde{z}^{(k+1)}) + (1-\alpha_k)\hat{\Psi}_k. \tag{3.7}$$

Proof. The first result $\Psi(x^{(k)}) \leq \hat{\Psi}_k$ follows directly from convexity of Ψ. We now prove (3.7). First we deal with the case $k = 0$. Using (2.6), (3.3), and the facts $y^{(0)} = z^{(0)}$ and $\hat{\Psi}_0 = \Psi(x^{(0)})$, we get

$$\begin{aligned}
\mathbf{E}_{i_0}[\hat{\Psi}_1] &= \mathbf{E}_{i_0}\left[n\alpha_0 \Psi(z^{(1)}) + (1-n\alpha_0)\Psi(z^{(0)})\right] \\
&= \mathbf{E}_{i_0}\left[n\alpha_0\Big(\Psi_{i_0}(\tilde{z}_{i_0}^{(1)}) + \textstyle\sum_{j\neq i_0} \Psi_j(z_j^0)\Big)\right] + (1-n\alpha_0)\Psi(z^{(0)}) \\
&= \alpha_0\Psi(\tilde{z}^{(1)}) + (n-1)\alpha_0\Psi(z^{(0)}) + (1-n\alpha_0)\Psi(x^{(0)}) \\
&= \alpha_0\Psi(\tilde{z}^{(1)}) + (1-\alpha_0)\Psi(x^{(0)}) \\
&= \alpha_0\Psi(\tilde{z}^{(1)}) + (1-\alpha_0)\hat{\Psi}_0.
\end{aligned}$$

For $k \geq 1$, we use (2.6) and the definition of β_k in (2.2) to obtain that

$$\begin{aligned}
\mathbf{E}_{i_k}\left[\Psi(z^{(k+1)})\right] &= \mathbf{E}_{i_k}\left[\Psi_{i_k}(z_{i_k}^{(k+1)}) + \sum_{j\neq i_k}\Psi_j(z_j^{(k+1)})\right] \\
&= \frac{1}{n}\Psi(\tilde{z}^{(k+1)}) + \left(1-\frac{1}{n}\right)\Psi\left(\frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}z^{(k)} + \frac{\alpha_k\mu}{\gamma_{k+1}}y^{(k)}\right). \qquad (3.8)
\end{aligned}$$

Using (2.2) and (2.3), one can observe that

$$\begin{aligned}
\frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}z^{(k)} + \frac{\alpha_k\mu}{\gamma_{k+1}}y^{(k)} &= \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}z^{(k)} + \frac{\alpha_k\mu}{\gamma_{k+1}(\alpha_k\gamma_k+\gamma_{k+1})}\left(\alpha_k\gamma_k z^{(k)} + \gamma_{k+1}x^{(k)}\right) \\
&= \left(1 - \frac{\alpha_k\mu}{\alpha_k\gamma_k+\gamma_{k+1}}\right)z^{(k)} + \frac{\alpha_k\mu}{\alpha_k\gamma_k+\gamma_{k+1}}x^{(k)}.
\end{aligned}$$

It follows from the above equation and convexity of Ψ that

$$\Psi\left(\frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}z^{(k)} + \frac{\alpha_k\mu}{\gamma_{k+1}}y^{(k)}\right) \leq \left(1 - \frac{\alpha_k\mu}{\alpha_k\gamma_k+\gamma_{k+1}}\right)\Psi(z^{(k)}) + \frac{\alpha_k\mu}{\alpha_k\gamma_k+\gamma_{k+1}}\Psi(x^{(k)}),$$

which together with (3.8) yields

$$\mathbf{E}_{i_k}\left[\Psi(z^{(k+1)})\right] \leq \frac{1}{n}\Psi(\tilde{z}^{(k+1)}) + \left(1-\frac{1}{n}\right)\left[\left(1 - \frac{\alpha_k\mu}{\alpha_k\gamma_k+\gamma_{k+1}}\right)\Psi(z^{(k)}) + \frac{\alpha_k\mu}{\alpha_k\gamma_k+\gamma_{k+1}}\Psi(x^{(k)})\right]. \tag{3.9}$$

In addition, from the definition of $\hat{\Psi}_k$ and $\theta_k^{(k)} = n\alpha_{k-1}$, we have

$$\sum_{l=0}^{k-1}\theta_l^{(k)}\Psi(z^{(l)}) = \hat{\Psi}_k - n\alpha_{k-1}\Psi(z^{(k)}). \tag{3.10}$$

Next, using the definition of $\hat{\Psi}_k$ and (3.3), we obtain

$$\begin{aligned}
\mathbf{E}_{i_k}[\hat{\Psi}_{k+1}] = n\alpha_k\mathbf{E}_{i_k}\left[\Psi(z^{(k+1)})\right] &+ \left[\left(1-\frac{\mu}{n}\right)\frac{\alpha_k\gamma_k + n\alpha_{k-1}\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}} - \frac{(1-\alpha_k)\gamma_k}{n\alpha_k}\right]\Psi(z^{(k)}) \\
&+ \left(1-\frac{\mu}{n}\right)\frac{\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}}\sum_{l=0}^{k-1}\theta_l^{(k)}\Psi(z^{(l)}). \qquad (3.11)
\end{aligned}$$

Plugging (3.9) and (3.10) into (3.11) yields

$$\begin{aligned}\mathbf{E}_{i_k}[\hat{\Psi}_{k+1}] \le{}& \alpha_k \Psi(\tilde{z}^{(k+1)}) + (n-1)\alpha_k \left[\left(1 - \frac{\alpha_k \mu}{\alpha_k\gamma_k + \gamma_{k+1}}\right)\Psi(z^{(k)}) + \frac{\alpha_k\mu}{\alpha_k\gamma_k+\gamma_{k+1}}\Psi(x^{(k)})\right] \\ &+ \left[\left(1-\frac{\mu}{n}\right)\frac{\alpha_k\gamma_k + n\alpha_{k-1}\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}} - \frac{(1-\alpha_k)\gamma_k}{n\alpha_k}\right]\Psi(z^{(k)}) \\ &+ \left(1-\frac{\mu}{n}\right)\frac{\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}}\left(\hat{\Psi}_k - n\alpha_{k-1}\Psi(z^{(k)})\right)\end{aligned} \tag{3.12}$$

$$\begin{aligned}\le{}& \alpha_k\Psi(\tilde{z}^{(k+1)}) + \underbrace{\frac{(n-1)\alpha_k^2\mu + \left(1-\frac{\mu}{n}\right)\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}}}_{\Gamma}\hat{\Psi}_k \\ &+ \underbrace{\left[(n-1)\alpha_k\left(1-\frac{\alpha_k\mu}{\alpha_k\gamma_k+\gamma_{k+1}}\right) + \left(1-\frac{\mu}{n}\right)\frac{\alpha_k\gamma_k}{\alpha_k\gamma_k+\gamma_{k+1}} - \frac{(1-\alpha_k)\gamma_k}{n\alpha_k}\right]}_{\Delta}\Psi(z^{(k)}),\end{aligned} \tag{3.13}$$

where the second inequality is due to $\Psi(x^{(k)}) \le \hat{\Psi}_k$. Notice that the right-hand side of (3.9) is an affine combination of $\Psi(\tilde{z}^{(k+1)})$, $\Psi(z^{(k)})$, and $\Psi(x^{(k)})$, and the right-hand side of (3.11) is an affine combination of $\Psi(z^{(0)}), \dots, \Psi(z^{(k+1)})$. In addition, all operations in (3.12) and (3.13) preserve the affine combination property. Using these facts, one can observe that the right-hand side of (3.13) is also an affine combination of $\Psi(\tilde{z}^{(k+1)})$, $\Psi(z^{(k)})$, and $\hat{\Psi}_k$, namely, $\alpha_k + \Delta + \Gamma = 1$, where Δ and Γ are defined above.

We next show that $\Gamma = 1 - \alpha_k$ and $\Delta = 0$. Indeed, notice that from (2.2) we have

$$\alpha_k\gamma_k + \gamma_{k+1} = \alpha_k\mu + \gamma_k. \tag{3.14}$$

Using this relation, $\gamma_{k+1} = n^2\alpha_k^2$ (Lemma 3.1(iv)), and the definition of Γ in (3.13), we get

$$\begin{aligned}\Gamma &= \frac{(n-1)\alpha_k^2\mu + \left(1-\frac{\mu}{n}\right)\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}} = \frac{(n-1)\alpha_k^2\mu + \gamma_{k+1} - \frac{\mu}{n}\gamma_{k+1}}{\alpha_k\gamma_k+\gamma_{k+1}} \\ &= \frac{(n-1)\alpha_k^2\mu + \gamma_{k+1} - \frac{\mu}{n}(n^2\alpha_k^2)}{\alpha_k\gamma_k+\gamma_{k+1}} = \frac{\gamma_{k+1} - \alpha_k^2\mu}{\alpha_k\gamma_k+\gamma_{k+1}} \\ &= 1 - \alpha_k\left(\frac{\alpha_k\mu+\gamma_k}{\alpha_k\gamma_k+\gamma_{k+1}}\right) = 1 - \alpha_k,\end{aligned}$$

where the last equality is due to (3.14). Finally, $\Delta = 0$ follows from $\Gamma = 1 - \alpha_k$ and $\alpha_k + \Delta + \Gamma = 1$. These together with the inequality (3.13) yield the desired result. □

3.2. Proof of Theorem 2.1. We are now ready to present a proof for Theorem 2.1. We note that the proof in this subsection can also be recast into the framework of randomized estimate sequence developed in [21, 14], but here we give a straightforward proof without using that machinery.

Dividing both sides of (2.1) by $n\alpha_k$ gives

$$n\alpha_k = \frac{(1-\alpha_k)\gamma_k}{n\alpha_k} + \frac{\mu}{n}. \tag{3.15}$$

Observe from (2.3) that

$$z^{(k)} - y^{(k)} = -\frac{\gamma_{k+1}}{\alpha_k \gamma_k}\left(x^{(k)} - y^{(k)}\right). \tag{3.16}$$

It follows from (2.4) and (3.15) that

$$\begin{aligned} x^{(k+1)} - y^{(k)} &= n\alpha_k z^{(k+1)} - \frac{(1-\alpha_k)\gamma_k}{n\alpha_k} z^{(k)} - \frac{\mu}{n} y^{(k)} \\ &= n\alpha_k z^{(k+1)} - \frac{(1-\alpha_k)\gamma_k}{n\alpha_k}(z^{(k)} - y^{(k)}) - \left(\frac{(1-\alpha_k)\gamma_k}{n\alpha_k} + \frac{\mu}{n}\right) y^{(k)}, \end{aligned}$$

which together with (3.15), (3.16), and $\gamma_{k+1} = n^2\alpha_k^2$ (Lemma 3.1(iv)) gives

$$\begin{aligned} x^{(k+1)} - y^{(k)} &= n\alpha_k z^{(k+1)} + \frac{(1-\alpha_k)\gamma_{k+1}}{n\alpha_k^2}\left(x^{(k)} - y^{(k)}\right) - n\alpha_k y^{(k)} \\ &= n\alpha_k z^{(k+1)} + n(1-\alpha_k)(x^{(k)} - y^{(k)}) - n\alpha_k y^{(k)} \\ &= n\left[\alpha_k(z^{(k+1)} - y^{(k)}) + (1-\alpha_k)(x^{(k)} - y^{(k)})\right]. \end{aligned}$$

Using this relation, (2.7), and Assumption 1, we have

$$\begin{aligned} f(x^{(k+1)}) &\le f(y^{(k)}) + \left\langle \nabla_{i_k} f(y^{(k)}),\ x_{i_k}^{(k+1)} - y_{i_k}^{(k)} \right\rangle + \ \frac{L_{i_k}}{2}\left\| x_{i_k}^{(k+1)} - y_{i_k}^{(k)} \right\|_2^2 \\ &= f(y^{(k)}) + n\left\langle \nabla_{i_k} f(y^{(k)}),\ \left[\alpha_k(z^{(k+1)} - y^{(k)}) + (1-\alpha_k)(x^{(k)} - y^{(k)})\right]_{i_k} \right\rangle \\ &\quad + \frac{n^2 L_{i_k}}{2}\left\| \left[\alpha_k(z^{(k+1)} - y^{(k)}) + (1-\alpha_k)(x^{(k)} - y^{(k)})\right]_{i_k} \right\|_2^2 \\ &= (1-\alpha_k)\left[f(y^{(k)}) + n\left\langle \nabla_{i_k} f(y^{(k)}), (x_{i_k}^{(k)} - y_{i_k}^{(k)}) \right\rangle\right] \\ &\quad + \alpha_k\left[f(y^{(k)}) + n\left\langle \nabla_{i_k} f(y^{(k)}), (z_{i_k}^{(k+1)} - y_{i_k}^{(k)}) \right\rangle\right] \\ &\quad + \frac{n^2 L_{i_k}}{2}\left\| \left[\alpha_k(z^{(k+1)} - y^{(k)}) + (1-\alpha_k)(x^{(k)} - y^{(k)})\right]_{i_k} \right\|_2^2. \end{aligned}$$

Taking expectation on both sides of the above inequality with respect to i_k, and noticing that $z_{i_k}^{(k+1)} = \tilde{z}_{i_k}^{(k+1)}$, we get

$$\begin{aligned} \mathbf{E}_{i_k}\left[f(x^{(k+1)})\right] &\le (1-\alpha_k)\left[f(y^{(k)}) + \left\langle \nabla f(y^{(k)}), (x^{(k)} - y^{(k)}) \right\rangle\right] \\ &\quad + \alpha_k\left[f(y^{(k)}) + \left\langle \nabla f(y^{(k)}), (\tilde{z}^{(k+1)} - y^{(k)}) \right\rangle\right] \\ &\quad + \frac{n}{2}\left\| \alpha_k(\tilde{z}^{(k+1)} - y^{(k)}) + (1-\alpha_k)(x^{(k)} - y^{(k)}) \right\|_L^2 \\ &\le (1-\alpha_k) f(x^{(k)}) + \alpha_k\left[f(y^{(k)}) + \left\langle \nabla f(y^{(k)}), (\tilde{z}^{(k+1)} - y^{(k)}) \right\rangle\right] \\ &\quad + \frac{n}{2}\left\| \alpha_k(\tilde{z}^{(k+1)} - y^{(k)}) + (1-\alpha_k)(x^{(k)} - y^{(k)}) \right\|_L^2, \end{aligned} \tag{3.17}$$

where the second inequality follows from the convexity of f.

In addition, by (2.2), (3.16), and $\gamma_{k+1} = n^2\alpha_k^2$ (Lemma 3.1(iv)), we have

$$\begin{aligned}
&\frac{n}{2}\left\|\alpha_k(\tilde{z}^{(k+1)} - y^{(k)}) + (1-\alpha_k)(x^{(k)} - y^{(k)})\right\|_L^2 \\
&= \frac{n}{2}\left\|\alpha_k(\tilde{z}^{(k+1)} - y^{(k)}) - \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}(z^{(k)} - y^{(k)})\right\|_L^2 \\
&= \frac{n\alpha_k^2}{2}\left\|\tilde{z}^{(k+1)} - y^{(k)} - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}(z^{(k)} - y^{(k)})\right\|_L^2 \\
&= \frac{\gamma_{k+1}}{2n}\left\|\tilde{z}^{(k+1)} - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}}y^{(k)}\right\|_L^2, \qquad (3.18)
\end{aligned}$$

where the first equality is due to (3.16), the third one is due to (2.1) and (2.2), and $\gamma_{k+1} = n^2\alpha_k^2$. This equation together with (3.17) yields

$$\begin{aligned}
\mathbf{E}_{i_k}\left[f(x^{(k+1)})\right] \le (1-\alpha_k)f(x^{(k)}) + \alpha_k\Bigg[& f(y^{(k)}) + \left\langle \nabla f(y^{(k)}), \tilde{z}^{(k+1)} - y^{(k)}\right\rangle \\
&+ \frac{\gamma_{k+1}}{2n\alpha_k}\left\|\tilde{z}^{(k+1)} - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}}y^{(k)}\right\|_L^2\Bigg].
\end{aligned}$$

Using Lemma 3.3, we have

$$\mathbf{E}_{i_k}\left[f(x^{(k+1)}) + \hat{\Psi}_{k+1}\right] \le \mathbf{E}_{i_k}[f(x^{(k+1)})] + \alpha_k\Psi(\tilde{z}^{(k+1)}) + (1-\alpha_k)\hat{\Psi}_k.$$

Combining the above two inequalities, one can obtain that

$$\mathbf{E}_{i_k}\left[f(x^{(k+1)}) + \hat{\Psi}_{k+1}\right] \le (1-\alpha_k)\left(f(x^{(k)}) + \hat{\Psi}_k\right) + \alpha_k V(\tilde{z}^{(k+1)}), \qquad (3.19)$$

where

$$V(x) = f(y^{(k)}) + \left\langle \nabla f(y^{(k)}), x - y^{(k)}\right\rangle + \frac{\gamma_{k+1}}{2n\alpha_k}\left\|x - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}}y^{(k)}\right\|_L^2 + \Psi(x).$$

Comparing with the definition of $\tilde{z}^{(k+1)}$ in (2.5), we see that

$$\tilde{z}^{(k+1)} = \arg\min_{x\in\Re^N} V(x). \qquad (3.20)$$

Notice that V has convexity parameter $\frac{\gamma_{k+1}}{n\alpha_k} = n\alpha_k$ with respect to $\|\cdot\|_L$. By the optimality condition of (3.20), we have that for any $x^\star \in X^*$,

$$V(x^\star) \ge V(\tilde{z}^{(k+1)}) + \frac{\gamma_{k+1}}{2n\alpha_k}\|x^\star - \tilde{z}^{(k+1)}\|_L^2.$$

Using the above inequality and the definition of V, we obtain

$$\begin{aligned}
V(\tilde{z}^{(k+1)}) &\le V(x^\star) - \frac{\gamma_{k+1}}{2n\alpha_k}\|x^\star - \tilde{z}^{(k+1)}\|_L^2 \\
&= f(y^{(k)}) + \left\langle \nabla f(y^{(k)}), x^\star - y^{(k)}\right\rangle + \frac{\gamma_{k+1}}{2n\alpha_k}\left\|x^\star - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}}y^{(k)}\right\|_L^2 \\
&\quad + \Psi(x^\star) - \frac{\gamma_{k+1}}{2n\alpha_k}\|x^\star - \tilde{z}^{(k+1)}\|_L^2.
\end{aligned}$$

Now using the assumption that f has convexity parameter μ with respect to $\|\cdot\|_L$, we have

$$V(\tilde{z}^{(k+1)}) \le f(x^\star) - \frac{\mu}{2}\|x^\star - y^{(k)}\|_L^2 + \frac{\gamma_{k+1}}{2n\alpha_k}\left\|x^\star - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}} z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}} y^{(k)}\right\|_L^2 + \Psi(x^\star) \\ - \frac{\gamma_{k+1}}{2n\alpha_k}\|x^\star - \tilde{z}^{(k+1)}\|_L^2.$$

Combining this inequality with (3.19), one sees that

(3.21)
$$\mathbf{E}_{i_k}\Big[f(x^{(k+1)}) + \hat{\Psi}_{k+1}\Big] \le (1-\alpha_k)\Big(f(x^{(k)}) + \hat{\Psi}_k\Big) + \alpha_k F^\star - \frac{\alpha_k\mu}{2}\|x^\star - y^{(k)}\|_L^2 \\ + \frac{\gamma_{k+1}}{2n}\left\|x^\star - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}} z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}} y^{(k)}\right\|_L^2 - \frac{\gamma_{k+1}}{2n}\|x^\star - \tilde{z}^{(k+1)}\|_L^2.$$

In addition, it follows from (2.2) and the convexity of $\|\cdot\|_L^2$ that

(3.22)
$$\left\|x^\star - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}} z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}} y^{(k)}\right\|_L^2 \le \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\|x^\star - z^{(k)}\|_L^2 + \frac{\alpha_k\mu}{\gamma_{k+1}}\|x^\star - y^{(k)}\|_L^2.$$

Using this relation and (2.6), we observe that

$$\begin{aligned}
&\mathbf{E}_{i_k}\left[\frac{\gamma_{k+1}}{2}\|x^\star - z^{(k+1)}\|_L^2\right] \\
&= \frac{\gamma_{k+1}}{2}\left[\frac{n-1}{n}\left\|x^\star - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}} z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}} y^{(k)}\right\|_L^2 + \frac{1}{n}\|x^\star - \tilde{z}^{(k+1)}\|_L^2\right] \\
&= \frac{\gamma_{k+1}(n-1)}{2n}\left\|x^\star - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}} z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}} y^{(k)}\right\|_L^2 + \frac{\gamma_{k+1}}{2n}\|x^\star - \tilde{z}^{(k+1)}\|_L^2 \\
&= \frac{\gamma_{k+1}}{2}\left\|x^\star - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}} z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}} y^{(k)}\right\|_L^2 \\
&\quad - \frac{\gamma_{k+1}}{2n}\left\|x^\star - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}} z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}} y^{(k)}\right\|_L^2 + \frac{\gamma_{k+1}}{2n}\|x^\star - \tilde{z}^{(k+1)}\|_L^2 \\
&\le \frac{(1-\alpha_k)\gamma_k}{2}\|x^\star - z^{(k)}\|_L^2 + \frac{\alpha_k\mu}{2}\|x^\star - y^{(k)}\|_L^2 \\
&\quad - \frac{\gamma_{k+1}}{2n}\left\|x^\star - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}} z^{(k)} - \frac{\alpha_k\mu}{\gamma_{k+1}} y^{(k)}\right\|_L^2 + \frac{\gamma_{k+1}}{2n}\|x^\star - \tilde{z}^{(k+1)}\|_L^2,
\end{aligned}$$

where the inequality follows from (3.22). Summing up this inequality and (3.21) gives

$$\mathbf{E}_{i_k}\left[f(x^{(k+1)}) + \hat{\Psi}_{k+1} + \frac{\gamma_{k+1}}{2}\|x^\star - z^{(k+1)}\|_L^2\right] \\ \le (1-\alpha_k)\left(f(x^{(k)}) + \hat{\Psi}_k + \frac{\gamma_k}{2}\|x^\star - z^{(k)}\|_L^2\right) + \alpha_k F^\star.$$

Subtracting $F^\star$ from both sides and taking expectation with respect to ξ_{k-1} yields

$$\mathbf{E}_{\xi_k}\left[f(x^{(k+1)}) + \hat{\Psi}_{k+1} - F^\star + \frac{\gamma_{k+1}}{2}\|x^\star - z^{(k+1)}\|_L^2\right] \\ \le (1-\alpha_k)\mathbf{E}_{\xi_{k-1}}\left[f(x^{(k)}) + \hat{\Psi}_k - F^\star + \frac{\gamma_k}{2}\|x^\star - z^{(k)}\|_L^2\right],$$

which together with $\hat{\Psi}_0 = \Psi(x^{(0)})$, $z^{(0)} = x^{(0)}$, and $\lambda_k = \Pi_{i=0}^{k-1}(1-\alpha_i)$ gives

$$\mathbf{E}_{\xi_{k-1}}\left[f(x^{(k)})+\hat{\Psi}_k - F^\star + \frac{\gamma_k}{2}\|x^\star - z^{(k)}\|_L^2\right] \le \lambda_k\left[F(x^{(0)}) - F^\star + \frac{\gamma_0}{2}\|x^\star - x^{(0)}\|_L^2\right].$$

The conclusion of Theorem 2.1 immediately follows from $F(x^{(k)}) \le f(x^{(k)}) + \hat{\Psi}_k$, Lemma 3.1(v), the arbitrariness of $x^\star$, and the definition of R_0.

4. Efficient implementation.

The APCG methods we presented in section 2 all need to perform full-dimensional vector operations at each iteration. In particular, $y^{(k)}$ is updated as a convex combination of $x^{(k)}$ and $z^{(k)}$, and this can be very costly since in general they are dense vectors in $\mathbb{R}^N$. Moreover, in the strongly convex case (Algorithms 1 and 2), all blocks of $z^{(k+1)}$ also need to be updated at each iteration, although only the i_kth block needs to compute the partial gradient and perform an proximal mapping of Ψ_{i_k}. These full-dimensional vector updates cost $O(N)$ operations per iteration and may cause the overall computational cost of APCG to be comparable or even higher than the full gradient methods (see discussions in [26]).

In order to avoid full-dimensional vector operations, Lee and Sidford [14] proposed a change of variables scheme for accelerated coordinated gradient methods for unconstrained smooth minimization. Fercoq and Richtárik [8] devised a similar scheme for efficient implementation in the non-strongly convex case ($\mu = 0$) for composite minimization. Here we show that full vector operations can also be avoided in the strongly convex case for minimizing composite functions. For simplicity, we only present an efficient implementation of the simplified APCG method with $\mu > 0$ (Algorithm 2), which is given as Algorithm 4.

Proposition 4.1. *The iterates of Algorithms* 2 *and* 4 *satisfy the following relationships:*

$$\begin{aligned} x^{(k)} &= \rho^k u^{(k)} + v^{(k)}, \\ y^{(k)} &= \rho^{k+1} u^{(k)} + v^{(k)}, \\ z^{(k)} &= -\rho^k u^{(k)} + v^{(k)} \end{aligned} \tag{4.1}$$

for all $k \ge 0$. That is, these two algorithms are equivalent.

Proof. We prove by induction. Notice that Algorithm 2 is initialized with $z^{(0)} = x^{(0)}$, and its first step implies $y^{(0)} = \frac{x^{(0)}+\alpha z^{(0)}}{1+\alpha} = x^{(0)}$; Algorithm 4 is initialized with

Algorithm 4. Efficient implementation of APCG with $\gamma_0 = \mu > 0$.

```
input: $x^{(0)} \in \mathrm{dom}\,(\Psi)$ and convexity parameter $\mu > 0$.
initialize: set $\alpha = \frac{\sqrt{\mu}}{n}$ and $\rho = \frac{1-\alpha}{1+\alpha}$, and initialize $u^{(0)} = 0$ and $v^{(0)} = x^{(0)}$.
iterate: repeat for $k = 0, 1, 2, \ldots$
  1. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random and compute
       $h_{i_k}^{(k)} = \arg\min_{h \in \mathbb{R}^{N_{i_k}}} \left\{ \frac{n\alpha L_{i_k}}{2}\|h\|_2^2 + \left\langle \nabla_{i_k} f(\rho^{k+1}u^{(k)} + v^{(k)}), h \right\rangle + \Psi_{i_k}\left(-\rho^{k+1}u_{i_k}^{(k)} + v_{i_k}^{(k)} + h\right) \right\}.$
  2. Let $u^{(k+1)} = u^{(k)}$ and $v^{(k+1)} = v^{(k)}$, and update
       (4.2)   $u_{i_k}^{(k+1)} = u_{i_k}^{(k)} - \frac{1-n\alpha}{2\rho^{k+1}} h_{i_k}^{(k)}, \qquad v_{i_k}^{(k+1)} = v_{i_k}^{(k)} + \frac{1+n\alpha}{2} h_{i_k}^{(k)}.$
output: $x^{(k+1)} = \rho^{k+1}u^{(k+1)} + v^{(k+1)}$
```

$u^{(0)} = 0$ and $v^{(0)} = x^{(0)}$. Therefore we have

$$x^{(0)} = \rho^0 u^{(0)} + v^{(0)}, \qquad y^{(0)} = \rho^1 u^{(0)} + v^{(0)}, \qquad z^{(0)} = -\rho^0 u^{(0)} + v^{(0)},$$

which means that (4.1) holds for $k = 0$. Now suppose it holds for some $k \geq 0$; then

$$\begin{aligned}(1-\alpha)z^{(k)} + \alpha y^{(k)} &= (1-\alpha)\left(-\rho^k u^{(k)} + v^{(k)}\right) + \alpha\left(\rho^{k+1}u^{(k)} + v^{(k)}\right)\\ &= -\rho^k\left((1-\alpha) - \alpha\rho\right)u^{(k)} + (1-\alpha)v^{(k)} + \alpha v^{(k)}\\ &= -\rho^{k+1}u^{(k)} + v^{(k)}. \qquad (4.3)\end{aligned}$$

So $h_{i_k}^{(k)}$ in Algorithm 4 can be written as

$$h_{i_k}^{(k)} = \underset{h\in\mathbb{R}^{N_{i_k}}}{\arg\min}\left\{\frac{n\alpha L_{i_k}}{2}\|h\|_2^2 + \langle\nabla_{i_k} f(y^{(k)}), h\rangle + \Psi_{i_k}\left((1-\alpha)z_{i_k}^{(k)} + \alpha y_{i_k}^{(k)} + h\right)\right\}.$$

Comparing with (2.5), and using $\beta_k = \alpha$, we obtain

$$h_{i_k}^{(k)} = \tilde{z}_{i_k}^{(k+1)} - \left((1-\alpha)z_{i_k}^{(k)} + \alpha y_{i_k}^{(k)}\right).$$

In terms of the full-dimensional vectors, using (2.6) and (4.3), we have

$$\begin{aligned}z^{(k+1)} &= (1-\alpha)z^{(k)} + \alpha y^{(k)} + U_{i_k}h_{i_k}^{(k)} \;=\; -\rho^{k+1}u^{(k)} + v^{(k)} + U_{i_k}h_{i_k}^{(k)}\\ &= -\rho^{k+1}u^{(k)} + v^{(k)} + \frac{1-n\alpha}{2}U_{i_k}h_{i_k}^{(k)} + \frac{1+n\alpha}{2}U_{i_k}h_{i_k}^{(k)}\\ &= -\rho^{k+1}\left(u^{(k)} - \frac{1-n\alpha}{2\rho^{k+1}}U_{i_k}h_{i_k}^{(k)}\right) + \left(v^{(k)} + \frac{1+n\alpha}{2}U_{i_k}h_{i_k}^{(k)}\right)\\ &= -\rho^{k+1}u^{(k+1)} + v^{(k+1)}.\end{aligned}$$

Using step 3 of Algorithm 2, we get

$$\begin{aligned}x^{(k+1)} &= y^{(k)} + n\alpha(z^{(k+1)} - z^{(k)}) + n\alpha^2(z^{(k)} - y^{(k)})\\ &= y^{(k)} + n\alpha\left(z^{(k+1)} - \left((1-\alpha)z^{(k)} + \alpha y^{(k)}\right)\right)\\ &= y^{(k)} + n\alpha U_{i_k}h_{i_k}^{(k)},\end{aligned}$$

where the last step used (2.6). Using the induction hypothesis $y^{(k)} = \rho^{k+1}u^{(k)} + v^{(k)}$, we have

$$\begin{aligned}x^{(k+1)} &= \rho^{k+1}u^{(k)} + v^{(k)} + \frac{1-n\alpha}{2}U_{i_k}h_{i_k}^{(k)} + \frac{1+n\alpha}{2}U_{i_k}h_{i_k}^{(k)}\\ &= \rho^{k+1}\left(u^{(k)} - \frac{1-n\alpha}{2\rho^{k+1}}U_{i_k}h_{i_k}^{(k)}\right) + \left(v^{(k)} + \frac{1+n\alpha}{2}U_{i_k}h_{i_k}^{(k)}\right)\\ &= \rho^{k+1}u^{(k+1)} + v^{(k+1)}.\end{aligned}$$

Finally,

$$\begin{aligned}y^{(k+1)} &= \frac{1}{1+\alpha}\left(x^{(k+1)} + \alpha z^{(k+1)}\right)\\ &= \frac{1}{1+\alpha}\left(\rho^{k+1}u^{(k+1)} + v^{(k+1)}\right) + \frac{\alpha}{1+\alpha}\left(-\rho^{k+1}u^{(k+1)} + v^{(k+1)}\right)\\ &= \frac{1-\alpha}{1+\alpha}\rho^{k+1}u^{(k+1)} + \frac{1+\alpha}{1+\alpha}v^{(k+1)} \;=\; \rho^{k+2}u^{(k+1)} + v^{(k+1)}.\end{aligned}$$

We just showed that (4.1) also holds for $k+1$. This finishes the induction. □

We note that in Algorithm 4, only single block coordinates of the vectors $u^{(k)}$ and $v^{(k)}$ are updated at each iteration, which costs $O(N_{i_k})$. However, computing the partial gradient $\nabla_{i_k} f(\rho^{k+1}u^{(k)}+v^{(k)})$ may still cost $O(N)$ in general. In section 5.2, we show how to further exploit the problem structure in regularized ERM to completely avoid full-dimensional vector operations.

5. Application to regularized ERM.

In this section, we show how to apply the APCG method to solve the regularized ERM problems associated with linear predictors.

Let $A_1, \ldots, A_n$ be vectors in $\mathbb{R}^d$, $\phi_1, \ldots, \phi_n$ be a sequence of convex functions defined on $\mathbb{R}$, and g be a convex function defined on $\mathbb{R}^d$. The goal of regularized ERM with linear predictors is to solve the following (convex) optimization problem:

$$\underset{w\in\mathbb{R}^d}{\text{minimize}} \quad \left\{ P(w) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \phi_i(A_i^T w) + \lambda g(w) \right\}, \tag{5.1}$$

where $\lambda > 0$ is a regularization parameter. For binary classification, given a label $b_i \in \{\pm 1\}$ for each vector A_i, for $i = 1, \ldots, n$, we obtain the linear support vector machine (SVM) problem by setting $\phi_i(z) = \max\{0, 1 - b_i z\}$ and $g(w) = (1/2)\|w\|_2^2$. Regularized logistic regression is obtained by setting $\phi_i(z) = \log(1 + \exp(-b_i z))$. This formulation also includes regression problems. For example, ridge regression is obtained by setting $\phi_i(z) = (1/2)(z - b_i)^2$ and $g(w) = (1/2)\|w\|_2^2$, and we get the lasso if $g(w) = \|w\|_1$. Our method can also be extended to cases where each A_i is a matrix, thus covering multiclass classification problems as well (see, e.g., [39]).

For each $i = 1, \ldots, n$, let ϕ_i^* be the convex conjugate of ϕ_i, that is,

$$\phi_i^*(u) = \max_{z\in\mathbb{R}} \ \{zu - \phi_i(z)\}.$$

The dual of the regularized ERM problem (5.1), which we call the primal, is to solve the problem (see, e.g., [40])

$$\underset{x\in\mathbb{R}^n}{\text{maximize}} \quad \left\{ D(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} -\phi_i^*(-x_i) - \lambda g^*\left(\frac{1}{\lambda n}Ax\right) \right\}, \tag{5.2}$$

where $A = [A_1, \ldots, A_n]$. This is equivalent to minimizing $F(x) \overset{\text{def}}{=} -D(x)$, that is,

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \left\{ F(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \phi_i^*(-x_i) + \lambda g^*\left(\frac{1}{\lambda n}Ax\right) \right\}. \tag{5.3}$$

The structure of $F(x)$ above matches our general formulation of minimizing composite convex functions in (1.1) and (1.2) with

$$f(x) = \lambda g^*\left(\frac{1}{\lambda n}Ax\right), \qquad \Psi(x) = \frac{1}{n}\sum_{i=1}^{n} \phi_i^*(-x_i). \tag{5.4}$$

Therefore, we can directly apply the APCG method to solve problem (5.3), i.e., to solve the dual of the regularized ERM problem. Here we assume that the proximal mappings of the conjugate functions ϕ_i^* can be computed efficiently, which is indeed the case for many regularized ERM problems (see, e.g., [40, 39]).

In order to obtain accelerated linear convergence rates, we make the following assumption.

Assumption 3. Each function ϕ_i is $1/\gamma$ smooth, and the function g has unit convexity parameter 1.

Here we slightly abuse the notation by overloading γ and λ, which appeared in sections 2 and 3. In this section γ represents the (inverse) smoothness parameter of ϕ_i, and λ denotes the regularization parameter on g. Assumption 3 implies that each ϕ_i^* has strong convexity parameter γ (with respect to the local Euclidean norm) and g^* is differentiable and ∇g^* has Lipschitz constant 1.

We note that in the linear SVM problem, the hinge loss $\phi_i(z) = \max\{0, 1 - b_i z\}$ is not differentiable; thus it does not satisfy Assumption 3 and $F(x)$ is not strongly convex. In this case, we can apply APCG with $\mu = 0$ (Algorithm 3) to obtain accelerated sublinear rates, which is better than the convergence rate of SDCA under the same assumption [40]. In this section, we focus on problems that satisfy Assumption 3, which enjoy accelerated linear convergence rates. In order to apply these results, a standard trick in machine learning is to replace the hingle loss with a smoothed approximation; see section 5.3 and more examples in [39, section 5.1]. On the other hand, in the Lasso formulation we have $g(w) = \|w\|_1$, which is not strongly convex. In this case, a common scheme in practice is to combine it with a small $\|w\|_2^2$ regularization; see [39, section 5.2] for further details.

In order to match the condition in Assumption 2, i.e., $f(x)$ needs to be strongly convex, we can apply the technique in section 2.2 to relocate the strong convexity from Ψ to f. Without loss of generality, we can use the following splitting of the composite function $F(x) = f(x) + \Psi(x)$:

$$f(x) = \lambda g^*\left(\frac{1}{\lambda n} Ax\right) + \frac{\gamma}{2n}\|x\|_2^2, \qquad \Psi(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\phi_i^*(-x_i) - \frac{\gamma}{2}\|x_i\|_2^2\right). \tag{5.5}$$

Under Assumption 3, the function f is smooth and strongly convex and each Ψ_i, for $i = 1, \ldots, n$, is still convex. As a result, we have the following complexity guarantee when applying the APCG method to minimize the function $F(x) = -D(x)$.

THEOREM 5.1. *Suppose Assumption* 3 *holds and* $\|A_i\|_2 \le R$ *for all* $i = 1, \ldots, n$. *In order to obtain an expected dual optimality gap* $\mathbf{E}[D^\star - D(x^{(k)})] \le \epsilon$ *using the APCG method, it suffices to have*

$$k \ge \left(n + \sqrt{\frac{nR^2}{\lambda\gamma}}\right)\log(C/\epsilon), \tag{5.6}$$

where $D^\star = \max_{x\in\mathbb{R}^n} D(x)$ *and*

$$C = D^\star - D(x^{(0)}) + \frac{\gamma}{2n}\|x^{(0)} - x^\star\|_2^2. \tag{5.7}$$

Proof. First, we notice that the function $f(x)$ defined in (5.5) is differentiable. Moreover, for any $x \in \mathbb{R}^n$ and $h_i \in \mathbb{R}$,

$$\begin{aligned}
\|\nabla_i f(x+U_i h_i) - \nabla_i f(x)\|_2 &= \left\|\frac{1}{n}A_i^T\left[\nabla g^*\left(\frac{1}{\lambda n}A(x+U_i h_i)\right) - \nabla g^*\left(\frac{1}{\lambda n}Ax\right)\right] + \frac{\gamma}{n}h_i\right\|_2 \\
&\le \frac{\|A_i\|_2}{n}\left\|\nabla g^*\left(\frac{1}{\lambda n}A(x+U_i h_i)\right) - \nabla g^*\left(\frac{1}{\lambda n}Ax\right)\right\|_2 + \frac{\gamma}{n}\|h_i\|_2 \\
&\le \frac{\|A_i\|_2}{n}\left\|\frac{1}{\lambda n}A_i h_i\right\|_2 + \frac{\gamma}{n}\|h_i\|_2 \;\le\; \left(\frac{\|A_i\|_2^2}{\lambda n^2} + \frac{\gamma}{n}\right)\|h_i\|_2,
\end{aligned}$$

where the second inequality used the assumption that g has convexity parameter 1 and thus ∇g^* has Lipschitz constant 1. The coordinatewise Lipschitz constants as defined in Assumption 1 are

$$L_i \;=\; \frac{\|A_i\|_2^2}{\lambda n^2} + \frac{\gamma}{n} \;\leq\; \frac{R^2 + \lambda\gamma n}{\lambda n^2}, \qquad i = 1, \ldots, n.$$

The function f has convexity parameter $\frac{\gamma}{n}$ with respect to the Euclidean norm $\|\cdot\|_2$. Let μ be its convexity parameter with respect to the norm $\|\cdot\|_L$ defined in (1.6). Then

$$\mu \;\geq\; \frac{\gamma}{n}\Big/\frac{R^2+\lambda\gamma n}{\lambda n^2} \;=\; \frac{\lambda\gamma n}{R^2+\lambda\gamma n}.$$

According to Theorem 2.1, the APCG method converges geometrically:

$$\mathbf{E}\left[D^\star - D(x^{(k)})\right] \;\leq\; \left(1 - \frac{\sqrt{\mu}}{n}\right)^k C \;\leq\; \exp\left(-\frac{\sqrt{\mu}}{n}k\right) C,$$

where the constant C is given in (5.7). Therefore, in order to have $\mathbf{E}[D^\star - D(x^{(k)})] \leq \epsilon$, it suffices to have the number of iterations k be larger than

$$\frac{n}{\sqrt{\mu}}\log\left(\frac{C}{\epsilon}\right) \leq n\sqrt{\frac{R^2+\lambda\gamma n}{\lambda\gamma n}}\log\left(\frac{C}{\epsilon}\right) = \sqrt{n^2+\frac{nR^2}{\lambda\gamma}}\log\left(\frac{C}{\epsilon}\right) \leq \left(n+\sqrt{\frac{nR^2}{\lambda\gamma}}\right)\log\left(\frac{C}{\epsilon}\right).$$

This finishes the proof. □

Let us compare the result in Theorem 5.1 with the complexity of solving the dual problem (5.3) using the AFG method of Nesterov [27]. Using the splitting in (5.4) and under Assumption 3, the gradient $\nabla f(x)$ has Lipschitz constant $\frac{\|A\|_2^2}{\lambda n^2}$, where $\|A\|_2$ denotes the spectral norm of A, and $\Psi(x)$ has convexity parameter $\frac{\gamma}{n}$ with respect to $\|\cdot\|_2$. So the condition number of the problem is

$$\kappa = \frac{\|A\|_2^2}{\lambda n^2}\Big/\frac{\gamma}{n} = \frac{\|A\|_2^2}{\lambda\gamma n}.$$

Suppose each iteration of the AFG method costs as much as n times the APCG method (as we will see in section 5.2); then the complexity of the AFG method [27, Theorem 6] measured in terms of number of coordinate gradient steps is

$$O\left(n\sqrt{\kappa}\log(1/\epsilon)\right) \;=\; O\left(\sqrt{\frac{n\|A\|_2^2}{\lambda\gamma}}\log(1/\epsilon)\right) \;\leq\; O\left(\sqrt{\frac{n^2R^2}{\lambda\gamma}}\log(1/\epsilon)\right).$$

The inequality above is due to $\|A\|_2^2 \leq \|A\|_F^2 \leq nR^2$. Therefore in the ill-conditioned case (assuming $n \leq \frac{R^2}{\lambda\gamma}$), the complexity of AFG can be a factor of $\sqrt{n}$ worse than that of APCG.

Several state-of-the-art algorithms for regularized ERM, including SDCA [40], SAG [35, 37], and SVRG [11, 48], have the iteration complexity

$$O\left(\left(n + \frac{R^2}{\lambda\gamma}\right)\log(1/\epsilon)\right).$$

Here the ratio $\frac{R^2}{\lambda\gamma}$ can be interpreted as the condition number of the regularized ERM problem (5.1) and its dual (5.2). We note that our result in (5.6) can be much better for ill-conditioned problems, i.e., when the condition number $\frac{R^2}{\lambda\gamma}$ is much larger than n.

Most recently, Shalev-Shwartz and Zhang [39] developed an accelerated SDCA method which achieves the same complexity $O((n+\sqrt{\frac{n}{\lambda\gamma}})\log(1/\epsilon))$ as our method. Their method is an inner-outer iteration procedure, where the outer loop is a full-dimensional accelerated gradient method in the primal space $w \in \mathbb{R}^d$. At each iteration of the outer loop, the SDCA method [40] is called to solve the dual problem (5.2) with customized regularization parameter and precision. In contrast, our APCG method is a straightforward single loop coordinate gradient method.

We note that the complexity bounds for the aforementioned work are either for the primal optimality $P(w^{(k)})-P^\star$ (SAG and SVRG) or for the primal-dual gap $P(w^{(k)})-D(x^{(k)})$ (SDCA and accelerated SDCA). Our results in Theorem 5.1 are in terms of the dual optimality $D^\star - D(x^{(k)})$. In section 5.1, we show how to recover primal solutions with the same order of convergence rate. In section 5.2, we show how to exploit the problem structure of regularized ERM to compute the partial gradient $\nabla_i f(x)$, which together with the efficient implementation proposed in section 4 completely avoids full-dimensional vector operations. The experiments in section 5.3 illustrate that our method has superior performance in reducing both the primal objective value and the primal-dual gap.

5.1. Recovering the primal solution.

Under Assumption 3, the primal problem (5.1) and dual problem (5.2) each have a unique solution, say, $w^\star$ and $x^\star$, respectively. Moreover, we have $P(w^\star) = D(x^\star)$. With the definition

$$\omega(x) = \nabla g^*\left(\frac{1}{\lambda n}Ax\right), \tag{5.8}$$

we have $w^\star = \omega(x^\star)$. When applying the APCG method to solve the dual regularized ERM problem, which generates a dual sequence $x^{(k)}$, we can obtain a primal sequence $w^{(k)} = \omega(x^{(k)})$. Here we discuss the relationship between the primal-dual gap $P(w^{(k)}) - D(x^{(k)})$ and the dual optimality $D^\star - D(x^{(k)})$.

Let $a = (a_1, \ldots, a_n)$ be a vector in $\mathbb{R}^n$. We consider the saddle-point problem

$$\max_x \ \min_{a,w} \ \left\{\Phi(x,a,w) \ \overset{\text{def}}{=} \ \frac{1}{n}\sum_{i=1}^n \phi_i(a_i) + \lambda g(w) - \frac{1}{n}\sum_{i=1}^n x_i(A_i^T w - a_i)\right\}, \tag{5.9}$$

so that

$$D(x) \ = \ \min_{a,w} \ \Phi(x,a,w).$$

Given an approximate dual solution $x^{(k)}$ (generated by the APCG method), we can find a pair of primal solutions $(a^{(k)}, w^{(k)}) = \arg\min_{a,w} \Phi(x^{(k)}, a, w)$, or more specifically,

$$a_i^{(k)} = \arg\max_{a_i}\left\{-x_i^{(k)}a_i - \phi_i(a_i)\right\} \ \in \ \partial\phi_i^*(-x_i^{(k)}), \quad i = 1, \ldots, n, \tag{5.10}$$

$$w^{(k)} = \arg\max_w\left\{w^T\left(\frac{1}{\lambda n}Ax^{(k)}\right) - g(w)\right\} \ = \ \nabla g^*\left(\frac{1}{\lambda n}Ax^{(k)}\right). \tag{5.11}$$

As a result, we obtain a subgradient of D at $x^{(k)}$, denoted $D'(x^{(k)})$, and

$$\|D'(x^{(k)})\|_2^2 = \frac{1}{n^2}\sum_{i=1}^{n}\left(A_i^T w^{(k)} - a_i^{(k)}\right)^2. \tag{5.12}$$

We note that $\|D'(x^{(k)})\|_2^2$ is not only a measure of the dual optimality of $x^{(k)}$ but also a measure of the primal feasibility of $(a^{(k)}, w^{(k)})$. In fact, it can also bound the primal-dual gap, which is the result of the following lemma.

LEMMA 5.2. *Given any dual solution $x^{(k)}$, let $(a^{(k)}, w^{(k)})$ be defined as in* (5.10) *and* (5.11). *Then*

$$P(w^{(k)}) - D(x^{(k)}) \;\le\; \frac{1}{2n\gamma}\sum_{i=1}^{n}\left(A_i^T w^{(k)} - a_i^{(k)}\right)^2 \;=\; \frac{n}{2\gamma}\|D'(x^{(k)})\|_2^2.$$

Proof. Because of (5.10), we have $\nabla\phi_i(a_i^{(k)}) = -x_i^{(k)}$. The $1/\gamma$-smoothness of $\phi_i(a)$ implies

$$\begin{aligned}
P(w^{(k)}) &= \frac{1}{n}\sum_{i=1}^{n}\phi_i(A_i^T w^{(k)}) + \lambda g(w^{(k)})\\
&\le \frac{1}{n}\sum_{i=1}^{n}\left(\phi_i(a_i^{(k)}) + \nabla\phi_i(a_i^{(k)})^T\left(A_i^T w^{(k)} - a_i^{(k)}\right) + \frac{1}{2\gamma}\left(A_i^T w^{(k)} - a_i^{(k)}\right)^2\right) + \lambda g(w^{(k)})\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\phi_i(a_i^{(k)}) - \left(x_i^{(k)}\right)^T\left(A_i^T w^{(k)} - a_i^{(k)}\right) + \frac{1}{2\gamma}\left(A_i^T w^{(k)} - a_i^{(k)}\right)^2\right) + \lambda g(w^{(k)})\\
&= \Phi(x^{(k)}, a^{(k)}, w^{(k)}) + \frac{1}{2n\gamma}\sum_{i=1}^{n}\left(A_i^T w^{(k)} - a_i^{(k)}\right)^2\\
&= D(x^{(k)}) + \frac{1}{2n\gamma}\sum_{i=1}^{n}(A_i^T w^{(k)} - a_i^{(k)})^2,
\end{aligned}$$

which leads to the inequality in the conclusion. The equality in the conclusion is due to (5.12). ▯

The following theorem states that under a stronger assumption than Assumption 3, the primal-dual gap can be bounded directly by the dual optimality gap; hence they share the same order of convergence rate.

THEOREM 5.3. *Suppose g is 1-strongly convex and each ϕ_i is $1/\gamma$-smooth and also $1/\eta$-strongly convex (all with respect to the Euclidean norm $\|\cdot\|_2$). Given any dual point $x^{(k)}$, let the primal correspondence be $w^{(k)} = \omega(x^{(k)})$, i.e., generated from* (5.11). *Then we have*

$$P(w^{(k)}) - D(x^{(k)}) \;\le\; \frac{\lambda\eta n + \|A\|_2^2}{\lambda\gamma n}\left(D^\star - D(x^{(k)})\right), \tag{5.13}$$

where $\|A\|_2$ denotes the spectral norm of A.

Proof. Since $g(w)$ is 1-strongly convex, the function $f(x) = \lambda g^*\left(\frac{Ax}{\lambda n}\right)$ is differentiable and $\nabla f(x)$ has Lipschitz constant $\frac{\|A\|_2^2}{\lambda n^2}$. Similarly, since each ϕ_i is $1/\eta$ strongly convex, the function $\Psi(x) = \frac{1}{n}\sum_{i=1}^{n}\phi_i^*(-x_i)$ is differentiable and $\nabla\Psi(x)$ has Lipschitz constant $\frac{\eta}{n}$. Therefore, the function $-D(x) = f(x) + \Psi(x)$ is smooth and its gradient has Lipschitz constant

$$\frac{\|A\|_2^2}{\lambda n^2} + \frac{\eta}{n} = \frac{\lambda\eta n + \|A\|_2^2}{\lambda n^2}.$$

It is known that (e.g., [25, Theorem 2.1.5]) if a function $F(x)$ is convex and L-smooth, then

$$F(y) \geq F(x) + \nabla F(x)^T(y-x) + \frac{1}{2L}\|\nabla F(x) - \nabla F(y)\|_2^2$$

for all $x, y \in \mathbb{R}^n$. Applying the above inequality to $F(x) = -D(x)$, we get for all x and y,

$$-D(y) \geq -D(x) - \nabla D(x)^T(y-x) + \frac{\lambda n^2}{2(\lambda\eta n + \|A\|_2^2)}\|\nabla D(x) - \nabla D(y)\|_2^2. \tag{5.14}$$

Under our assumptions, the saddle-point problem (5.9) has a unique solution $(x^\star, a^\star, w^\star)$, where $w^\star$ and $x^\star$ are the solutions to the primal and dual problems (5.1) and (5.2), respectively. Moreover, they satisfy the optimality conditions

$$A_i^T w^\star - a_i^\star = 0, \qquad a_i^\star = \nabla\phi_i^*(-x_i^\star), \qquad w^\star = \nabla g^*\left(\frac{1}{\lambda n}Ax^\star\right).$$

Since D is differentiable in this case, we have $D'(x) = \nabla D(x)$ and $\nabla D(x^\star) = 0$. Now we choose x and y in (5.14) to be $x^\star$ and $x^{(k)}$, respectively. This leads to

$$\|\nabla D(x^{(k)})\|_2^2 \;=\; \|\nabla D(x^{(k)}) - \nabla D(x^\star)\|_2^2 \;\leq\; \frac{2(\lambda\eta n + \|A\|_2^2)}{\lambda n^2}(D(x^\star) - D(x^{(k)})).$$

Then the conclusion can be derived from Lemma 5.2. □

The assumption that each ϕ_i is $1/\gamma$-smooth and $1/\eta$-strongly convex implies that $\gamma \leq \eta$. Therefore the coefficient on the right-hand side of (5.13) satisfies $\frac{\lambda\eta n + \|A\|_2^2}{\lambda\gamma n} > 1$. This is consistent with the fact that for any pair of primal and dual points $w^{(k)}$ and $x^{(k)}$, we always have $P(w^{(k)}) - D(x^{(k)}) \geq D^\star - D(x^{(k)})$.

Corollary 5.4. *Under the assumptions of Theorem* 5.3*, in order to obtain an expected primal-dual gap* $\mathbf{E}\left[P(w^{(k)}) - D(x^{(k)})\right] \leq \epsilon$ *using the APCG method, it suffices to have*

$$k \geq \left(n + \sqrt{\frac{nR^2}{\lambda\gamma}}\right)\log\left(\frac{(\lambda\eta n + \|A\|_2^2)}{\lambda\gamma n}\frac{C}{\epsilon}\right),$$

where the constant C is defined in (5.7).

The above results require that each ϕ_i be both smooth and strongly convex. One example that satisfies such assumptions is ridge regression, where $\phi_i(a_i) = \frac{1}{2}(a_i - b_i)^2$ and $g(w) = \frac{1}{2}\|w\|_2^2$. For problems that only satisfy Assumption 3, we may add a small strongly convex term $\frac{1}{2\eta}a_i^2$ to each loss $\phi_i(a_i)$ and obtain that the primal-dual gap (of a slightly perturbed problem) shares the same accelerated linear convergence rate as the dual optimality gap. Alternatively, we can obtain the same guarantee with the extra cost of a proximal full gradient step. This is summarized in the following theorem.

Theorem 5.5. *Suppose Assumption* 3 *holds. Given any dual point* $x^{(k)}$*, define*

$$T(x^{(k)}) = \arg\min_{x\in\mathbb{R}^n}\left\{\langle\nabla f(x^{(k)}), x\rangle + \frac{\|A\|_2^2}{2\lambda n^2}\|x - x^{(k)}\|_2^2 + \Psi(x)\right\}, \tag{5.15}$$

where f and Ψ are defined in the simple splitting (5.4)*. Let*

$$w^{(k)} = \omega(T(x^{(k)})) = \nabla g^*\left(\frac{1}{\lambda n}AT(x^{(k)})\right). \tag{5.16}$$

Then we have

$$P(w^{(k)}) - D(T(x^{(k)})) \leq \frac{4\|A\|_2^2}{\lambda\gamma n}\left(D(x^\star) - D(x^{(k)})\right). \tag{5.17}$$

Proof. Notice that the Lipschitz constant of $\nabla f(x)$ is $L_f = \frac{\|A\|_2^2}{\lambda n^2}$, which is used in calculating $T(x^{(k)})$. The corresponding *gradient mapping* [27] at $x^{(k)}$ is

$$G(x^{(k)}) = L_f\left(x^{(k)} - T(x^{(k)})\right) = \frac{\|A\|_2^2}{\lambda n^2}\left(x^{(k)} - T(x^{(k)})\right).$$

Applying [27, Theorem 1] to $F(x) = -D(x)$, we have for any $x \in \operatorname{dom} F$,

$$\langle -D'\big(T(x^{(k)})\big), x - T(x^{(k)})\rangle \geq -2\left\|G(x^{(k)})\right\|_2 \cdot \left\|x - T(x^{(k)})\right\|_2.$$

Letting $x = T(x^{(k)}) + D'\big(T(x^{(k)})\big)$, we have $\left\|D'\big(T(x^{(k)})\big)\right\|_2 \leq 2\left\|G(x^{(k)})\right\|_2$. Another consequence of [27, Theorem 1] is

$$\begin{aligned}\left\|G(x^{(k)})\right\|_2^2 &\leq 2L_f\left(F(x^{(k)}) - F\big(T(x^{(k)})\big)\right)\\ &\leq 2L_f\left(F(x^{(k)}) - F(x^\star)\right) = 2L_f\left(D(x^\star) - D(x^{(k)})\right).\end{aligned}$$

Combining these two inequalities yields

$$\left\|D'\big(T(x^{(k)})\big)\right\|_2^2 \leq 4\left\|G(x^{(k)})\right\|_2^2 \leq 8L_f\Big(D(x^\star) - D(x^{(k)})\Big) = \frac{8\|A\|_2^2}{\lambda n^2}\Big(D(x^\star) - D(x^{(k)})\Big).$$

The conclusion can then be derived from Lemma 5.2. □

Here the coefficient in the right-hand side of (5.17), $\frac{4\|A\|_2^2}{\lambda\gamma n}$, can be less than 1. This does not contradict the fact that the primal-dual gap should be no less than the dual optimality gap, because the primal-dual gap on the left-hand side of (5.17) is measured at $T(x^{(k)})$ rather than $x^{(k)}$.

Corollary 5.6. *Suppose Assumption* 3 *holds. In order to obtain a primal-dual pair $w^{(k)}$ and $x^{(k)}$ such that $\mathbf{E}\left[P(w^{(k)}) - D(T(x^{(k)}))\right] \leq \epsilon$, it suffices to run the APCG method for*

$$k \geq \left(n + \sqrt{\frac{nR^2}{\lambda\gamma}}\right)\log\left(\frac{4\|A\|_2^2}{\lambda\gamma n}\frac{C}{\epsilon}\right)$$

steps and follow with a proximal full gradient step (5.15) *and* (5.16), *where C is defined in* (5.7).

We note that the computational cost of the proximal full gradient step (5.15) is comparable with n proximal coordinate gradient steps. Therefore the overall complexity of this scheme is on the same order as necessary for the expected dual optimality gap to reach ϵ. Actually the numerical experiments in section 5.3 show that running the APCG method alone without the final full gradient step is sufficient to reduce the primal-dual gap at a very fast rate.

Algorithm 5. APCG for solving regularized ERM with $\mu > 0$.

```
input: $x^{(0)} \in \mathrm{dom}\,(\Psi)$ and convexity parameter $\mu = \frac{\lambda\gamma n}{R^2+\lambda\gamma n}$.
initialize: set $\alpha = \frac{\sqrt{\mu}}{n}$ and $\rho = \frac{1-\alpha}{1+\alpha}$, and let $u^{(0)} = 0$, $v^{(0)} = x^{(0)}$, $p^{(0)} = 0$, and $q^{(0)} = Ax^{(0)}$.
iterate: repeat for $k = 0, 1, 2, \ldots$
  1. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random, compute the coordinate gradient
       $\nabla_{i_k}^{(k)} = \frac{1}{\lambda n^2}\left(\rho^{k+1}A_{i_k}^T p^{(k)} + A_{i_k}^T q^{(k)}\right) + \frac{\gamma}{n}\left(\rho^{k+1}u_{i_k}^{(k)} + v_{i_k}^{(k)}\right).$
  2. Compute coordinate increment
     (5.18)
       $h_{i_k}^{(k)} = \arg\min_{h\in\mathbb{R}^{N_{i_k}}}\left\{\frac{\alpha(\|A_{i_k}\|^2+\lambda\gamma n)}{2\lambda n}\|h\|_2^2 + \langle\nabla_{i_k}^{(k)}, h\rangle + \Psi_{i_k}\left(-\rho^{k+1}u_{i_k}^{(k)}+v_{i_k}^{(k)}+h\right)\right\}.$
  3. Let $u^{(k+1)} = u^{(k)}$ and $v^{(k+1)} = v^{(k)}$, and update
     (5.19)
       $u_{i_k}^{(k+1)} = u_{i_k}^{(k)} - \frac{1-n\alpha}{2\rho^{k+1}}h_{i_k}^{(k)}, \qquad v_{i_k}^{(k+1)} = v_{i_k}^{(k)} + \frac{1+n\alpha}{2}h_{i_k}^{(k)},$
       $p^{(k+1)} = p^{(k)} - \frac{1-n\alpha}{2\rho^{k+1}}A_{i_k}h_{i_k}^{(k)}, \qquad q^{(k+1)} = q^{(k)} + \frac{1+n\alpha}{2}A_{i_k}h_{i_k}^{(k)}.$
output: approximate dual and primal solutions
  $x^{(k+1)} = \rho^{k+1}u^{(k+1)} + v^{(k+1)}, \qquad w^{(k+1)} = \frac{1}{\lambda n}\left(\rho^{k+1}p^{(k+1)} + q^{(k+1)}\right).$
```

5.2. Implementation details. Here we show how to exploit the structure of the regularized ERM problem to efficiently compute the coordinate gradient $\nabla_{i_k} f(y^{(k)})$ and totally avoid full-dimensional updates in Algorithm 4.

We focus on the special case $g(w) = \frac{1}{2}\|w\|_2^2$ and show how to compute $\nabla_{i_k} f(y^{(k)})$. In this case, $g^*(v) = \frac{1}{2}\|v\|_2^2$ and $\nabla g^*(\cdot)$ is the identity map. According to (5.5),

$$\nabla_{i_k} f(y^{(k)}) = \frac{1}{\lambda n^2}A_{i_k}^T(Ay^{(k)}) + \frac{\gamma}{n}y_{i_k}^{(k)}.$$

Notice that we do not form $y^{(k)}$ in Algorithm 4. By Proposition 4.1, we have

$$y^{(k)} = \rho^{k+1}u^{(k)} + v^{(k)}.$$

So we can store and update the two vectors

$$p^{(k)} = Au^{(k)}, \qquad q^{(k)} = Av^{(k)}$$

and obtain

$$Ay^{(k)} = \rho^{k+1}p^{(k)} + q^{(k)}.$$

Since the update of both $u^{(k)}$ and $v^{(k)}$ at each iteration involves only the single coordinate i_k, we can update $p^{(k)}$ and $q^{(k)}$ by adding or subtracting a scaled column A_{i_k}, as given in (5.19). The resulting method is detailed in Algorithm 5.

Table 1
Characteristics of three binary classification datasets obtained from [7].

Dataset	Source	Number of samples n	Number of features d	Sparsity
RCV1	[16]	20,242	47,236	0.16%
covtype	[4]	581,012	54	22%
News20	[12, 13]	19,996	1,355,191	0.04%

In Algorithm 5, we use $\nabla_{i_k}^{(k)}$ to represent $\nabla_{i_k} f(y^{(k)})$ to reflect the fact that we never form $y^{(k)}$ explicitly. The function Ψ_i in (5.18) is the one given in (5.5), i.e.,

$$\Psi_i(x_i) = \frac{1}{n}\phi_i^*(-x_i) - \frac{\gamma}{2n}\|x_i\|_2^2.$$

Each iteration of Algorithm 5 involves only the two inner products $A_{i_k}^T p^{(k)}$ and $A_{i_k}^T q^{(k)}$ in computing $\nabla_{i_k}^{(k)}$ and the two vector additions in (5.19). They all cost $O(d)$ rather than $O(n)$. When the A_i's are sparse (the case of most large-scale problems), these operations can be carried out very efficiently. Basically, each iteration of Algorithm 5 only costs twice as much as that of SDCA [10, 40].

In step 3 of Algorithm 5, the division by ρ^{k+1} in updating $u^{(k)}$ and $p^{(k)}$ may cause numerical problems because $\rho^{k+1} \to 0$ as the number of iterations k gets large. To fix this issue, we notice that $u^{(k)}$ and $p^{(k)}$ are always accessed in Algorithm 5 in the forms of $\rho^{k+1}u^{(k)}$ and $\rho^{k+1}p^{(k)}$. So we can replace $u^{(k)}$ and $p^{(k)}$ by

$$\bar{u}^{(k)} = \rho^{k+1}u^{(k)}, \qquad \bar{p}^{(k)} = \rho^{k+1}p^{(k)},$$

which can be updated without numerical problems. To see this, we have

$$\bar{u}^{(k+1)} = \rho^{k+2}u^{(k+1)} = \rho^{k+2}\left(u^{(k)} - \frac{1-n\alpha}{2\rho^{k+1}}U_{i_k}h_{i_k}^{(k)}\right) = \rho\left(\bar{u}^{(k)} - \frac{1-n\alpha}{2}U_{i_k}h_{i_k}^{(k)}\right).$$

Similarly, we have

$$\bar{p}^{(k+1)} = \rho\left(\bar{p}^{(k)} - \frac{1-n\alpha}{2}A_{i_k}h_{i_k}^{(k)}\right).$$

5.3. Numerical experiments. In our experiments, we solve the regularized ERM problem (5.1) with a smoothed hinge loss for binary classification. Specifically, we premultiply each feature vector A_i by its label $b_i \in \{\pm 1\}$ and let

$$\phi_i(a) = \begin{cases} 0 & \text{if } a \geq 1, \\ 1 - a - \frac{\gamma}{2} & \text{if } a \leq 1-\gamma, \\ \frac{1}{2\gamma}(1-a)^2 & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n.$$

This smoothed hinge loss is obtained by adding a strong convex perturbation to the conjugate function of the hinge loss (e.g., [40]). The resulting conjugate function of ϕ_i is $\phi_i^*(b) = b + \frac{\gamma}{2}b^2$ if $b \in [-1, 0]$ and ∞ otherwise. Therefore we have

$$\Psi_i(x_i) = \frac{1}{n}\left(\phi_i^*(-x_i) - \frac{\gamma}{2}\|x_i\|_2^2\right) = \begin{cases} \frac{-x_i}{n} & \text{if } x_i \in [0,1], \\ \infty & \text{otherwise.} \end{cases}$$

For the regularization term, we use $g(w) = \frac{1}{2}\|w\|_2^2$. We used three publicly available datasets obtained from [7]. Their characteristics are summarized in Table 1.

In our experiments, we comparing the APCG method (Algorithm 5) with SDCA [40] and the AFG method [25] with an additional line search procedure to improve efficiency. For AFG, each iteration involves a single pass over the whole dataset. When the regularization parameter λ is not too small (around 10^{-4}), then APCG performs similarly to SDCA, as predicted by our complexity results, and they both outperform AFG by a substantial margin.

Figure 1 shows the reduction of primal optimality $P(w^{(k)}) - P^\star$ by the three methods in the ill-conditioned setting, with λ varying from 10^{-5} to 10^{-8}. For APCG, the primal points $w^{(k)}$ are generated simply as $w^{(k)} = \omega(x^{(k)})$ defined in (5.8). Here we see that APCG has superior performance in reducing the primal objective value compared with SDCA and AFG, even without performing the final proximal full gradient step described in Theorem 5.5.

Figure 2 shows the reduction of primal-dual gap $P(w^{(k)}) - D(x^{(k)})$ by the two methods APCG and SDCA. We can see that in the ill-conditioned setting, the APCG method is more effective in reducing the primal-dual gap as well.

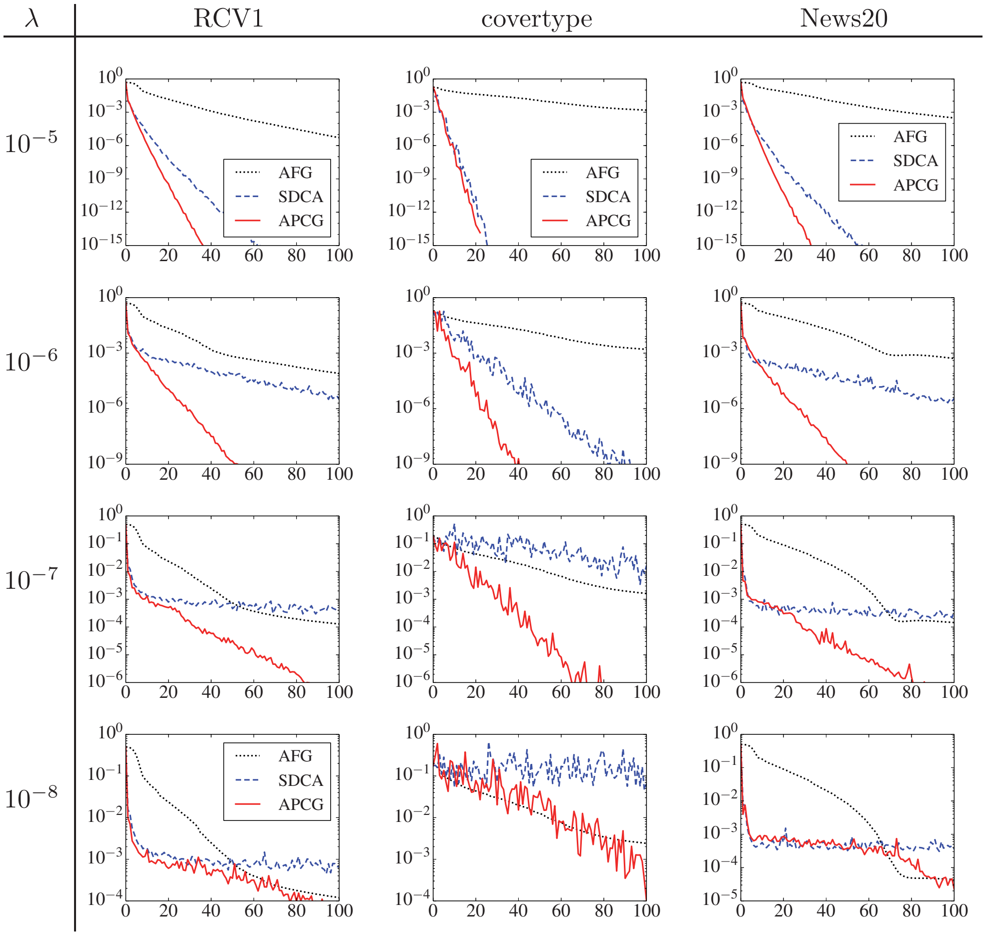

FIG. 1. *Comparing the APCG method with SDCA and the AFG method. In each plot, the vertical axis is the primal objective value gap, i.e.,* $P(w^{(k)}) - P^\star$*, and the horizontal axis is the number of passes through the entire dataset. The three columns correspond to the three datasets, and each row corresponds to a particular value of* λ.

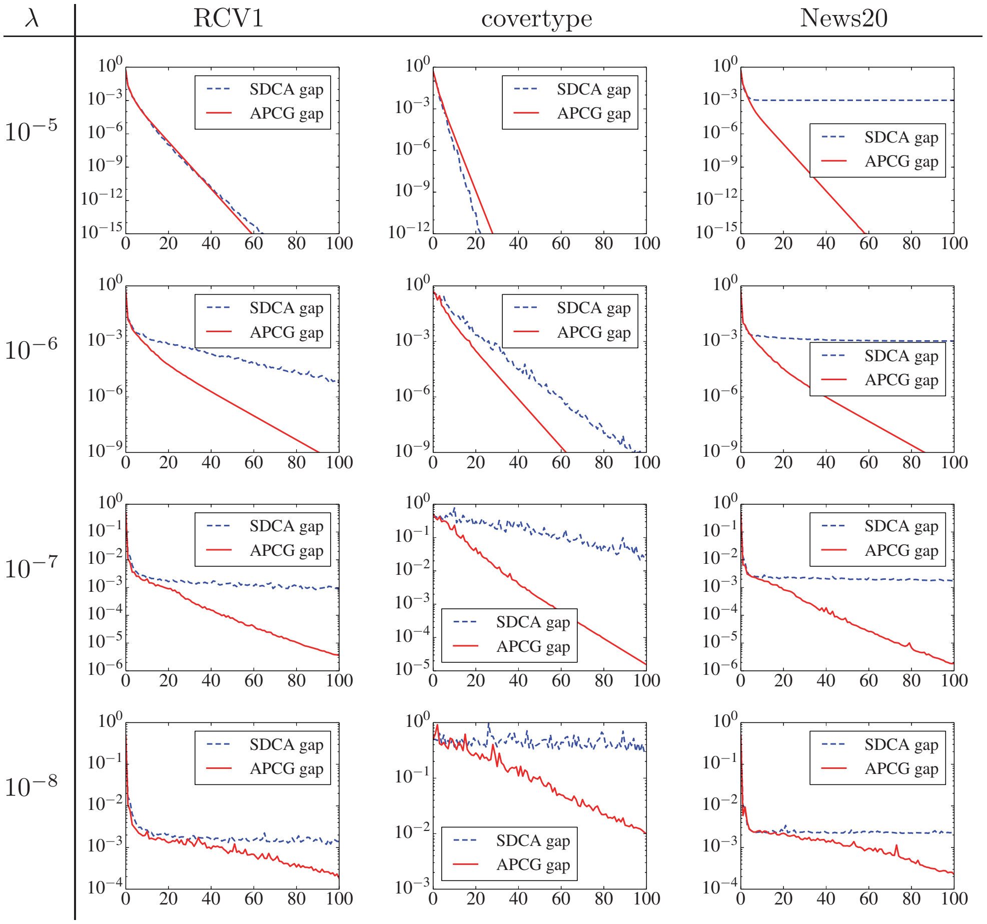

Fig. 2. *Comparing the primal-dual objective gap produced by APCG and SDCA. In each plot, the vertical axis is the primal-dual objective value gap, i.e.,* $P(w^{(k)}) - D(x^{(k)})$*, and the horizontal axis is the number of passes through the entire dataset. The three columns correspond to the three datasets, and each row corresponds to a particular value of* λ*.*

REFERENCES

[1] A. Beck and M. Teboulle, *A fast iterative shrinkage-threshold algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[2] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 13 (2013), pp. 2037–2060.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[4] J. A. Blackard, D. J. Dean, and C. W. Anderson, *Covertype data set*, in UCI Machine Learning Repository, K. Bache and M. Lichman, eds., School of Information and Computer Sciences, University of California, Irvine, http://archive.ics.uci.edu/ml (2013).

[5] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, *Parallel coordinate descent for* l_1*-regularized loss minimization*, in Proceedings of the 28th International Conference on Machine Learning (ICML), 2011, pp. 321–328.

[6] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, *Coordinate descent method for large-scale* l_2*-loss linear support vector machines*, J. Mach. Learn. Res., 9 (2008), pp. 1369–1398.

[7] R.-E. Fan and C.-J. Lin, *LIBSVM Data: Classification, Regression and Multi-label*, http://www.csie.ntu.edu.tw/˜cjlin/libsvmtools/datasets (2011).

[8] O. Fercoq and P. Richtárik, *Accelerated, parallel, and proximal coordinate descent.* SIAM J. Optim., 25 (2015), pp. 1997–2023.

[9] M. Hong, X. Wang, M. Razaviyayn, and Z. Q. Luo, *Iteration complexity analysis of block coordinate descent methods*, arXiv:1310.6957, 2013.

[10] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, *A dual coordinate descent method for large-scale linear svm*, in Proceedings of the 25th International Conference on Machine Learning (ICML), 2008, pp. 408–415.

[11] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems 26, MIT Press, Cambridge, MA, 2013, pp. 315–323.

[12] S. S. Keerthi and D. DeCoste, *A modified finite Newton method for fast solution of large scale linear svms*, J. Mach. Learn. Res., 6 (2005), pp. 341–361.

[13] K. Lang, *Newsweeder: Learning to filter netnews*, in Proceedings of the 12th International Conference on Machine Learning (ICML), 1995, pp. 331–339.

[14] Y. T. Lee and A. Sidford, *Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems*, in Proceedings of IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS), Berkeley, CA, 2013, pp. 147–156.

[15] D. Leventhal and A. S. Lewis, *Randomized methods for linear constraints: Convergence rates and conditioning*, Math. Oper. Res., 35 (2010), pp. 641–654.

[16] D. D. Lewis, Y. Yang, T. Rose, and F. Li, *RCV1: A new benchmark collection for text categorization research*, J. Mach. Learn. Res., 5 (2004), pp. 361–397.

[17] Y. Li and S. Osher, *Coordinate descent optimization for ℓ_1 minimization with application to compressed sensing: A greedy algorithm*, Inverse Probl. Imaging, 3 (2009), pp. 487–503.

[18] J. Liu and S. J. Wright, *An accelerated randomized Kacamarz algorithm*, Math. Comp., 85 (2016), pp. 153–178.

[19] J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar, *An asynchronous parallel stochastic coordinate descent algorithm*, JMLR W&CP, 32 (2014), pp. 469–477.

[20] Z. Lu and L. Xiao, *Randomized block coordinate non-monotone gradient method for a class of nonlinear programming*, arXiv:1306.5918, 2013.

[21] Z. Lu and L. Xiao, *On the complexity analysis of randomized block-coordinate descent methods*, Math. Program. Ser. A, 152 (2015), pp. 615–642.

[22] Z. Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (2002), pp. 7–35.

[23] I. Necoara and D. Clipici, *Parallel Random Coordinate Descent Method for Composite Minimization*, Technical Report 1-41, University Politehnica Bucharest, 2013.

[24] I. Necoara and A. Patrascu, *A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints*, Comput. Optim. Appl., 57 (2014), pp. 307–377.

[25] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer, Boston, 2004.

[26] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM J. Optim., 22 (2012), pp. 341–362.

[27] Yu. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program. Ser. B, 140 (2013), pp. 125–161.

[28] A. Patrascu and I. Necoara, *Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization*, J. Global Optim., 61 (2015), pp. 19–46.

[29] J. Platt, *Fast training of support vector machine using sequential minimal optimization*, in Advances in Kernel Methods—Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press, Cambridge, MA, 1999, pp. 185–208.

[30] Z. Qin, K. Scheinberg, and D. Goldfarb, *Efficient block-coordinate descent algorithms for the group lasso*, Math. Program. Comput., 5 (2013), pp. 143–169.

[31] P. Richtárik and M. Takáč, *Parallel coordinate descent methods for big data optimization*, Math. Program., DOI:10.1007/s10107-015-0901-6.

[32] P. Richtárik and M. Takáč, *Distributed coordinate descent method for learning with big data*, arXiv:1310.2059, 2013.

[33] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Math. Program., 144 (2014), pp. 1–38.

[34] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[35] N. Le Roux, M. Schmidt, and F. Bach, *A stochastic gradient method with an exponential convergence rate for finite training sets*, in Advances in Neural Information Processing Systems 25, 2012, pp. 2672–2680.

[36] A. Saha and A. Tewari, *On the non-asymptotic convergence of cyclic coordinate descent methods*, SIAM J. Optim., 23 (2013), pp. 576–601.

[37] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing Finite Sums with the Stochastic Average Gradient*, Technical Report HAL 00860051, INRIA, Paris, 2013.

[38] S. Shalev-Shwartz and A. Tewari, *Stochastic methods for ℓ_1 regularized loss minimization*, in Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, Canada, 2009, pp. 929–936.

[39] S. Shalev-Shwartz and T. Zhang, *Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization*, Math. Program., DOI:10.1007/s10107-014-0839-0.

[40] S. Shalev-Shwartz and T. Zhang, *Stochastic dual coordinate ascent methods for regularized loss minimization*, J. Mach. Learn. Res., 14 (2013), pp. 567–599.

[41] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, J. Optim. Theory Appl., 140 (2001), pp. 513–535.

[42] P. Tseng, *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, unpublished manuscript, 2008.

[43] P. Tseng and S. Yun, *Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization*, J. Optim. Theory Appl., 140 (2009), pp. 513–535.

[44] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423.

[45] Z. Wen, D. Goldfarb, and K. Scheinberg, *Block coordinate descent methods for semidefinite programming*, in Handbook on Semidefinite, Conic and Polynomial Optimization: Theory, Algorithms, Software and Applications, M. F. Anjos and J. B. Lasserre, eds., Internat. Ser. Oper. Res. Management Sci. 166, Springer, Berlin, 2012, pp. 533–564.

[46] S. J. Wright, *Accelerated block-coordinate relaxation for regularized optimization*, SIAM J. Optim., 22 (2012), pp. 159–186.

[47] T. Wu and K. Lange, *Coordinate descent algorithms for Lasso penalized regression*, Ann. Appl. Stat., 2 (2008), pp. 224–244.

[48] L. Xiao and T. Zhang, *A proximal stochastic gradient method with progressive variance reduction*, SIAM J. Optim., 24 (2014), pp. 2057–2075.