First-Order Algorithms Converge Faster than $O(1/k)$ on Convex Problems

Ching-pei Lee [1] **Stephen J. Wright** [1]

Abstract

It is well known that both gradient descent and stochastic coordinate descent achieve a global convergence rate of $O(1/k)$ in the objective value, when applied to a scheme for minimizing a Lipschitz-continuously differentiable, unconstrained convex function. In this work, we improve this rate to $o(1/k)$. We extend the result to proximal gradient and proximal coordinate descent on regularized problems to show similar $o(1/k)$ convergence rates. The result is tight in the sense that a rate of $O(1/k^{1+\epsilon})$ is not generally attainable for any $\epsilon > 0$, for any of these methods.

1. Introduction

Consider the unconstrained optimization problem

$$\min_{x} \; f(x), \tag{1}$$

where f has domain in an inner-product space and is convex and L-Lipschitz continuously differentiable for some $L > 0$. We assume throughout that the solution set Ω is nonempty. (Elementary arguments based on the convexity and continuity of f show that Ω is a closed convex set.) Classical convergence theory for gradient descent on this problem indicates a $O(1/k)$ global convergence rate in the function value. Specifically, if

$$x_{k+1} \coloneqq x_k - \alpha_k \nabla f(x_k), \quad k = 0, 1, 2, \ldots, \tag{2}$$

and $\alpha_k \equiv \bar{\alpha} \in (0, 1/L]$, we have

$$f(x_k) - f^* \le \frac{\operatorname{dist}(x_0, \Omega)^2}{2\bar{\alpha}k}, \quad k = 1, 2, \ldots, \tag{3}$$

where f^* is the optimal objective value and $\operatorname{dist}(x, \Omega)$ denotes the distance from x to the solution set. The proof of (3) relies on showing that

$$\begin{aligned} k(f(x_k) - f^*) &\le \sum_{T=1}^{k} (f(x_T) - f^*) \\ &\le \frac{1}{2\bar{\alpha}} \operatorname{dist}(x_0, \Omega)^2, \quad k = 1, 2, \ldots, \end{aligned} \tag{4}$$

where the first inequality utilizes the fact that gradient descent is a descent method (yielding a nonincreasing sequence of function values $\{f(x_k)\}$). We demonstrate in this paper that the bound (3) is not tight, in the sense that $k(f(x_k) - f^*) \to 0$, and thus $f(x_k) - f^* = o(1/k)$. This result is a consequence of the following technical lemma.

Lemma 1. *Let $\{\Delta_k\}$ be a nonnegative sequence satisfying the following conditions:*

1. *$\{\Delta_k\}$ is monotonically decreasing;*
2. *$\{\Delta_k\}$ is summable, that is, $\sum_{k=0}^{\infty} \Delta_k < \infty$.*

Then $k\Delta_k \to 0$, so that $\Delta_k = o(1/k)$.

Our claim about the fixed-step gradient descent method follows immediately by setting $\Delta_k = f(x_k) - f^*$ in Lemma 1. We state the result formally as follows, and prove it at the start of Section 2.

Theorem 2. *Consider (1) with f convex and L-Lipschitz continuously differentiable and nonempty solution set Ω. If the step sizes satisfy $\alpha_k \equiv \bar{\alpha} \in (0, 1/L]$ for all k, then gradient descent (2) generates objective values $f(x_k)$ that converge to f^* at an asymptotic rate of $o(1/k)$.*

This result shows that the $o(1/k)$ rate for gradient descent with a fixed short step size is universal on convex problems, without any additional requirements such as the boundedness of Ω assumed in Bertsekas (2016, Proposition 1.3.3). In the remainder of the paper, we show that this faster rate holds for several other smooth optimization algorithms, including gradient descent with fixed steps in the larger range $(0, 2/L)$, gradient descent with various line-search strategies, and stochastic coordinate descent with arbitrary sampling strategies. We then extend the result to algorithms for regularized convex optimization problems, including proximal gradient and stochastic proximal coordinate descent.

Except for the cases of coordinate descent and proximal coordinate descent which require a finite-dimensional space so

[1] Department of Computer Sciences and Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, Wisconsin, USA. Correspondence to: Ching-pei Lee <ching-pei@cs.wisc.edu>, Stephen J. Wright <swright@cs.wisc.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

that all the coordinates can be processed, our results apply to any inner-product spaces. Assumptions such as bounded solution set, bounded level set, or bounded distance to the solution set, which are commonly assumed in the literature, are all unnecessary. We can remove these assumptions because an implicit regularization property causes the iterates to stay within a bounded area.

In our description, the Euclidean norm is used for simplicity, but our results can be extended directly to any norms induced by an inner product,[1] provided that Lipschitz continuity of ∇f is defined with respect to the corresponding norm and its dual norm.

Related Work. Our work was inspired by Peng et al. (2018, Corollary 2) and Bertsekas (2016, Proposition 1.3.3), which improve convergence for certain algorithms and problems on convex problems in a Euclidean space from $O(1/k)$ to $o(1/k)$ when the level set is compact. Our paper develops improved convergence rates of several algorithms on convex problems without the assumption on the level set, with most of our results applying to non-Euclidean Hilbert spaces. The main proof techniques in this work are somewhat different from those in the works cited here.

For an accelerated version of proximal gradient on convex problems, it is proved in (Attouch & Peypouquet, 2016) that the convergence rate can be improved from $O(1/k^2)$ to $o(1/k^2)$. Accelerated proximal gradient is a more complicated algorithm than the nonaccelerated versions we discuss, and thus Attouch & Peypouquet (2016) require a more complicated analysis that is quite different from ours.

Deng et al. (2017) have stated a version of Lemma 1 with a proof different from the proof that we present in the supplementary material, using it to show the convergence rate of the quantity $\|x_k - x_{k+1}\|$ of a version of the alternating-directions method of multipliers (ADMM). Our work differs in the range of algorithms considered and the nature of the convergence. We also provide a discussion of the tightness of the $o(1/k)$ convergence rate.

2. Main Results on Unconstrained Smooth Problems

We start by detailing the procedure for obtaining (4), to complete the proof of Theorem 2. First, we define

$$M(\alpha) := \alpha - \tfrac{1}{2}L\alpha^2. \tag{5}$$

From the Lipschitz continuity of ∇f, we have for any point x and any real number α that

$$\begin{aligned} & f(x - \alpha\nabla f(x)) \\ & \le f(x) - \nabla f(x)^\top (\alpha \nabla f(x)) + \frac{L}{2}\|\alpha\nabla f(x)\|^2 \\ & = f(x) - M(\alpha)\|\nabla f(x)\|^2. \end{aligned} \tag{6}$$

Clearly,

$$\alpha \in (0, 1/L] \quad \Rightarrow \quad M(\alpha) \ge \tfrac{1}{2}\alpha > 0, \tag{7}$$

so in this case, we have by rearranging (6) that

$$\|\nabla f(x)\|^2 \le \frac{2}{\alpha}\left(f(x) - f(x - \alpha\nabla f(x))\right). \tag{8}$$

Considering any solution $\bar{x} \in \Omega$ and any $T \ge 0$, we have for gradient descent (2) that

$$\begin{aligned} \|x_{T+1} - \bar{x}\|^2 &= \|x_T - \alpha_T\nabla f(x_T) - \bar{x}\|^2 \\ &= \|x_T - \bar{x}\|^2 + \alpha_T^2\|\nabla f(x_T)\|^2 \\ &\quad - 2\alpha_T\nabla f(x_T)^\top (x_T - \bar{x}). \end{aligned} \tag{9}$$

Since $\alpha_T \in (0, 1/L]$ in (9), from (8) and the convexity of f (implying $\nabla f(x_T)^T(\bar{x} - x_T) \le f^* - f(x_T)$), we have

$$\begin{aligned} \|x_{T+1} - \bar{x}\|^2 \le \|x_T - \bar{x}\|^2 + 2\alpha_T\left(f\left(x_T\right) - f\left(x_{T+1}\right)\right) \\ + 2\alpha_T\left(f^* - f\left(x_T\right)\right). \end{aligned} \tag{10}$$

By rearranging (10) and using $\alpha_T \equiv \bar{\alpha} \in (0, 1/L]$,

$$f\left(x_{T+1}\right) - f^* \le \frac{1}{2\bar{\alpha}}\left(\|x_T - \bar{x}\|^2 - \|x_{T+1} - \bar{x}\|^2\right). \tag{11}$$

We then obtain (4) by summing (11) from $T = 0$ to $T = k - 1$ and noticing that $\bar{x}$ is arbitrary in Ω.

Theorem 2 applies to step sizes in the range $(0, 1/L]$ only, but it is known that gradient descent converges at the rate of $O(1/k)$ for both the fixed step size scheme with $\bar{\alpha} \in (0, 2/L)$ and line-search schemes. Next, we show that $o(1/k)$ rates hold for these variants too. We then extend the result to stochastic coordinate descent with arbitrary sampling of coordinates.

2.1. Gradient Descent with Longer Steps

In this subsection, we allow the steplengths α_k for (2) to vary from iteration to iteration, according to the following conditions, for some $\gamma \in (0, 1]$:

$$\alpha_k \in [C_2, C_1],\ C_2 \in \left(0, \frac{2-\gamma}{L}\right],\ C_1 \ge C_2, \tag{12a}$$

$$f\left(x_{k+1}\right) \le f\left(x_k\right) - \frac{\gamma\alpha_k}{2}\|\nabla f(x_k)\|^2. \tag{12b}$$

[1] We meant that given an inner product $< \cdot, \cdot >$, the norm $\|\cdot\|$ is defined as $\|x\| := \sqrt{< x, x >}$.

Note that these conditions encompass a fixed-steplength strategy with $\alpha_k \equiv C_2$ as a special case, by setting $C_1 = C_2$, and noting that condition (12b) is a consequence of (6). (Note too that $\alpha_k \equiv C_2 \in (0, (2-\gamma)/L]$ can be almost twice as large as the bound $1/L$ considered above.)

The main result for this subsection is as follows.

Theorem 3. *Consider (1) with f convex and L-Lipschitz continuously differentiable and nonempty solution set Ω. If the step sizes α_k satisfy (12), then gradient descent (2) generates objective values $f(x_k)$ converging to f^* at an asymptotic rate of $o(1/k)$.*

We give two alternative proofs of this result to provide different insights. The first proof is similar to the one we presented for Theorem 2 at the start of this section. The second proof holds only for Euclidean spaces. This proof improves the standard proof of Nesterov (2004, Section 2.1.5).

We start from the following lemma, which verifies that the iterates remain in a bounded set and is used in both proofs.

Lemma 4. *Consider algorithm (2) with any initial point x_0, and assume that f is convex and L-Lipschitz-continuously differentiable for some $L > 0$. Then when the sequence of steplengths α_k is chosen to satisfy (12), all iterates x_k lie in a bounded set. In particular, for any $\bar{x} \in \Omega$ and any $k \geq 0$, we have that*

$$\begin{aligned}\|x_{k+1}-\bar{x}\|^2 &\leq \|x_0-\bar{x}\|^2 + \frac{2C_1}{\gamma}\left(f(x_0) - f(x_{k+1})\right)\\ &\quad + 2C_2\sum_{T=0}^{k}\left(f^* - f(x_T)\right) \qquad (13)\\ &\leq \|x_0-\bar{x}\|^2 + \frac{2C_1}{\gamma}\left(f(x_0)-f^*\right). \qquad (14)\end{aligned}$$

Proof. By (12b) and the convexity of f, (9) further implies that for any $T \geq 0$,

$$\begin{aligned}&\|x_{T+1}-\bar{x}\|^2 - \|x_T-\bar{x}\|^2 \qquad (15)\\ &\leq \frac{2\alpha_T}{\gamma}\left(f(x_T) - f(x_{T+1})\right) + 2\alpha_T\left(f^* - f(x_T)\right).\end{aligned}$$

We know that the first term is nonnegative from (12b), while the second term is nonpositive from the optimality of f^*. Therefore, (15) implies

$$\begin{aligned}&\|x_{T+1}-\bar{x}\|^2 - \|x_T-\bar{x}\|^2 \qquad (16)\\ &\leq \frac{2C_1}{\gamma}\left(f(x_T) - f(x_{T+1})\right) + 2C_2\left(f^* - f(x_T)\right).\end{aligned}$$

We then obtain (13) by summing (16) for $T = 0, 1, \dots, k$ and telescoping. By noting that $f(x_k) \geq f^*$ for all k, (14) follows. □

The first proof of Theorem 3 is as follows.

First Proof of Theorem 3. We again consider Lemma 1 with $\Delta_k := f(x_k) - f^*$, which is always nonnegative from the optimality of f^*. Monotonicity is clear from (12b), so we just need to show summability. By rearranging (13) and noting $f(x_{k+1}) \geq f^*$, we obtain

$$\begin{aligned}2C_2\sum_{T=0}^{k}\Delta_T &\leq \|x_0-\bar{x}\|^2 - \|x_{k+1}-\bar{x}\|^2 + \frac{2C_1}{\gamma}\Delta_0\\ &\leq \|x_0-\bar{x}\|^2 + \frac{2C_1}{\gamma}\Delta_0.\end{aligned}$$ □

For the second proof of Theorem 3, we first outline the analysis from Nesterov (2004, Section 2.1.5) and then show how it can be modified to produce the desired $o(1/k)$ rate. Denote by $\bar{x}_T$ the projection of x_T onto Ω (which is well defined because Ω is nonempty, closed, and convex). We can utilize the convexity of f to obtain

$$\Delta_T \leq \nabla f(x_T)^\top (x_T - \bar{x}_T) \leq \|\nabla f(x_T)\| \text{dist}(x_T, \Omega),$$

so that

$$\|\nabla f(x_T)\| \geq \frac{\Delta_T}{\text{dist}(x_T,\Omega)}. \qquad (17)$$

By subtracting f^* from both sides of (12b) and using $\alpha_k \geq C_2$ and (17), we obtain

$$\Delta_{T+1} \leq \Delta_T - \frac{C_2\gamma\Delta_T^2}{2\text{dist}(x_T,\Omega)^2}.$$

By dividing both sides of this expression by $\Delta_T\Delta_{T+1}$ and using $\Delta_{T+1} \leq \Delta_T$, we obtain

$$\begin{aligned}\frac{1}{\Delta_{T+1}} &\geq \frac{1}{\Delta_T} + \frac{C_2\gamma\Delta_T}{2\text{dist}(x_T,\Omega)^2\Delta_{T+1}}\\ &\geq \frac{1}{\Delta_T} + \frac{C_2\gamma}{2\text{dist}(x_T,\Omega)^2}. \qquad (18)\end{aligned}$$

By summing (18) over $T = 0, 1, \dots, k-1$, we obtain

$$\begin{aligned}&\frac{1}{\Delta_k} \geq \frac{1}{\Delta_0} + \sum\nolimits_{T=0}^{k-1}\frac{C_2\gamma}{2\text{dist}(x_T,\Omega)^2}\\ &\Rightarrow \Delta_k \leq \frac{1}{\sum_{T=0}^{k-1}\frac{C_2\gamma}{2\text{dist}(x_T,\Omega)^2}}. \qquad (19)\end{aligned}$$

A $O(1/k)$ rate is obtained by noting from Lemma 4 that $\text{dist}(x_T,\Omega) \leq R_0$ for some $R_0 > 0$ and all T, so that

$$\sum\nolimits_{T=0}^{k-1}\frac{1}{\text{dist}(x_T,\Omega)^2} \geq \frac{k}{R_0^2}. \qquad (20)$$

Our alternative proof uses the fact that (20) is a loose bound for Euclidean spaces and that an improved result can be obtained by working directly with (19). We first use the Bolzano-Weierstrass theorem (a bounded and closed set is sequentially compact in a Euclidean space) together with Lemma 4, to show that the sequence $\{x_k\}$ approaches the solution set Ω.

Lemma 5. *Assume the conditions in Lemma 4 and in addition that f has domain in a Euclidean space $f : \Re^n \to \Re$. We have*

$$\lim_{k \to \infty} \text{dist}\,(x_k, \Omega) = 0. \tag{21}$$

Proof. The proof is similar to (Peng et al., 2018, Proposition 1). Assume for contradiction that (21) does not hold. Then there are $\epsilon > 0$ and an infinite increasing sequence $\{k_i\}$, $i = 1, 2, \dots$, such that

$$\text{dist}\,(x_{k_i}, \Omega) \geq \epsilon, \quad i = 1, 2, \dots . \tag{22}$$

From Lemma 4 and that $\{x_{k_i}\} \subset \Re^n$, the sequence $\{x_{k_i}\}$ lies in a compact set and therefore has an accumulation point x^*. From (18), we have $1/\Delta_{k_{i+1}} \geq 1/\Delta_{k_i} + C_2\gamma/(2\epsilon^2)$, so that $1/\Delta_k \uparrow \infty$ and hence $\Delta_k \downarrow 0$. By continuity of f, it follows that $f(x^*) = f^*$, so that $x^* \in \Omega$ by definition, contradicting (22). □

We note that a result similar to Lemma 5 has been given in (Burachik et al., 1995) using a more complicated argument with more restricted choices of α.

Second Proof of Theorem 3, for Euclidean Spaces. We start with (19) and show that

$$\lim_{k \to \infty} \frac{\frac{1}{\frac{C_2\gamma}{2} \sum_{T=0}^{k-1} \frac{1}{\text{dist}(x_T, \Omega)^2}}}{\frac{1}{k}} = 0,$$

or, equivalently,

$$\lim_{k \to \infty} \frac{k}{\sum_{T=0}^{k-1} \frac{1}{\text{dist}(x_T, \Omega)^2}} = 0. \tag{23}$$

From the arithmetic-mean / harmonic-mean inequality,[2] we have that

$$0 \leq \frac{k}{\sum_{T=0}^{k-1} \frac{1}{\text{dist}(x_T, \Omega)^2}} \leq \frac{\sum_{T=0}^{k-1} \text{dist}(x_T, \Omega)^2}{k}. \tag{24}$$

Lemma 5 shows that $\text{dist}(x_T, \Omega) \to 0$, so by the Stolz-Cesàro theorem (see, for example, (Mureşan, 2009)), the right-hand side of (24) converges to 0. Therefore, from the sandwich lemma, (23) holds. □

2.2. Coordinate Descent

We now extend Theorem 2 to the case of randomized coordinate descent. Our results can extend immediately to block-coordinate descent with fixed blocks. Our analysis for coordinate descent requires Euclidean spaces so that coordinate descent can go through all coordinates.

The standard short-step coordinate descent procedure requires knowledge of coordinate-wise Lipschitz constants. Denoting by e_i the ith unit vector, we denote by $L_i \geq 0$ the constants such that:

$$\begin{gathered} \left|\nabla_i f(x) - \nabla_i f(x + h e_i)\right| \leq L_i \left|h\right|, \\ \text{for all } x \in \Re^n \text{ and all } h \in \Re, \end{gathered} \tag{25}$$

where $\nabla_i f(\cdot)$ denotes the ith coordinate of the gradient. Note that if $\nabla f(x)$ is L-Lipschitz continuous, there always exist $L_1, \dots, L_n \in [0, L]$ such that (25) holds. Without loss of generality, we assume $L_i > 0$ for all i. Given parameters $\{\bar{L}_i\}_{i=1}^n$ such that $\bar{L}_i \geq L_i$ for all i, the coordinate descent update is

$$x_{k+1} \leftarrow x_k - \frac{\nabla_{i_k} f\left(x_k\right)}{\bar{L}_{i_k}} e_{i_k}, \tag{26}$$

where i_k is the coordinate selected for updating at the kth iteration. We consider the general case of stochastic coordinate descent in which each i_k is independently identically distributed following a fixed prespecified probability distribution $p_1, \dots, p_n$ satisfying

$$p_i \geq p_{\min}, \quad i = 1, 2, \dots, n; \quad \sum_{i=1}^n p_i = 1, \tag{27}$$

for some constant $p_{\min} > 0$. Nesterov (2012) proves that stochastic coordinate descent has a $O(1/k)$ convergence rate (in expectation of f) on convex problems. We show below that this rate can be improved to $o(1/k)$.

Theorem 6. *Consider (1) with f convex and nonempty solution set Ω, and that (25) holds with some $L_1, \dots, L_n > 0$. If we apply coordinate descent (26) and at each iteration, i_k is independently picked at random following a probability distribution satisfying (27), then the expected objective $\mathbb{E}_{i_0, i_1, \dots, i_{k-1}}[f(x_k)]$ converges to f^* at an asymptotic rate of $o(1/k)$.*

Proof. From (25) and that $\bar{L}_i \geq L_i$, by treating all other coordinates as non-variables, we have that for any $T \geq 0$,

$$f\left(x_T - \frac{\nabla_i f\left(x_T\right)}{\bar{L}_i} e_i\right) - f\left(x_T\right) \leq -\frac{\left\|\nabla_i f\left(x_T\right)\right\|^2}{2\bar{L}_i}, \forall i, \tag{28}$$

showing that the algorithm decreases f at each iteration. Consider any $\bar{x} \in \Omega$, by defining

$$r_T^2 \coloneqq \sum\nolimits_{i=1}^n \frac{\bar{L}_i}{p_i} \left\|\left(x_T - \bar{x}\right)_i\right\|^2, \tag{29}$$

[2] For any real numbers $a_1, \dots, a_n > 0$, their harmonic mean does not exceed their arithmetic mean. Namely,

$$\frac{n}{\sum_{i=1}^n a_i^{-1}} \leq \frac{\sum_{i=1}^n a_i}{n}.$$

we have from (26) that

$$r_{T+1}^2 = r_T^2 + \frac{\|\nabla_{i_T} f(x_T)\|^2}{\bar{L}_{i_T} p_{i_T}} - \frac{2\nabla_{i_T} f(x_T)^\top (x_T - \bar{x})_{i_T}}{p_{i_T}}.$$

By taking expectation over i_T on both sides of the above expression, we obtain from the convexity of f and (28) that

$$\begin{aligned} &\frac{1}{2}\left(\mathbb{E}_{i_T}\left[r_{T+1}^2\right] - r_T^2\right) \\ &\overset{(28)}{\leq} \frac{1}{p_{\min}} \sum\nolimits_{i=1}^{n} p_i \left(f(x_T) - f\left(x_T - \frac{\nabla_i f(x_T)}{\bar{L}_i} e_i \right)\right) \\ &\quad - \nabla f(x_T)^\top (x_T - \bar{x}) \\ &\leq \frac{f(x_T) - \mathbb{E}_{i_T}[f(x_{T+1})]}{p_{\min}} + (f^* - f(x_T)). \end{aligned} \tag{30}$$

By taking expectation over $i_0, i_1, \ldots, i_{T-1}$ on (30) and summing (30) over $T = 0, 1, \ldots, k$, we obtain

$$\begin{aligned} &2\sum\nolimits_{T=0}^{k} \left(\mathbb{E}_{i_0,\ldots,i_{T-1}}[f(x_T)] - f^*\right) \\ &\leq r_0^2 - \mathbb{E}_{i_0,\ldots,i_k}\left[r_{k+1}^2\right] + \frac{2\left(f(x_0) - \mathbb{E}_{i_0,\ldots,i_k}[f(x_{k+1})]\right)}{p_{\min}} \\ &\leq r_0^2 + \frac{2\left(f(x_0) - f^*\right)}{p_{\min}}. \end{aligned}$$

The result now follows from Lemma 1. □

3. Regularized Problems

We turn to regularized optimization in an inner-product space:

$$\min_x \; F(x) := f(x) + \psi(x), \tag{31}$$

where both terms are convex, f is L-Lipschitz-continuously differentiable, and ψ is extended-valued, proper, and closed, but possibly nondifferentiable. We also assume that ψ is such that the prox-operator can be applied easily, by solving the following problem for any given y and any $\lambda > 0$:

$$\min_x \; \psi(x) + \frac{1}{2\lambda}\|x - y\|^2.$$

We assume further that the solution set Ω of (31) is nonempty, and denote by F^* the value of F for all $x \in \Omega$. We discuss two algorithms to show how our techniques can be extended to regularized problems. They are proximal gradient (both with and without line search) and stochastic proximal coordinate descent with arbitrary sampling.

3.1. Short-Step Proximal Gradient

Given $\bar{L} \geq L$, the kth step of the proximal gradient algorithm is defined as follows:

$$\begin{aligned} &x_{k+1} \leftarrow x_k + d_k, \\ &d_k := \arg\min_d \; \nabla f(x_k)^\top d + \frac{\bar{L}}{2}\|d\|^2 + \psi(x_k + d). \end{aligned} \tag{32}$$

Note that d_k is uniquely defined here, since the subproblem is strongly convex. It is shown in (Beck & Teboulle, 2009; Nesterov, 2013) that $F(x_k)$ converges to F^* at a rate of $O(1/k)$ for this algorithm, under our assumptions. We prove that a $o(1/k)$ rate can be attained.

Theorem 7. *Consider (31) with f convex and L-Lipschitz continuously differentiable, ψ convex, and nonempty solution set Ω. Given any $\bar{L} \geq L$, the proximal gradient method (32) generates iterates whose objective value converges to F^* at a $o(1/k)$ rate.*

Proof. The method (32) can be shown to be a descent method from the Lipschitz continuity of ∇f and the fact that $\bar{L} \geq L$. From the optimality of the solution to (32) and that $x_{k+1} = x_k + d_k$,

$$-\left(\nabla f(x_k) + \bar{L} d_k\right) \in \partial\psi(x_{k+1}), \tag{33}$$

where $\partial\psi$ denotes the subdifferential of ψ. Consider any $\bar{x} \in \Omega$. We have from (32) that for any $T \geq 0$, the following chain of relationships holds:

$$\begin{aligned} &\|x_{T+1} - \bar{x}\|^2 - \|x_T - \bar{x}\|^2 \\ &= 2d_T^\top (x_T - \bar{x}) + \|d_T\|^2 \\ &= 2d_T^\top (x_T + d_T - \bar{x}) - \|d_T\|^2 \\ &= 2\left(d_T + \frac{\nabla f(x_T)}{\bar{L}}\right)^\top (x_{T+1} - \bar{x}) \\ &\quad - \frac{2}{\bar{L}}\nabla f(x_T)^\top (x_T + d_T - \bar{x}) - \|d_T\|^2 \\ &\overset{(33)}{\leq} 2\frac{\psi(\bar{x}) - \psi(x_{T+1})}{\bar{L}} - \frac{2}{\bar{L}}\nabla f(x_T)^\top (x_T - \bar{x}) \\ &\quad - \frac{2}{\bar{L}}\nabla f(x_T)^\top d_T - \|d_T\|^2 \\ &\leq \frac{2}{\bar{L}}\left(\psi(\bar{x}) - \psi(x_{T+1})\right) + \frac{2f(\bar{x})}{\bar{L}} \\ &\quad - \frac{2}{\bar{L}}\left(f(x_T) + \nabla f(x_T)^\top d_T + \frac{\bar{L}\|d_T\|^2}{2}\right) \\ &\leq \frac{2\left(F^* - F(x_{T+1})\right)}{\bar{L}}, \end{aligned} \tag{34}$$

where in the last inequality, we have used

$$\begin{aligned} f(x + d) &\leq f(x) + \nabla f(x)^\top d + \frac{L}{2}\|d\|^2 \\ &\leq f(x) + \nabla f(x)^\top d + \frac{\bar{L}}{2}\|d\|^2. \end{aligned} \tag{35}$$

By rearranging (34), we obtain

$$F(x_{T+1}) - F^* \leq \frac{\bar{L}}{2}\left(\|x_T - \bar{x}\|^2 - \|x_{T+1} - \bar{x}\|^2\right).$$

The result follows by summing both sides of this expression over $T = 0, 1, \ldots, k-1$ and applying Lemma 1. □

3.2. Proximal Gradient with Line Search

We discuss a line-search variant of proximal gradient, where the update is defined as follows:

$$\begin{aligned} &x_{k+1} \leftarrow x_k + d_k, \\ &d_k := \arg\min\nolimits_d \ \nabla f(x_k)^\top d + \frac{\|d\|^2}{2\alpha_k} + \psi\left(x_k + d\right), \end{aligned} \tag{36}$$

where α_k is chosen such that for given $\gamma \in (0, 1]$ and $C_1 \geq C_2 > 0$ defined as in (12a), we have

$$\alpha_k \in [C_2, C_1],\ F\left(x_k + d_k\right) \leq F\left(x_k\right) - \frac{\gamma}{2\alpha_k}\|d_k\|^2. \tag{37}$$

This framework is a generalization of that in Section 2.1, and includes the SpaRSA algorithm of Wright et al. (2009), which obtains an initial choice of α_k from a Barzilai-Borwein approach and adjusts it until (37) holds. The approach of the previous subsection can also be seen as a special case of (36)-(37) through the following elementary result, whose proof is omitted.

Lemma 8. *Consider a convex function ψ, a positive scalar $a > 0$ and two vectors b and x. If d is the unique solution of the strictly convex problem*

$$\min\nolimits_d \ b^\top d + \frac{a}{2}\|d\|^2 + \psi(x + d),$$

then

$$b^\top d + \frac{a}{2}\|d\|^2 + \psi(x + d) - \psi(x) \leq -\frac{a}{2}\|d\|^2.$$

By setting $b = \nabla f(x)$, $1/\alpha_k \equiv a = \bar{L} > 0$ (where $\bar{L} \geq L$), this lemma together with (35) implies that (37) holds for any $\gamma \in (0, 1]$. Moreover, it also implies that for any $k \geq 0$,

$$\begin{aligned} &F\left(x_{k+1}\right) - F\left(x_k\right) \\ &\overset{(35)}{\leq} \nabla f\left(x_k\right)^\top d_k + \frac{1}{2\alpha_k}\left\|d_k\right\|^2 + \psi\left(x_k + d_k\right) - \psi\left(x_k\right) \\ &\qquad + \left(\frac{L}{2} - \frac{1}{2\alpha_k}\right)\left\|d_k\right\|^2 \\ &\leq -\left(\frac{1}{\alpha_k} - \frac{L}{2}\right)\left\|d_k\right\|^2. \end{aligned}$$

Therefore, for any $\gamma \in (0, 1]$, (37) holds whenever $\alpha > 0$ and $\gamma/(2\alpha_k) \leq 1/\alpha_k - L/2$, or equivalently $\alpha_k \in (0, (2 - \gamma)/L]$, which is how the upper bound for C_2 is set.

We show now that this approach also has a $o(1/k)$ convergence rate on convex problems.

Theorem 9. *Consider (31) with f convex and L-Lipschitz continuously differentiable, ψ convex, and nonempty solution set Ω. Given some $\gamma \in (0, 1]$ and C_2 and C_1 such that $C_1 \geq C_2$ and $C_2 \in (0, (2 - \gamma)/L]$, then the algorithm (36) with α_k satisfying (37) generates iterates $\{x_k\}$ whose objective values converge to F^* at a rate of $o(1/k)$. Moreover, the sequence of iterates is bounded.*

Proof. From the optimality conditions of (36), we have

$$-\left(\nabla f(x_T) + \frac{1}{\alpha_T}d_T\right) \in \partial\psi\left(x_{T+1}\right). \tag{38}$$

Now consider any $\bar{x} \in \Omega$. We have from (36) that for any $T \geq 0$, the following chain of relationships holds:

$$\begin{aligned} &\left\|x_{T+1} - \bar{x}\right\|^2 - \left\|x_T - \bar{x}\right\|^2 \\ &= 2d_T^\top\left(x_T + d_T - \bar{x}\right) - \left\|d_T\right\|^2 \\ &= 2\left(d_T + \alpha_T\nabla f(x_T)\right)^\top\left(x_{T+1} - \bar{x}\right) \\ &\qquad - 2\alpha_T\nabla f(x_T)^\top\left(x_T + d_T - \bar{x}\right) - \left\|d_T\right\|^2 \\ &\overset{(38)}{\leq} 2\alpha_T\left(\psi\left(\bar{x}\right) - \psi\left(x_{T+1}\right)\right) \\ &\qquad - 2\alpha_T\nabla f(x_T)^\top\left(x_T - \bar{x}\right) - 2\alpha_T\nabla f(x_T)^\top d_T \\ &= 2\alpha_T\left(\psi\left(\bar{x}\right) - \psi\left(x_{T+1}\right)\right) - 2\alpha_T\nabla f(x_T)^\top\left(x_T - \bar{x}\right) \\ &\qquad - 2\alpha_T\nabla f(x_T)^\top d_T + \alpha_T L\|d_T\|^2 - \alpha_T L\|d_T\|^2 \\ &\leq 2\alpha_T\left(\psi\left(\bar{x}\right) - \psi\left(x_{T+1}\right) + f\left(\bar{x}\right)\right) \\ &\qquad - 2\alpha_T\left(\left(f\left(x_T\right) + \nabla f(x_T)^\top d_T + \frac{L}{2}\|d_T\|^2\right)\right) \\ &\qquad + \alpha_T L\|d_T\|^2 \\ &\overset{(37)}{\leq} 2\alpha_T\left(F^* - F\left(x_{T+1}\right)\right) + \frac{2L\alpha_T^2}{\gamma}\left(F(x_T) - F(x_{T+1})\right) \\ &\leq 2C_2\left(F^* - F\left(x_{T+1}\right)\right) \\ &\qquad + \frac{2LC_1^2}{\gamma}\left(F\left(x_T\right) - F\left(x_{T+1}\right)\right). \end{aligned} \tag{39}$$

By rearrangement, of this inequality, we obtain

$$\begin{aligned} F(x_{T+1}) - F^* \leq{} & \frac{LC_1^2}{\gamma C_2}(F(x_T) - F(x_{T+1})) \\ & + \frac{1}{2C_2}\left(\|x_T - \bar{x}\|^2 - \|x_{T+1} - \bar{x}\|^2\right), \end{aligned}$$

and by summing both sides and using telescoping sums, we find that $\sum_{T=0}^{\infty}(F(x_{T+1}) - F^*) < \infty$, thus the conditions of Lemma 1 are satisfied by $\Delta_T := F(x_T) - F^*$, and the $o(1/k)$ rate follows.

By summing the inequality above finitely over $T = 0, 1, \ldots, k-1$, we obtain

$$\begin{aligned} 0 &\leq \sum\nolimits_{T=0}^{k-1}(F(x_{T+1}) - F^*) \\ &\leq \frac{LC_1^2}{\gamma C_2}(F(x_0) - F^*) + \frac{1}{2C_2}\left(\|x_0 - \bar{x}\|^2 - \|x_k - \bar{x}\|^2\right). \end{aligned}$$

By rearranging this inequality, we obtain a uniform upper bound on $\|x_k - \bar{x}\|$, thus showing that the sequence $\{x_k\}$ is bounded. □

3.3. Proximal Coordinate Descent

We discuss the extension of coordinate descent to (31), with the assumption (25) on f, Euclidean domain of dimension

n, sampling weighted according to (27) , and the additional assumption of separability of the regularizer ψ, that is,

$$\psi(x) = \sum\nolimits_{i=1}^{n} \psi_i(x_i), \tag{40}$$

where each ψ_i is convex, extended valued, and possibly nondifferentiable. As in our discussion of Section 2.2, the results in this subsection can be extended directly to the case of block-coordinate descent.

Given the component-wise Lipschitz constants $\{L_i\}_{i=1}^n$ and algorithmic parameters $\{\bar{L}_i\}_{i=1}^n$ with $\bar{L}_i \geq L_i$ for all i, proximal coordinate descent updates have the form

$$\begin{aligned} & x_{k+1} \leftarrow x_k + d_{i_k}^k e_{i_k}, \\ & d_{i_k}^k := \arg\min_{d\in\Re} \nabla_{i_k} f(x_k) d + \frac{\bar{L}_{i_k}}{2} d^2 + \psi_{i_k}\left((x_k)_{i_k} + d\right). \end{aligned} \tag{41}$$

With $p_i \equiv 1/n$ for all i, Lu & Xiao (2015) showed that the expected objective value converges to F^* at a $O(1/k)$ rate. When arbitrary sampling (27) is considered, (41) is a special case of the general algorithmic framework described in (Lee & Wright, 2018). The latter paper shows the same $O(1/k)$ rate for convex problems under the additional assumption that for any x_0, we have

$$\max\nolimits_{x: F(x) \leq F(x_0)} \text{dist}\,(x, \Omega) < \infty. \tag{42}$$

We show here that with arbitrary sampling according to (27), (41) produces $o(1/k)$ convergence rates for the expected objective on convex problems, without the assumption (42).

The following result makes use of the quantity r_k defined in (29).

Theorem 10. *Consider (31) with f and ψ convex and nonempty solution set Ω. Assume further that (40) is true, and that (25) holds with some $L_1, L_2, \dots, L_n > 0$. Given $\{\bar{L}_i\}_{i=1}^n$ with $\bar{L}_i \geq L_i$ for all i, suppose that proximal coordinate descent defines iterates according to (41), with i_k chosen i.i.d. according to a probability distribution satisfying (27). Then $\mathbb{E}_{i_0, i_1, \dots, i_{k-1}}[F(x_k)]$ converges to F^* at an asymptotic rate of $o(1/k)$. Moreover, given any $\bar{x} \in \Omega$, the sequence of $\mathbb{E}_{i_0, \dots, i_{k-1}} r_k^2$ is bounded.*

Proof. From (25), we first notice that in the update (41),

$$\begin{aligned} & F\left(x_k + d_{i_k}^k e_{i_k}\right) - F\left(x_k\right) \\ & \leq \nabla_{i_k} f(x_k) d_{i_k}^k + \frac{\bar{L}_{i_k}}{2}\left(d_{i_k}^k\right)^2 \\ & \quad + \psi_{i_k}\left(\left(x_k\right)_{i_k} + d_{i_k}^k\right) - \psi_{i_k}\left(\left(x_k\right)_{i_k}\right). \end{aligned} \tag{43}$$

From Lemma 8, the method defined by (41) is a descent method. Optimality of the subproblem in (41) yields

$$-\left(\nabla_{i_T} f\left(x_T\right) + \bar{L}_{i_T} d_{i_T}^T\right) \in \partial \psi_{i_T}\left(\left(x_T\right)_{i_T} + d_{i_T}^T\right). \tag{44}$$

By taking any $\bar{x} \in \Omega$, and using the definition (29), we have:

$$\begin{aligned} & r_{T+1}^2 - r_T^2 \\ & = \frac{2\bar{L}_{i_T}}{p_{i_T}} \left(d_{i_T}^\top\right)^\top \left(x_T + d_{i_T}^T - \bar{x}\right)_{i_T} - \frac{\bar{L}_{i_T}}{p_{i_T}} \left(d_{i_T}^T\right)^2 \\ & = \frac{2}{p_{i_T}} \left(\nabla_{i_T} f\left(x_T\right) + \bar{L}_{i_T} d_{i_T}^T\right)^\top \left(x_T + d_{i_T}^T - \bar{x}\right)_{i_T} \\ & \qquad - \frac{\bar{L}_{i_T}}{p_{i_T}} \left(d_{i_T}^T\right)^2 - \frac{2}{p_{i_T}} \nabla_{i_T} f\left(x_T\right)^\top \left(x_T - \bar{x}\right)_{i_T} \\ & \qquad - \frac{2}{p_{i_T}} \nabla_{i_T} f\left(x_T\right)^\top d_{i_T}^T \\ & \overset{(44)}{\leq} \frac{2}{p_{i_T}} \left(\psi_{i_T}\left(\bar{x}_{i_T}\right) - \psi_{i_T}\left(\left(x_T\right)_{i_T}\right)\right) \\ & \qquad - \frac{2}{p_{i_T}} \left(\psi_{i_T}\left(\left(x_T\right)_{i_T} + d_{i_T}^T\right) - \psi_{i_T}\left(\left(x_T\right)_{i_T}\right)\right) \\ & \qquad - \frac{2}{p_{i_T}} \nabla_{i_T} f\left(x_T\right)^\top \left(x_T - \bar{x}\right)_{i_T} \\ & \qquad - \frac{2}{p_{i_T}} \left(\nabla_{i_T} f\left(x_T\right)^\top d_{i_T}^T + \frac{\bar{L}_{i_T}}{2} \left\|d_{i_T}^T\right\|^2\right). \end{aligned} \tag{45}$$

By taking expectation over i_T on both sides of (45) and using the convexity of f together with (43), we obtain

$$\begin{aligned} & \frac{1}{2}\left(\mathbb{E}_{i_T}\left[r_{T+1}^2\right] - r_T^2\right) \\ & \leq \psi\left(\bar{x}\right) - \psi\left(x_T\right) + f\left(\bar{x}\right) - f\left(x_T\right) \\ & \qquad + \left(\sum\nolimits_{i=1}^{n} F\left(x_T\right) - F\left(x_T + d_i^T e_i\right)\right) \\ & \leq \left(F^* - F\left(x_T\right)\right) \qquad\qquad\qquad\qquad\qquad \text{(46a)} \\ & \qquad + \frac{1}{p_{\min}} \sum\nolimits_{i=1}^{n} p_i \left(F\left(x_T\right) - F\left(x_T + d_i^T e_i\right)\right) \\ & = \left(F^* - F\left(x_T\right)\right) + \frac{\left(F\left(x_T\right) - \mathbb{E}_{i_T}\left[F\left(x_{T+1}\right)\right]\right)}{p_{\min}}, \qquad \text{(46b)} \end{aligned}$$

where in (46a) we used the fact that (41) is a descent method. By taking expectation over $i_0, \dots, i_{k-1}$ on (46b), summing over $T = 0, \dots, k$, and applying Lemma 1, we obtain the desired convergence rate.

Boundedness of $\mathbb{E}_{i_0, \dots, i_{k-1}}[r_k^2]$ follows from the same telescoping sum and the fact that $F(x_k)$ decreases monotonically with k. □

Our result shows that, similar to gradient descent and proximal gradient, proximal coordinate descent and coordinate descent also provide a form of implicit regularization in that the expected value of r_k is bounded. Since r_k can be viewed as a weighted Euclidean norm, this observation implies that the iterates are also in a sense expected to lie within a bounded region.

Our analysis here improves the rates in (Lu & Xiao, 2015; Lee & Wright, 2018) in terms of the dependency on k and removes the assumption of (12a) in (Lee & Wright, 2018). Even aside from the improvement from $O(1/k)$ to $o(1/k)$, Theorem 10 is the first time that a convergence rate for proximal stochastic coordinate descent with arbitrary sampling for the coordinates is proven without additional assumptions such as (42). By manipulating (46b), one can also observe how different probability distributions affect the upper bound, and it might also be possible to get better upper bounds by using norms different from (29).

4. Tightness of the $o(1/k)$ Estimate

We demonstrate that the $o(1/k)$ estimate of convergence of $\{f(x_k)\}$ is tight by showing that for any $\epsilon \in (0, 1]$, there is a convex smooth function for which the sequence of function values generated by gradient descent with a fixed step size converges slower than $O(1/k^{1+\epsilon})$. The example problem we provide is a simple one-dimensional function, so it serves also as a special case of stochastic coordinate descent and the proximal methods (where $\psi \equiv 0$) as well. Thus, this example shows tightness of our analysis for all methods without line search considered in this paper.

Consider the one-dimensional real convex function

$$f(x) = x^p, \tag{47}$$

where p is an even integer greater than 2. The minimizer of this function is clearly at $x^* = 0$, for which $f(0) = f^* = 0$. Suppose that the gradient descent method is applied starting from $x_0 = 1$. For any descent method, the iterates x_k are confined to $[-1, 1]$ and we have

$$\|\nabla^2 f(x)\| \leq p(p-1) \quad \text{for all } x \text{ with } |x| \leq 1,$$

so we set $L = p(p-1)$. Suppose that $\bar{\alpha} \in (0, 2/L)$ as above. Then the iteration formula is

$$x_{k+1} = x_k - \bar{\alpha}\nabla f(x_k) = x_k \left(1 - p\bar{\alpha}x_k^{p-2}\right), \tag{48}$$

and by Lemma 4, all iterates lie in a bounded set: the level set $[-1, 1]$ defined by x_0. In fact, since $p \geq 4$ and $\bar{\alpha} \in (0, 2/L)$, we have that

$$\begin{aligned} x_k \in (0,1] \Rightarrow 1 - p\bar{\alpha}x_k^{p-2} &\in \left(1 - \frac{2p}{p(p-1)}x_k^{p-2}, 1\right) \\ &\subseteq \left(1 - \frac{2}{p-1}, 1\right) \subseteq \left(\frac{2}{3}, 1\right), \end{aligned}$$

so that $x_{k+1} \in \left(\frac{2}{3}x_k, x_k\right)$ and the value of L remains valid for all iterates.

We show by an informal argument that there exists a constant C such that

$$f(x_k) \approx \frac{C}{k^{p/(p-2)}}, \quad \text{for all } k \text{ sufficiently large.} \tag{49}$$

From (48) we have

$$f(x_{k+1}) = f(x_k)\left(1 - p\bar{\alpha}f(x_k)^{(p-2)/p}\right)^p. \tag{50}$$

By substituting the hypothesis (49) into (50), and taking k to be large, we obtain the following sequence of equivalent approximate equalities:

$$\begin{aligned} \frac{C}{(k+1)^{p/(p-2)}} &\approx \frac{C}{k^{p/(p-2)}}\left(1 - p\bar{\alpha}\frac{C^{(p-2)/p}}{k}\right)^p \\ \Leftrightarrow \quad \left(\frac{k}{k+1}\right)^{p/(p-2)} &\approx \left(1 - p\bar{\alpha}\frac{C^{(p-2)/p}}{k}\right)^p \\ \Leftrightarrow \quad \left(1 - \frac{1}{k+1}\right)^{p/(p-2)} &\approx 1 - p^2\bar{\alpha}\frac{C^{(p-2)/p}}{k} \\ \Leftrightarrow \quad 1 - \frac{p}{p-2}\frac{1}{k+1} &\approx 1 - p^2\bar{\alpha}\frac{C^{(p-2)/p}}{k} \end{aligned}$$

This last expression is approximately satisfied for large k if C satisfies the expression

$$p/(p-2) = p^2\bar{\alpha}C^{(p-2)/p}.$$

Stated another way, our result (49) indicates that a convergence rate faster than $O(1/k^{1+\epsilon})$ is not possible when steepest descent with fixed steplength is applied to the function $f(x) = x^p$ provided that $p/(p-2) \leq 1 + \epsilon$, that is,

$$p \geq 2\frac{1+\epsilon}{\epsilon} \quad \text{and } p \text{ is a positive even integer.}$$

We follow Attouch et al. (2018) to provide a continuous-time analysis of the same objective function, using a gradient flow argument. For the function f defined by (47), consider the following differential equation:

$$x'(t) = -\alpha\nabla f(x(t)). \tag{51}$$

Suppose that

$$x(t) = t^{-\theta} \tag{52}$$

for some $\theta > 0$, which indicates that starting from any $t > 0$, $x(t)$ lies in a bounded area. Substituting (52) into (51), we obtain

$$-\theta t^{-\theta-1} = -\alpha p t^{-\theta(p-1)},$$

which holds true if and only if the following equations are satisfied:

$$\begin{cases} \theta &= \alpha p, \\ -\theta - 1 &= -\theta p + \theta, \end{cases}$$

from which we obtain $\theta = (p-2)^{-1}$, $\alpha = (p(p-2))^{-1}$. Since x decreases monotonically to zero, for all $t \geq (p-1)/(p-2)$, $L = p(p-1)((p-1)/(p-2))^{-\theta(p-2)} = p(p-2)$ is an appropriate value for a bound on $\|\nabla^2 f(x)\|$. These values of α and L satisfy $0 < \alpha \leq \frac{1}{L}$, making α a valid step size. The objective value is $f(x(t)) = t^{-p/(p-2)}$, matching the rate of (49).

Acknowledgements

This work was supported by NSF Awards IIS-1447449, 1628384, 1634597, and 1740707; Subcontract 8F-30039 from Argonne National Laboratory; and Award N660011824020 from the DARPA Lagrange Program. The authors thank Yixin Tao for a discussion that helped us to improve the clarity of this work.

References

Attouch, H. and Peypouquet, J. The rate of convergence of nesterov's accelerated forward-backward method is actually faster than 1/kˆ2. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

Attouch, H., Chbani, Z., Peypouquet, J., and Redont, P. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1-2):123–175, 2018.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Bertsekas, D. P. *Nonlinear programming*. Athena scientific Belmont, 3 edition, 2016.

Burachik, R., Graña Drummond, L., Iusem, A. N., and Svaiter, B. Full convergence of the steepest descent method with inexact line searches. *Optimization*, 32(2): 137–146, 1995.

Deng, W., Lai, M.-J., Peng, Z., and Yin, W. Parallel multi-block ADMM with $o(1/k)$ convergence. *Journal of Scientific Computing*, 71(2):712–736, 2017.

Lee, C.-p. and Wright, S. J. Inexact variable metric stochastic block-coordinate descent for regularized optimization. Technical report, 2018. URL http://www.optimization-online.org/DB_HTML/2018/08/6753.html.

Lu, Z. and Xiao, L. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015.

Mureşan, M. *A concrete approach to classical analysis*, volume 14. Springer, 2009.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Nesterov, Y. E. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Peng, W., Zhang, H., and Zhang, X. Global complexity analysis of inexact successive quadratic approximation methods for regularized optimization under mild assumptions. Technical report, 2018. arXiv:1808.04291.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. T. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Supplementary Materials for "First-Order Algorithms Converge Faster than $O(1/k)$ on Convex Problems"

Ching-pei Lee[1] **Stephen J. Wright**[1]

A. Proof of Lemma 1

Proof. The proof uses simplified elements of the proofs of Lemmas 2 and 9 of Section 2.2.1 from (Polyak, 1987). Define $s_k := k\Delta_k$ and $u_k := s_k + \sum_{i=k}^{\infty} \Delta_i$. Note that

$$s_{k+1} = (k+1)\Delta_{k+1} \leq k\Delta_k + \Delta_{k+1} \leq s_k + \Delta_k. \tag{1}$$

From (1) we have

$$\begin{aligned} u_{k+1} &= s_{k+1} + \sum_{i=k+1}^{\infty} \Delta_i \leq s_k + \Delta_k + \sum_{i=k+1}^{\infty} \Delta_i \\ &= s_k + \sum_{i=k}^{\infty} \Delta_i = u_k, \end{aligned}$$

so that $\{u_k\}$ is a monotonically decreasing nonnegative sequence. Thus there is $u \geq 0$ such that $u_k \to u$, and since $\lim_{k\to\infty} \sum_{i=k}^{\infty} \Delta_i = 0$, we have $s_k \to u$ also.

Assuming for contradiction that $u > 0$, there exists $k_0 > 0$ such that $s_k \geq u/2 > 0$ for all $k \geq k_0$, so that $\Delta_k \geq u/(2k)$ for all $k \geq k_0$. This contradicts the summability of $\{\Delta_k\}$. Therefore we have $u = 0$, so that $k\Delta_k = s_k \to 0$, proving the result. □

References

Polyak, B. T. *Introduction to Optimization.* Translation Series in Mathematics and Engineering. 1987.

[1] Department of Computer Sciences and Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, Wisconsin, USA. Correspondence to: Ching-pei Lee <ching-pei@cs.wisc.edu>, Stephen J. Wright <swright@cs.wisc.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).