

Global Convergence of Arbitrary-Block Gradient Methods for Generalized Polyak-Łojasiewicz Functions

Dominik Csiba* Peter Richtárik†

September 12, 2017

Abstract

In this paper we introduce two novel generalizations of the theory for gradient descent type methods in the proximal setting. First, we introduce the proportion function, which we further use to analyze all known (and many new) block-selection rules for block coordinate descent methods under a single framework. This framework includes randomized methods with uniform, non-uniform or even adaptive sampling strategies, as well as deterministic methods with batch, greedy or cyclic selection rules. We additionally introduce a novel block selection technique called greedy minibatches, for which we provide competitive convergence guarantees. Second, the theory of strongly-convex optimization was recently generalized to a specific class of non-convex functions satisfying the so-called Polyak-Łojasiewicz condition. To mirror this generalization in the weakly convex case, we introduce the Weak Polyak-Łojasiewicz condition, using which we give global convergence guarantees for a class of non-convex functions previously not considered in theory. Additionally, we establish (necessarily somewhat weaker) convergence guarantees for an even larger class of non-convex functions satisfying a certain smoothness assumption only.

By combining the two abovementioned generalizations we recover the state-of-the-art convergence guarantees for a large class of previously known methods and setups as special cases of our general framework. Moreover, our framework allows for the derivation of new guarantees for many new combinations of methods and setups, as well as a large class of novel non-convex objectives. The flexibility of our approach offers a lot of potential for future research, as a new block selection procedure will have a convergence guarantee for all objectives considered in our framework, while a new objective analyzed under our approach will have a whole fleet of block selection rules with convergence guarantees readily available.

Contents

1	Introduction	2
1.1	Contributions	3
1.2	Outline	5
1.3	Notation	5
2	Gradient Descent	5
2.1	Weakly Polyak-Łojasiewicz functions	6
2.2	Gradient dominated functions	8
2.3	Brief literature review	8

*University of Edinburgh

†KAUST and University of Edinburgh

3	General Setup	9
3.1	Smoothness and separability	9
3.2	Masking vectors and matrices	10
3.3	Algorithm	10
3.4	Forcing function	12
3.5	Proportion function	12
3.6	Generic descent lemma	13
4	Proportion Function	14
4.1	Smooth problems	14
4.2	Non-smooth problems	16
5	Strongly Polyak-Łojasiewicz Functions	19
5.1	Strongly PL functions	19
5.2	Strongly convex functions are strongly PL	19
5.3	Convergence	20
5.4	Applications of Theorem 16	20
6	Weakly Polyak-Łojasiewicz Functions	23
6.1	Weakly PL functions	23
6.2	Weakly convex functions are weakly PL	23
6.3	Convergence	24
6.4	Applications of Theorem 19	24
7	General Nonconvex Functions	26
7.1	Applications of Theorem 20	27
8	Experiments	28
8.1	Setup	28
8.2	Global convergence of serial coordinate descent	28
8.3	Local convergence of the gradient	29
A	Some basic properties of WPL functions	32
B	Proofs	34
B.1	Proof of Lemma 13	34
B.2	Proof of Theorem 15	34
B.3	Proof of Theorem 16	35
B.4	Proof of Theorem 18	35
B.5	Proof of Theorem 19	35
B.6	Proof of Theorem 20	36
C	Technical Lemmas	37
C.1	Lemma 24	37
C.2	Lemma 25	38
D	Notation Glossary	39

1 Introduction

During the last decade, gradient-type methods have become the methods of choice for solving optimization problems of very large sizes arising in fields such as machine learning, data science, engineering, and visual computing.

Consider the optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}),$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a differentiable function. Assume that this problem has a nonempty set of global minimizers \mathcal{X}^* (clearly, $\nabla f(x^*) = 0$ for all $x^* \in \mathcal{X}^*$). It is well known [1] that if f is L -smooth and μ -strongly convex, where $L \geq \mu > 0$, then the gradient descent method $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L} \nabla f(\mathbf{x}^k)$ for all $k \geq 0$ satisfies $\xi(\mathbf{x}^{k+1}) \leq (1 - \frac{\mu}{L}) \xi(\mathbf{x}^k)$, where $\mathbf{x}^* \in \mathcal{X}^*$ and

$$\xi(\mathbf{x}) := f(\mathbf{x}) - f(\mathbf{x}^*), \quad (1)$$

is the *optimality gap function*. Motivated by the rise of nonconvex models in fields such as image and signal processing and deep learning, there is interest in studying the performance of gradient-type methods for nonconvex functions.

As observed by Polyak in 1963 [2], and recently popularized and further studied by Karimi, Nutini and Schmidt [3] in the context of proximal methods, proofs of linear convergence rely on a certain *consequence* of strong convexity known as the *Polyak-Lojasiewicz* (PL) inequality. Since functions satisfying the PL inequality need not be convex, linear convergence of gradient methods to the global optimum extends beyond the realm of convex functions.

The (strong) PL inequality can be written in the form

$$\frac{1}{2} \|\nabla f(\mathbf{x})\|^2 \geq \mu \cdot \xi(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n. \quad (2)$$

We write $f \in \mathcal{S}_{PL}(\mu)$ if f satisfies (2). The PL inequality and methods based on it have been an inspiration for many researchers in recent years [4, 5, 6]. It is known that in order to guarantee $\xi(\mathbf{x}^k) \leq \epsilon$, it suffices to take $k = \mathcal{O}((L/\mu) \log(1/\epsilon))$.

The starting point of this paper is the realization that *while the PL inequality serves as a generalization of strong convexity, there is no equivalent generalization of (weak) convexity*. One of the key contributions of this paper is to remedy this situation by introducing the *weak PL inequality*:

$$\|\nabla f(\mathbf{x})\| \cdot \|\mathbf{x} - \mathbf{x}^*\| \geq \sqrt{\mu} \cdot \xi(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n. \quad (3)$$

We write $f \in \mathcal{W}_{PL}(\mu)$ if f satisfies (3). If f is convex, then $f \in \mathcal{W}_{PL}(1)$. Indeed, by convexity and Cauchy-Schwartz inequality, we have

$$\xi(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \leq \|\nabla f(\mathbf{x})\| \cdot \|\mathbf{x} - \mathbf{x}^*\|.$$

However, $\mathcal{W}_{PL}(1)$ contains nonconvex functions as well. As an example consider the function $f(x_1, x_2) = x_1^2 x_2^2$, for which it is straightforward to show that $f \in \mathcal{W}_{PL}(1)$ and it is apparently nonconvex. If we allow $0 < \mu < 1$, the inequality (3) becomes weaker, and holds for a larger family of functions still.

In this paper we prove that $\xi(\mathbf{x}^k) \leq \epsilon$ if $k \geq 2LR^2/\epsilon$, where R is a uniform upper bound on $\|\mathbf{x}^k - \mathbf{x}^0\|$. Since gradient descent is a monotonic method, such a bound exists if, for instance, the level set $\{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$ is bounded. This result extends standard convergence result for gradient descent for convex functions to weak PL functions.

1.1 Contributions

We now briefly summarize the main contributions of this work.

- (i) We consider a large family of gradient type methods. The methods include block coordinate descent with arbitrary block selection rules, such as cyclic, greedy, randomized, adaptive and

so on. Gradient descent arises as a special case when the active block at each iteration consists of all coordinates. Also, we introduce a novel method called greedy minibatch descent which we analyze using our developed theory to prove convergence for it in various setups mentioned in the next point.

- (ii) We extend all results (strong and weak PL inequality, algorithms and complexity results) to the proximal setup. That is, we consider *composite* optimization problems of the form

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + g(x),$$

where f is a differentiable function, and g is a simple (and possibly nonsmooth) function. For instance, the weak PL inequality (3) arises as a special case of the new *proximal weak PL inequality* when $g = 0$. The complexity results are the same: $\mathcal{O}(\log(1/\epsilon))$ for strongly PL functions (see Section 5), $\mathcal{O}(1/\epsilon)$ for weakly PL functions (see Section 6), and $\mathcal{O}(1/\epsilon \log(1/\epsilon))$ for general nonconvex functions (see Section 7.1). The specific rates can be found in Tables 1 and 2. The definitions of the various symbols and constants appearing in the tables is given in the rest of the text.

block selection rule	strongly PL	weakly PL	general nonconvex
gradient descent	$\frac{\lambda_{\max}(\mathbf{M})}{\mu} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{\lambda_{\max}(\mathbf{M})}{\rho(\mathbf{x}^0)\epsilon}$	$\frac{\lambda_{\max}(\mathbf{M})\xi(\mathbf{x}^0)}{\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$
uniform coordinate	$\frac{n \max_i \{M_{ii}\}}{\mu} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{n \max_i \{M_{ii}\}}{\rho(\mathbf{x}^0)\epsilon}$	$\frac{n \max_i \{M_{ii}\}\xi(\mathbf{x}^0)}{\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$
importance coordinate	$\frac{\sum_{i=1}^n M_{ii}}{\mu} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{\sum_{i=1}^n M_{ii}}{\rho(\mathbf{x}^0)\epsilon}$	$\frac{\sum_{i=1}^n M_{ii}\xi(\mathbf{x}^0)}{\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$
greedy coordinate	$\frac{\sum_{i=1}^n M_{ii}}{\mu} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{\sum_{i=1}^n M_{ii}}{\rho(\mathbf{x}^0)\epsilon}$	$\frac{\sum_{i=1}^n M_{ii}\xi(\mathbf{x}^0)}{\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$
uniform minibatch	$\frac{1}{\mu\lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1]})} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{1}{\rho(\mathbf{x}^0)\lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1}])\epsilon}$	$\frac{\xi(\mathbf{x}^0)}{\lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1}])\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$
greedy minibatch	$\frac{1}{\mu\lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1}])} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{1}{\rho(\mathbf{x}^0)\lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1}])\epsilon}$	$\frac{\xi(\mathbf{x}^0)}{\lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1}])\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$

Table 1: Iteration complexity guarantees for $\mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon$ in the smooth case ($g = 0$).

block selection rule	strongly PL	weakly PL	general nonconvex
gradient descent	$\frac{\lambda_{\max}(\mathbf{M})}{\mu} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{\lambda_{\max}(\mathbf{M})}{\rho(\mathbf{x}^0)\epsilon}$	$\frac{\lambda_{\max}(\mathbf{M})\xi(\mathbf{x}^0)}{\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$
uniform coordinate	$\frac{n \max_i \{M_{ii}\}}{\mu} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{n \max_i \{M_{ii}\}}{\rho(\mathbf{x}^0)\epsilon}$	$\frac{n \max_i \{M_{ii}\}\xi(\mathbf{x}^0)}{\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$
greedy coordinate	$\frac{n \max_i \{M_{ii}\}}{\mu} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{n \max_i \{M_{ii}\}}{\rho(\mathbf{x}^0)\epsilon}$	$\frac{n \max_i \{M_{ii}\}\xi(\mathbf{x}^0)}{\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$
uniform minibatch	$\frac{nL_\tau}{\tau\mu} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{nL_\tau}{\rho(\mathbf{x}^0)\tau\epsilon}$	$\frac{\xi(\mathbf{x}^0)nL_\tau}{\tau\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$
greedy minibatch	$\frac{nL_\tau}{\tau\mu} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$	$\frac{nL_\tau}{\rho(\mathbf{x}^0)\tau\epsilon}$	$\frac{\xi(\mathbf{x}^0)nL_\tau}{\tau\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right)$

Table 2: Iteration complexity guarantees for $\mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon$ in the non-smooth case ($g \neq 0$).

To the best of our knowledge, all the rates are novel, except those for strongly PL functions for gradient descent and uniform and greedy coordinate descent, in both smooth (i.e., $g = 0$) and non-smooth (i.e., $g \neq 0$) cases. These were already shown in [3]. Even in these cases, our class of strongly PL functions is somewhat larger than that considered in [3] in the non-smooth case ($g \neq 0$).

1.2 Outline

We first perform the our general analysis specified for smooth gradient descent in Section 2. In Section 3 we introduce the general setup considered in the rest of the work. In Section 4 we introduce the proportion function, which is a tool for the general analysis of block selection rules. In Sections 5, 6 and 7 we establish the main theory for strongly PL, weakly PL, and general non-convex functions, respectively. Finally, in Section 8 we perform numerical experiments confirming our theoretical findings.

1.3 Notation

We use boldface to denote a multi-dimensional object. As an example, we have a vector \mathbf{x} , a matrix \mathbf{X} , while a scalar entry of a vector x_i has a normal typeset. By $[n]$ we denote the set $\{1, \dots, n\}$. $\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2}$ is the L2 norm, where $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$ is the standard inner product. refer to Table 3 in the appendix for a summary of frequently used notation.

2 Gradient Descent

We assume throughout this section that f is L -smooth for some $L > 0$:

$$f(\mathbf{x} + \mathbf{h}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{L}{2} \|\mathbf{h}\|^2, \quad \mathbf{x}, \mathbf{h} \in \mathbb{R}^n. \quad (4)$$

We shall write $f \in C^1(L)$. In addition to this assumption, in our analysis we consider several classes of *nonconvex* objectives: PL functions, weak PL functions, and gradient dominated functions.

In this section we perform a novel analysis of the gradient descent¹ method for minimizing f :

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L} \nabla f(\mathbf{x}^k). \quad (5)$$

for the above classes of nonconvex functions. As we shall show, for these classes of objectives gradient descent converges to the global minimizer. By focusing on the notoriously known gradient descent method first, we illuminate some of the key insights of this paper without distractions from additional complications caused by the proximal setup and particularities of other algorithms, making the more general treatment in further sections more easily digestable.

A key role in the analysis is played by the *forcing function* associated with f , defined as

$$\mu(\mathbf{x}) := \frac{\|\nabla f(\mathbf{x})\|^2}{2\xi(\mathbf{x})}, \quad x \in \mathbb{R}^n / \mathcal{X}^*. \quad (6)$$

For any fixed value of this function, the smaller the gradient $\|\nabla f(\mathbf{x})\|^2$ is, the smaller the optimality gap $\xi(\mathbf{x})$. In other words, small gradients force the optimality gap to become small. The importance of this function is clear from the following simple lemma, which says that the larger $\mu(\mathbf{x}^k)$ is, the more reduction we get at iteration k in the optimality gap.

¹For simplicity, we consider gradient descent with fixed stepsize inversely proportional to the Lipschitz constant: $1/L$. While one can extend our results to other stepsize strategies using standard techniques, we avoid doing so as to present our results in a simple setting.

Lemma 1. Let $f \in C^1(L)$ and let $\{\mathbf{x}^k\}_{k \geq 0}$ be the sequence of iterates produced by the gradient descent method (5). As long as $\mathbf{x}^k \notin \mathcal{X}^*$, we have

$$\xi(\mathbf{x}^{k+1}) \leq \left(1 - \frac{\mu(\mathbf{x}^k)}{L}\right) \xi(\mathbf{x}^k).$$

Moreover, $\|\nabla f(\mathbf{x})\|^2 \leq 2L\xi(\mathbf{x})$ for all \mathbf{x} .

Proof. Let $\mathbf{h}^k = -\frac{1}{L}\nabla f(\mathbf{x}^k)$. Then

$$\begin{aligned} \xi(\mathbf{x}^{k+1}) &\stackrel{(5)}{=} \xi(\mathbf{x}^k + \mathbf{h}^k) \stackrel{(1)}{=} f(\mathbf{x}^k + \mathbf{h}^k) - f(\mathbf{x}^*) \\ &\stackrel{(4)}{\leq} f(\mathbf{x}^k) + \langle f(\mathbf{x}^k), \mathbf{h}^k \rangle + \frac{L}{2}\|\mathbf{h}^k\|^2 - f(\mathbf{x}^*) \stackrel{(1)}{=} \xi(\mathbf{x}^k) + \langle f(\mathbf{x}^k), \mathbf{h}^k \rangle + \frac{L}{2}\|\mathbf{h}^k\|^2 \\ &= \xi(\mathbf{x}^k) - \frac{1}{2L}\|\nabla f(\mathbf{x}^k)\|^2 \stackrel{(6)}{=} \left(1 - \frac{\mu(\mathbf{x}^k)}{L}\right)\xi(\mathbf{x}^k). \end{aligned}$$

Since $\xi(\mathbf{x}^{k+1}) \geq 0$, it must be the case that $\mu(\mathbf{x}) \leq L$ for all $\mathbf{x} \notin \mathcal{X}^*$. \square

It is well known that gradient descent is monotonic: $\xi(\mathbf{x}^{k+1}) \leq \xi(\mathbf{x}^k)$ for all k . Note that this property follows from the second-to-last identity in the proof, and relies on the assumption of L -smoothness only. If f is μ -strongly convex or, more generally, if $f \in \mathcal{S}_{PL}(\mu)$, then $\mu(\mathbf{x}) \geq \mu$ for all $\mathbf{x} \notin \mathcal{X}^*$, and Lemma 1 implies the linear rate $\xi(\mathbf{x}^{k+1}) \leq \left(1 - \frac{\mu}{L}\right)\xi(\mathbf{x}^k)$. This result was shown already by Polyak [2].

2.1 Weakly Polyak-Łojasiewicz functions

Consider now functions satisfying a weak version of the PL inequality. To the best of our knowledge, this is the first work where such functions are considered.

Definition 2 (Weak Polyak-Łojasiewicz functions). We say that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a weak Polyak-Łojasiewicz (WPL) function with parameter $\mu \geq 0$ if there exists $\mathbf{x}^* \in \mathcal{X}^*$ such that

$$\sqrt{\mu} \cdot \xi(\mathbf{x}) \leq \|\nabla f(\mathbf{x})\| \cdot \|\mathbf{x} - \mathbf{x}^*\|, \quad \mathbf{x} \in \mathbb{R}^n. \quad (7)$$

For simplicity, we write $f \in \mathcal{W}_{PL}(\mu)$.

Consider the Huber loss given by

$$H(z) := \begin{cases} z^2 & |z| < 1 \\ 2|z| - 1 & \text{otherwise,} \end{cases}$$

and the derived function f given by $f(x_1, x_2) = H(x_1)H(x_2)$. It is straightforward to show from the definition that f is smooth function for which $f \in \mathcal{W}_{PL}(\mu)$ for some $\mu > 0$, while $f \notin \mathcal{S}_{PL}(\mu)$ for all $\mu > 0$.

Note that all² functions belong to $\mathcal{W}_{PL}(0)$. As the next result shows, WPL functions admit a lower bound on $\mu(\mathbf{x})$ which is proportional to $\xi(\mathbf{x})$ and inversely proportional to $\|\mathbf{x} - \mathbf{x}^*\|^2$.

²By “all” we implicitly mean all functions for which the definition make sense. That is, differentiable and having a global minimizer \mathbf{x}^* .

Lemma 3. *If $f \in \mathcal{W}_{PL}(\mu)$, then*

$$\mu(\mathbf{x}) \geq \frac{\mu\xi(\mathbf{x})}{2\|\mathbf{x} - \mathbf{x}^*\|^2}, \quad \mathbf{x} \in \mathbb{R}^n / \mathcal{X}^*.$$

Proof. We have

$$\mu(\mathbf{x}) \stackrel{(6)}{=} \frac{\|\nabla f(\mathbf{x})\|^2}{2\xi(\mathbf{x})} \stackrel{(7)}{\geq} \frac{\mu\xi^2(\mathbf{x})/\|\mathbf{x} - \mathbf{x}^*\|^2}{2\xi(\mathbf{x})} = \frac{\mu\xi(\mathbf{x})}{2\|\mathbf{x} - \mathbf{x}^*\|^2}.$$

□

Several basic properties of WPL functions are summarized in Appendix A. Combining Lemma 1 and Lemma 3, we get the recursion

$$\xi(\mathbf{x}^{k+1}) \leq \left(1 - \frac{\mu\xi(\mathbf{x}^k)}{2L\|\mathbf{x}^k - \mathbf{x}^*\|^2}\right) \xi(\mathbf{x}^k). \quad (8)$$

The next lemma will be useful in the analysis of this recursion.

Lemma 4. *Let $\{\alpha^t\}_{t=0}^k$ and $\{\beta^t\}_{t=0}^k$ be two sequences of positive numbers satisfying the recursion*

$$\alpha^{t+1} \leq (1 - \alpha^t \beta^t) \alpha^t. \quad (9)$$

Then for all $k \geq 0$ we have the bound

$$\alpha^k \leq \frac{\alpha^0}{1 + \alpha^0 \sum_{t=0}^{k-1} \beta^t}.$$

Proof. As α^t and β^t are positive numbers, we have $\alpha^{t+1} \leq \alpha^t$ for all t using (9). Observe that

$$\frac{1}{\alpha^{t+1}} - \frac{1}{\alpha^t} = \frac{\alpha^t - \alpha^{t+1}}{\alpha^{t+1}\alpha^t} \geq \frac{\alpha^t - \alpha^{t+1}}{(\alpha^t)^2} \stackrel{(9)}{\geq} \beta^t,$$

which we can recursively use to show

$$\frac{1}{\alpha^k} \geq \frac{1}{\alpha^{k-1}} + \beta^{k-1} \geq \dots \geq \frac{1}{\alpha^0} + \sum_{t=0}^{k-1} \beta^t.$$

We get the result by inverting the last equation. □

By applying the above lemma to recursion (8), we get a global convergence result for gradient descent applied to a WPL function.

Theorem 5. *Let $f \in \mathcal{C}^1(L)$, and let $\{\mathbf{x}^k\}_{k \geq 0}$ be the sequence of iterates produced by the gradient descent method (5). Assume $f \in \mathcal{W}_{PL}(\mu)$ for $\mu > 0$. Then for all $k \geq 1$ we have*

$$\xi(\mathbf{x}^k) \leq \frac{\xi(\mathbf{x}^0)}{1 + \xi(\mathbf{x}^0) \frac{\mu}{2L} \sum_{t=0}^{k-1} \frac{1}{\|\mathbf{x}^t - \mathbf{x}^*\|^2}} \leq \frac{2L}{\mu} \cdot \frac{1}{\sum_{t=0}^{k-1} \frac{1}{\|\mathbf{x}^t - \mathbf{x}^*\|^2}}. \quad (10)$$

The second inequality is obtained from the first by neglecting the additive constant 1 in the denominator.

By monotonicity, all iterates of gradient descent stay in the level set $\mathcal{L}_0 := \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$. If this set is bounded, then $R := \max_{\mathbf{x} \in \mathcal{L}_0} \|\mathbf{x} - \mathbf{x}^*\| < +\infty$, and we have $\|\mathbf{x}^k - \mathbf{x}^*\| \leq R$ for all k . In this case, the bound (10) implies

$$\xi(\mathbf{x}^k) \leq \frac{2LR^2}{\mu k}.$$

2.2 Gradient dominated functions

We now consider a new class of (not necessarily convex) functions. To the best of our knowledge, this class was not considered in optimization before.

Definition 6. We say that function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is φ -gradient dominated if there exists a function $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\varphi(0) = 0$, $\lim_{t \rightarrow 0} \varphi(t) = 0$ and

$$\xi(\mathbf{x}) \leq \varphi(\|\nabla f(\mathbf{x})\|), \quad \mathbf{x} \in \mathbb{R}^n. \quad (11)$$

The above definition essentially says that for any sequence $\{\mathbf{x}^k\}$ (not necessarily related to iterates of gradient descent) such that $\|\nabla f(\mathbf{x}^k)\| \rightarrow 0$, we must have $f(\mathbf{x}^k) \rightarrow f(\mathbf{x}^*)$. In particular, if f has multiple minimizers, all must have the same function value.

As an example of the function φ , we might consider any function of the form $\varphi(t) = c \cdot |t|^p$, where $c > 0$ and $p > 0$. The specific choice of $p = 2$ was already considered before in [7].

Theorem 7. Assume that $f \in \mathcal{C}^1(L)$ is φ -gradient dominated. Pick $\epsilon > 0$ and let

$$k \geq \frac{2L\xi(\mathbf{x}^0)}{\epsilon} \log \left(\frac{\xi(\mathbf{x}^0)}{\varphi(\epsilon)} \right). \quad (12)$$

Then $\min\{\xi(\mathbf{x}^t) : t = 0, 1, \dots, k\} \leq \varphi(\epsilon)$.

Proof. If $\|\nabla f(\mathbf{x}^t)\|^2 \leq \epsilon$ for some $k = 0, 1, \dots, k - 1$, we are done by applying (11). Otherwise, $\|\nabla f(\mathbf{x}^t)\|^2 > \epsilon$ for all $t = 0, 1, \dots, k - 1$. Then in view of (6), $\mu(\mathbf{x}^t) > \epsilon/(2\xi(\mathbf{x}^t))$ for all such t . By monotonicity, $\xi(\mathbf{x}^{t+1}) \leq \xi(\mathbf{x}^t)$ for all t , whence $\mu(\mathbf{x}^t) > \epsilon/(2\xi(\mathbf{x}^0))$. By applying Lemma 1, we get

$$\xi(\mathbf{x}^{t+1}) \leq \left(1 - \frac{\epsilon}{2L\xi(\mathbf{x}^0)}\right) \xi(\mathbf{x}^t), \quad t = 0, 1, \dots, k - 1.$$

By unrolling the recurrence, and using the bound $(1 - s)^{1/s} \leq \exp(-1)$ (which holds for $0 < s \leq 1$), we get

$$\xi(\mathbf{x}^k) \leq \left(1 - \frac{\epsilon}{2L\xi(\mathbf{x}^0)}\right)^k \xi(\mathbf{x}^0) \leq \exp\left(-\frac{\epsilon k}{2L\xi(\mathbf{x}^0)}\right) \xi(\mathbf{x}^0) \stackrel{(12)}{\leq} \varphi(\epsilon).$$

□

2.3 Brief literature review

The original gradient descent method was developed by Cauchy [8] and it has seen a lot of development ever since. This development is documented in detail in [9, 1]. In the recent years, a version of gradient descent called coordinate descent was developed. The first developments of coordinate descent are due to [10] and it was first analyzed for general convex objectives by Nesterov in [11].

An important part of the coordinate descent is its ability to work with arbitrary block selection strategies. In the seminal work of Nesterov [11], there were three strategies introduced, which are known as coordinate descent with uniform probabilities, coordinate descent with importance sampling, and greedy coordinate descent. The first two strategies fall into the family of randomized strategies. These were further developed in [21, 12, 13, 14, 19, 20, 15]. The third selection rule is a

deterministic strategy similar in nature to batch gradient descent and it was shown to be superior over randomized methods in terms of iteration complexity in [16].

The Polyak-Lojasiewicz condition was first introduced by Boris Polyak in [2]. It was revived recently in [3] and applied to modern optimization approaches. Since then, multiple papers used the condition to develop new approaches [4, 5, 6]. Gradient dominated functions were recently considered in [7].

Lastly, we note that our framework considers the non-accelerated version of coordinate descent methods, although they play a key role in modern theory. If needed, the acceleration can be achieved non-directly by using approaches as proposed in [17] or [18]. We leave the accelerated counterpart of this framework to future work.

3 General Setup

In this section we move beyond the simplified setup considered in the previous section and introduce the setting considered in this paper in its full generality. Our general treatment differs from that in Section 2 in several ways.

First, we consider the *composite* optimization problem

$$\min_{x \in \mathbb{R}^n} \{F(\mathbf{x}) \quad := \quad f(\mathbf{x}) + g(\mathbf{x})\}, \quad (13)$$

where f is assumed to be smooth, and g is a simple (possibly nonconvex and nonsmooth) separable function. Second, we f is assumed to be smooth in a slightly more general sense than L -smoothness of Section 2. Third, we go beyond gradient descent and consider a large family of first order methods which include randomized, cyclic, adaptive and greedy coordinate descent, in serial and adaptive settings.

In the following, we will refer to the problem (13) with $g = 0$ as the smooth case and otherwise as the non-smooth case.

3.1 Smoothness and separability

We assume, that f is \mathbf{M} -smooth, which is formalized by the following assumption:

Assumption 8. *We say that a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is \mathbf{M} -smooth, if there exists a positive definite matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ such that*

$$f(\mathbf{x} + \mathbf{h}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2} \langle \mathbf{M} \mathbf{h}, \mathbf{h} \rangle, \quad \mathbf{x}, \mathbf{h} \in \mathbb{R}^n. \quad (14)$$

In the non-smooth case ($g \neq 0$), we will without loss of generality assume \mathbf{M} to be a multiple of the identity matrix; specifically $\mathbf{M} := L\mathbf{I}$, where \mathbf{I} is the identity matrix. If (8) holds for some \mathbf{M} , we can always replace \mathbf{M} by $L\mathbf{I}$, where $L = \lambda_{\max}(\mathbf{M})$. It is easy to verify, that if a function is \mathbf{M} -smooth, it is also $L\mathbf{I}$ -smooth. Note that in some cases we might choose L to be smaller. This will be in detail explained in Section 4.2.

For simplicity, we will use the notation with \mathbf{M} also in the non-smooth case, but we will always treat it as the diagonal matrix $L\mathbf{I}$.

In addition to smoothness of f , we assume that the function g is separable, which is defined as follows:

Assumption 9. We say that a function $g : \mathbb{R}^n \rightarrow \bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is separable, if there exist n scalar functions $g_1, \dots, g_n : \mathbb{R} \rightarrow \bar{\mathbb{R}}$, such that

$$g(\mathbf{x}) = \sum_{i=1}^n g_i(x_i). \quad (15)$$

Note that the function g is treated as the non-smooth part of the problem (e.g., L1 norm, box constraints, and so on). When we refer to a smooth problem, we assume the setup with $g = 0$, while all the other setups are referred as non-smooth problems.

The problem described in (13) is encountered in many areas, ranging from machine learning and signal processing to biology and beyond. We believe it does not need to be motivated further, as it was already considered in a lot of previous works.

3.2 Masking vectors and matrices

Let $\mathbf{x} \in \mathbb{R}^n$ be an arbitrary vector and let $\mathbf{X} \in \mathbb{R}^{n \times n}$ be an arbitrary matrix. We will need to index vectors and matrices by subsets of coordinates $\emptyset \neq S \subseteq [n]$. The indexing has two distinct forms. By \mathbf{x}_S we denote the $|S|$ -dimensional vector constructed by taking the entries of \mathbf{x} with indices in S , while the notation $\mathbf{x}_{[S]}$ is used to zero out every entry of \mathbf{x} not appearing in S without changing its length. We have similar notation for matrices, where \mathbf{X}_S denotes the $|S| \times |S|$ matrix of the entries with both column and row indices in S , while $\mathbf{X}_{[S]}$ is used for the matrix \mathbf{X} with entries zeroes out outside of columns and rows with indices in S . As a rule of thumb, the subscript S changes the dimensions of the object, while the subscript $[S]$ maintains its dimensions.

To illustrate the notation, consider the following example.

Example 1. Let $\mathbf{x} \in \mathbb{R}^3$, $\mathbf{X} \in \mathbb{R}^{3 \times 3}$ and $S = \{1, 3\}$. Then

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \Rightarrow \mathbf{x}_S = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \quad \mathbf{x}_{[S]} = \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}$$

and

$$\mathbf{X} = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \Rightarrow \mathbf{X}_S = \begin{bmatrix} 1 & 7 \\ 3 & 9 \end{bmatrix}, \quad \mathbf{X}_{[S]} = \begin{bmatrix} 1 & 0 & 7 \\ 0 & 0 & 0 \\ 3 & 0 & 9 \end{bmatrix}.$$

3.3 Algorithm

We now propose and analyze a wide class of block descent algorithms for solving (13). For a non-empty block of coordinates $S \subseteq [n]$, we define

$$U_S(\mathbf{x}, \mathbf{u}) := \langle (\nabla f(\mathbf{x}))_{[S]}, \mathbf{u} \rangle + \frac{1}{2} \mathbf{u}^\top \mathbf{M}_{[S]} \mathbf{u} + \sum_{i \in S} [g_i(x_i + u_i) - g_i(x_i)]. \quad (16)$$

We assume that finding a minimizer of $U_S(\mathbf{x}, \mathbf{h})$ in \mathbf{h} is cheap (e.g., there exists a closed form solution, or an efficient algorithm). Given an iterate \mathbf{x}^k , in iteration k of our method we select a block $S_k \subseteq [n]$ of active coordinates, according to an *arbitrary block selection procedure* \mathcal{P}_k , and subsequently minimize $U_{S_k}(\mathbf{x}^k, \mathbf{h})$ in \mathbf{h} . The result is denoted \mathbf{u}^k . Due to the structure of problem

(16), the minimizer of U_{S_k} does not depend on u_i for $i \notin S_k$. Hence, only the active coordinates \mathbf{u}_i^k for $i \in S_k$ are relevant, and we set $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{u}_{[S_k]}^k$. Equivalently, using other alternative notation, we can write this as

$$\mathbf{x}_i^{k+1} = \begin{cases} \mathbf{x}_i^k + \mathbf{u}_i^k, & i \in S_k, \\ \mathbf{x}_i^k, & i \notin S_k, \end{cases}$$

or

$$\mathbf{x}_{S_k}^{k+1} = \mathbf{x}_{S_k}^k + \mathbf{u}_{S_k}^k, \quad \mathbf{x}_{[n]/S_k}^{k+1} = \mathbf{x}_{[n]/S_k}^k.$$

This is Algorithm 1.

Algorithm 1 Proximal Arbitrary-Block Descent Method

Input: Initial iterate $\mathbf{x}^0 \in \text{dom}(g)$, arbitrary block selection procedures $\{\mathcal{P}_k\}_{k=0}^{K-1}$
for $k = 0, \dots, K - 1$ **do**
 Pick a non-empty subset of coordinates $S_k \subset [n]$ using the procedure \mathcal{P}_k
 Compute $\mathbf{u}^k \in \arg \min_{\mathbf{u}} \{U_{S_k}(\mathbf{x}^k, \mathbf{u})\}$
 Update $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{u}_{[S_k]}^k$
end for
Output \mathbf{x}^K

Observe, that in the smooth case we have

$$\mathbf{u}_{S_k}^k = -\mathbf{M}_{S_k}^{-1} \nabla_{S_k} f(\mathbf{x}^k), \quad (17)$$

where $\nabla_S f(\mathbf{x}) := (\nabla f(\mathbf{x}))_S$. The iteration can be computed as a solution of a linear system of a size $|S_k| \times |S_k|$, which is very cheap for small $|S_k|$. In the non-smooth case, the iterate does not have a closed-form solution in general but can be solved fast for a lot of forms of g , e.g., box constraints or L1 norm. Since we assume $\mathbf{M} = L\mathbf{I}$, we can write

$$\mathbf{u}_{S_k}^k = \arg \min_{\mathbf{u} \in \mathbb{R}^{|S_k|}} \left\{ \langle \nabla_S f(\mathbf{x}^k), \mathbf{u} \rangle + \frac{L}{2} \|\mathbf{u}\|^2 + \sum_{i \in S_k} g_i(x_i + u_i) \right\}. \quad (18)$$

For simplicity, assume $\mathbf{M} = L\mathbf{I}$ for some constant $L > 0$. If $g = 0$, and we always pick $S_k = [n]$, then Algorithm 1 reduces to gradient descent, considered in Section 2. If $g \neq 0$, and we always pick $S_k = [n]$, then Algorithm 1 reduces to proximal gradient descent. On the other hand, if we always pick $|S_k| = 1$, then we obtain coordinate descent ($g = 0$) or proximal coordinate descent ($g \neq 0$). The selection procedure \mathcal{P}_k may be set to choose the coordinates in a cyclic manner, greedily, randomly according to any (fixed or evolving) probability law, and even adaptively to the entire history of the iterative process. There are many other possibilities between the two extremes of always selecting $|S_k| = 1$ and $|S_k| = n$. Such methods can be considered block coordinate descent methods, subspace descent methods, or parallel coordinate descent methods (as the updates to individual coordinates can be performed in parallel). We stress that unlike all other methods considered in the literature, in our method we allow for the block selection procedure \mathcal{P}_k to be *arbitrary*, without any restrictions whatsoever.

By removing these restrictions, we allow for several new possibilities in the sampling procedure. These are: 1) the block selection procedure might change from iteration to iteration, allowing for

adaptive strategies as [19], 2) the procedure might depend on previous iterations, which opens up the possibility of cyclic and other similar selections, and 3) the procedure does not have to be randomized, which allows for greedy selection procedures [16].

In all our convergence results we shall enforce several key common assumptions, together with some additional assumptions. In order to avoid repeating the common core, we shall summarize them here.

Assumption 10 (Common Assumptions). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an \mathbf{M} -smooth (14) function, let $g : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ be separable (15), and let $F : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ be function defined using f and g as in (13). Assume F has a global minimizer \mathbf{x}^* , such that $F(\mathbf{x}^*) > -\infty$. Let $\mathbf{x}^0 \in \text{dom}(g)$ be an initial point and let the sequence $\{\mathbf{x}^k\}_{k=1}^K$ be generated using Algorithm 1, where $\{S_k\}_{k=0}^{K-1}$ is an arbitrary sequence of non-empty (possibly random) subsets of $[n]$.*

3.4 Forcing function

As before, let us define the *optimality gap function*

$$\xi(\mathbf{x}) \quad := \quad F(\mathbf{x}) - F(\mathbf{x}^*), \quad (19)$$

where \mathbf{x}^* is a minimizer of F . Observe that $\xi(\mathbf{x}) \geq 0$ with equality only if $\mathbf{x} \in \mathcal{X}^* := \arg \min_{\mathbf{x}} F(\mathbf{x})$.

We now extend the definition of the forcing function to the proximal setting.

Definition 11 (Forcing function: proximal version). Let

$$\lambda(\mathbf{x}) \quad := \quad -L \cdot \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x} + \mathbf{y}) - g(\mathbf{x}) \right\}, \quad (20)$$

where L is chosen such that (14) holds with $\mathbf{M} = L\mathbf{I}$. Specific choices of L are explained in detail in Section 4.2. The non-negative function

$$\mu(\mathbf{x}) \quad := \quad \frac{\lambda(\mathbf{x})}{\xi(\mathbf{x})} \quad (21)$$

is the *proximal forcing function*. The domain of \mathbf{x} is $\text{dom}(g)/\mathcal{X}^*$.

Pick any $\mathbf{x} \in \text{dom}(g)/\mathcal{X}^*$. The minimum in (20) is non-positive since setting $\mathbf{y} = \mathbf{0}$ gives a zero value. Hence, $\lambda(\mathbf{x}) \geq 0$. Since the denominator in (21) is positive, $\mu(\mathbf{x})$ is always nonnegative.

In the smooth case ($g = 0$) we have

$$\mu(\mathbf{x}) \quad = \quad \frac{\|\nabla f(\mathbf{x})\|^2}{2\xi(\mathbf{x})}, \quad (22)$$

which is simply the forcing function (6).

3.5 Proportion function

We introduce one more notion, which we call the *proportion function*. This function plays an important role in our theory.

Definition 12. Let

$$\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^n : \lambda(\mathbf{x}) \neq 0\}. \quad (23)$$

The *proportion function* is defined by

$$\theta(S, \mathbf{x}) := \frac{-\min_{\mathbf{u} \in \mathbb{R}^n} \mathbf{U}_S(\mathbf{x}, \mathbf{u})}{\lambda(\mathbf{x})} \quad (24)$$

for all $S \in \mathcal{S}$ and $\mathbf{x} \in \mathcal{X}$. For $\mathbf{x} \in \mathcal{X}$, we set $\theta(S, \mathbf{x}) = 0$.

Similarly as in the definition of the forcing function, both of the minimizations in (24) are non-positive, as $\mathbf{u} = 0$ and $\mathbf{y} = 0$ gives zero value in the numerator and denominator, respectively. Also observe that in the smooth case ($g = 0$) we have

$$\theta(S, \mathbf{x}) = \frac{(\nabla_S f(\mathbf{x}))^\top (\mathbf{M}_S)^{-1} \nabla_S f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|^2}. \quad (25)$$

Note that the matrix $(\mathbf{M}_S)^{-1}$ exists since all principal submatrices of a positive definite matrix are also positive definite. A more detailed treatment of the proportion function can be found in Section 4.

Also note that \mathcal{X} (23) might differ from \mathcal{X}^* in the case of local minimizers.

3.6 Generic descent lemma

We now formulate a simple but important descent lemma which bounds the progress gained by a single iteration of Algorithm 1. Our bound applies to arbitrary block selection rules, and will enable us to prove global convergence results for Algorithm 1 for new classes of nonconvex functions.

Lemma 13. *Let \mathbf{x}^{k+1} be the next iterate of Algorithm 1 generated from \mathbf{x}^k by picking a nonempty set of coordinates $S_k \subseteq [n]$. Then*

$$\xi(\mathbf{x}^{k+1}) \leq [1 - \theta(S_k, \mathbf{x}^k) \cdot \mu(\mathbf{x}^k)] \cdot \xi(\mathbf{x}^k). \quad (26)$$

Applying this repeatedly, for all $K \geq 1$ we obtain the estimate

$$\xi(\mathbf{x}^K) \leq \left(\prod_{k=0}^{K-1} [1 - \theta(S_k, \mathbf{x}^k) \cdot \mu(\mathbf{x}^k)] \right) \xi(\mathbf{x}^0). \quad (27)$$

Proof. See Section B.1 □

Recursion (26) is a direct generalization of Lemma 1. Indeed, if $\mathbf{M} = L\mathbf{I}$, $g = 0$, and $S_k = [n]$, then in the view of (25), $\theta([n], \mathbf{x}^k) = 1/L$. Note that as $\theta \geq 0$ and $\mu \geq 0$, we can be sure that $\xi(\mathbf{x}^{k+1}) \leq \xi(\mathbf{x}^k)$, which means that we are not getting worse by iterating the Algorithm 1. The proof of Lemma 13 is straightforward. The difficulty will lie in bounding the forcing and proportion functions so as to obtain convergence. In Sections 5.2, 6 and 7 we apply this lemma to prove the main results of this paper.

The K -step bound (27) provides us with a compact and generic bound on the optimality gap at the K -th iterate of Algorithm 1 dependent on the iterates $\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^{K-1}$ and the selected sets S_0, S_1, \dots, S_{k-1} . This result does not immediately imply convergence as at this level of generality, it is possible for the product appearing in (27) not to converge to zero. Indeed, this corollary also covers the situation where $S_k = 1$ for all k , which clearly can't result in convergence. We will need to introduce further restrictions in order to establish convergence.

4 Proportion Function

In this section we show standard bounds on the proportion function, which are independent of a given iterate. This will be important further, to recover the convergence rates given by standard theory. We note, that for stochastic methods we bound the expectation of the proportion function conditioned on the last iterate, instead of directly bounding the proportion function. This quantity will be important in the theory for the convergence of stochastic methods, as specified in following sections.

The proportion function is the only quantity in the convergence theory, which is dependent on the choice of the set of coordinates S . Therefore, to analyze a new sampling strategy for coordinate descent, one only has to show a bound on the proportion function. This opens up a possible venue of novel techniques for coordinate selection.

In the following we tackle all the known cases of samplings, which we break down by their smoothness. Also, we introduce and analyze a new sampling procedure to showcase the generality of our framework.

4.1 Smooth problems

In the case of $g = 0$, we have the proportion function equal to (25), i.e.,

$$\theta(S, \mathbf{x}) = \frac{(\nabla_S f(\mathbf{x}))^\top (\mathbf{M}_S)^{-1} \nabla_S f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|^2}$$

for all $\mathbf{x} \in \mathcal{X}^*$, and all $\emptyset \neq S \subseteq [n]$. Let us break-down the cases according to specific choices of the set S .

- **Batch Gradient descent:** In the case when $S = [n]$, we recover the standard gradient descent strategy, which dates back to the work of Cauchy [8]. In this case we can lower bound the proportion function by

$$\theta([n], \mathbf{x}) = \frac{(\nabla f(\mathbf{x}))^\top \mathbf{M}^{-1} \nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|^2} \geq \frac{1}{\lambda_{\max}(\mathbf{M})} \quad (28)$$

for all $\mathbf{x} \in \mathcal{X}^*$.

- **Serial Coordinate Descent:** Suppose $S = \{i\}$, for some given i . In this case we get

$$\theta(\{i\}, \mathbf{x}) = \frac{(\nabla_i f(\mathbf{x}))^2}{M_{ii} \|\nabla f(\mathbf{x})\|^2}. \quad (29)$$

There are multiple strategies for choosing the coordinate i , and we tackle them one by one.

- **Uniform probabilities:** Suppose we choose the coordinate i with the probability p_i given by $p_i = 1/n$, independently of \mathbf{x} . This strategy was originally analyzed in [11]. The expectation of θ can be lower bounded as

$$\mathbf{E}_i[\theta(\{i\}, \mathbf{x}) \mid \mathbf{x}] \stackrel{(29)}{=} \frac{1}{n} \sum_{i=1}^n \frac{(\nabla_i f(\mathbf{x}))^2}{M_{ii} \|\nabla f(\mathbf{x})\|^2} \geq \frac{1}{n \cdot \max_i \{M_{ii}\}}. \quad (30)$$

- **Importance sampling:** Suppose we choose the coordinate i with the probability p_i given by

$$p_i = \frac{M_{ii}}{\sum_{i=1}^n M_{ii}}, \quad (31)$$

independently of \mathbf{x} . Again, this strategy was originally analyzed in [11]. The expectation of θ can be lower bounded as

$$\mathbf{E}_i[\theta(\{i\}, \mathbf{x}) \mid \mathbf{x}] \stackrel{(29)}{=} \frac{1}{\sum_{i=1}^n M_{ii}} \sum_{i=1}^n \frac{M_{ii}(\nabla_i f(\mathbf{x}))^2}{M_{ii}\|\nabla f(\mathbf{x})\|^2} \geq \frac{1}{\sum_{i=1}^n M_{ii}} \quad (32)$$

- **Greedy choice:** Suppose we choose the coordinate i deterministically as

$$i = \arg \max_j \left\{ \frac{(\nabla_j f(\mathbf{x}))^2}{M_{jj}} \right\}. \quad (33)$$

It is straightforward that this strategy maximizes the proportion function (29) for a single iteration, given that we choose only a single coordinate. It was originally proposed in [11] and further improved in [16]. However, in this case we do not have a better bound than (32), which would be independent of \mathbf{x}^3 . We can get this bound using that the maximum of some quantity is more than its average weighted by M_{ii}

$$\theta(\{i\}, \mathbf{x}) = \frac{\max_j \{(\nabla_j f(\mathbf{x}))^2 / M_{jj}\}}{\|\nabla f(\mathbf{x})\|^2} \geq \frac{1}{\sum_{j=1}^n M_{jj}}. \quad (34)$$

Observe that in the case of $\|\nabla f(\mathbf{x})\|^2 = (\nabla_i f(\mathbf{x}))^2$ we have that the above lower bound (34) could be $\sum_{j=1}^n M_{jj} / M_{ii}$ larger to still hold. Therefore on some iterations, greedy coordinate descent is much better than coordinate descent with importance sampling, although their global bounds are the same. This usually leads to superiority of greedy rules in practice, in the case that they can be implemented cheaply.

- **Minibatch Coordinate Descent:** Suppose $|S| = \tau$, for some given $n > \tau > 1$. There are currently two analyzed strategies in this case and we introduce a third.

- **τ -nice sampling:** Assume that we want to sample a subset of $\tau > 1$ coordinates at each iteration, uniformly at random from all subsets of cardinality τ . It can be inferred from results established in [20] that the expectation of the proportion function can be lower bounded by the quantity

$$\mathbf{E}[\theta(S, \mathbf{x}) \mid \mathbf{x}] \stackrel{(25)}{=} \frac{(\nabla f(\mathbf{x}))^\top \mathbf{E}[\mathbf{M}_{[S]}^{-1}] \nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|^2} \geq \lambda_{\min} \left(\mathbf{E} \left[\mathbf{M}_{[S]}^{-1} \right] \right), \quad (35)$$

where $\mathbf{M}_{[S]}^{-1}$ is the $n \times n$ matrix constructed by putting the matrix \mathbf{M}_S^{-1} on the columns and rows specified by S and zero out the rest of the entries. Additionally, assuming that we have a factorization $\mathbf{M} = \mathbf{A}^\top \mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$, using results from [14], this can be further lower bounded as

$$\lambda_{\min} \left(\mathbf{E} \left[\mathbf{M}_{[S]}^{-1} \right] \right) \stackrel{[20]}{\geq} \frac{1}{n \cdot \max_i \{v_i\}}, \quad (36)$$

³In [16] they proved a slightly better bound using ℓ_1 -strong convexity, which can be achieved by replacing the ℓ_2 -norm by ℓ_∞ -norm in the definition of the proportion and forcing functions.

where

$$v_i \stackrel{[14]}{:=} \sum_{j=1}^m \left[1 + \frac{(\|\mathbf{A}_j\|_0 - 1)(\tau - 1)}{n - 1} \right] A_{ji}^2.$$

- **Importance sampling for minibatches:** A minibatch version of importance sampling was recently proposed in [15]. The main idea is as follows: we randomly partition the coordinates into τ approximately equally sized “buckets”, and subsequently and independently perform standard importance sampling (as described above) for each bucket. The sampling is then generated as the union of all sampled coordinates. For specific bounds, we recommend discussing the original paper [15], as they are not available in a compact form.
- **Greedy minibatches:** To showcase the power of our framework, we introduce a brand-new selection rule called *greedy minibatches*. This selection rule aims to select the set S such that it minimizes the proportion function on the current iteration, i.e.,

$$S^g := \arg \max_{S:|S|=\tau} \left\{ (\nabla_S f(\mathbf{x}))^\top (\mathbf{M}_S)^{-1} \nabla_S f(\mathbf{x}) \right\}. \quad (37)$$

The above selection procedure is a difficult problem in general, but it might be feasible for some specific problems, e.g., for diagonal \mathbf{M} or the function f with a special structure (see [16] for examples).

To get a lower bound on the proportion function independent of the current iterate \mathbf{x} , we use the argument that the maximum of some quantities is always at least equal to any weighted mean of the same quantities. Using this, we get that $\theta(S^g, \mathbf{x}) \geq \mathbf{E}[\theta(S, \mathbf{x}) \mid \mathbf{x}]$, where the sets S were selected according to the τ -nice sampling, importance sampling for minibatches introduced above, or any other sampling. Therefore, using (35) we get that

$$\theta(S^g, \mathbf{x}) \stackrel{(25)+(35)}{\geq} \lambda_{\min} \left(\mathbf{E} \left[\mathbf{M}_{[S]}^{-1} \right] \right). \quad (38)$$

Observe that both the selection rule and the above bound is a generalization of the greedy coordinate sampling to minibatches. Also observe that the above bound is potentially very loose. As an example, consider a diagonal matrix \mathbf{M} with all the elements equal to 1. The right-hand side of (38) is then equal to τ/n , while the left-hand side is equal to 1. Even in this very special case, the bound is disregarding a factor of n/τ , which is potentially huge. For this reason, it is expected that the greedy minibatches will actually perform much better in practice than in theory (38).

4.2 Non-smooth problems

Let us define for all i and all $\mathbf{x} \in \mathbb{R}^n$ the function

$$\lambda_i(\mathbf{x}) := -L \cdot \min_{v \in \mathbb{R}} \left\{ \nabla_i f(\mathbf{x})v + \frac{L}{2}v^2 + g_i(x_i + v) - g_i(x_i) \right\}. \quad (39)$$

Using the definition of proportion function (24) with the diagonal $\mathbf{M} := \mathbf{L}\mathbf{I}$ we get

$$\theta(S, \mathbf{x}) \stackrel{(24)+(39)}{=} \frac{\sum_{i \in S} \lambda_i(\mathbf{x})}{L \sum_{j=1}^n \lambda_j(\mathbf{x})}. \quad (40)$$

As mentioned in Section 3.1, the value L can be chosen as $L = \lambda_{\max}(\mathbf{M})$ to satisfy the inequality (14). However, this choice of L might be suboptimal in many cases. As an example, when analyzing coordinate descent methods, the vector \mathbf{h} will always be of the form $h\mathbf{e}^i$, where \mathbf{e}^i is the i -th coordinate vector (see Appendix B.1). Therefore, it would be sufficient for L to satisfy (14) for the specific choice $\mathbf{h} = h\mathbf{e}^i$ which leads to

$$f(\mathbf{x} + h\mathbf{e}^i) \leq f(\mathbf{x}) + \nabla_i f(\mathbf{x})h + \frac{M_{ii}}{2}h^2 \leq f(\mathbf{x}) + \nabla_i f(\mathbf{x})h + \frac{L}{2}h^2$$

for each $i \in \{1, \dots, n\}$ and $h \in \mathbb{R}$. We can easily observe from the above that this will be satisfied with $L := \max_i \{M_{ii}\} \leq \lambda_{\max}(\mathbf{M})$. To account for this in general, we need to take some additional measures. Specifically, let \mathcal{S} be the collection of all sets $S \subset [n]$ which can be possibly generated during the iterative process by \mathcal{P}_k . As an example, $\mathcal{S} = \{\{1\}, \{2\}, \dots, \{n\}\}$ corresponds to all such sets for coordinate descent. We define L as the smallest number for which

$$f(\mathbf{x} + \mathbf{h}_{[S]}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h}_{[S]} \rangle + \frac{1}{2} \langle \mathbf{M} \mathbf{h}_{[S]}, \mathbf{h}_{[S]} \rangle$$

is satisfied for every $\mathbf{x}, \mathbf{h} \in \mathbb{R}^n$ and every $S \in \mathcal{S}$. It is straightforward to see that if the size of possible sets $S \in \mathcal{S}$ is upper bounded as $|S| \leq \tau$, then we can safely choose

$$L = L_\tau := \max_{S: |S|=\tau} \{\lambda_{\max}(\mathbf{M}_S)\} \leq \max_{S: |S|=\tau} \left\{ \sum_{i \in S} M_{ii} \right\}, \quad (41)$$

where the last inequality is due to the eigenvalues being positive and their sum being the trace. We can easily verify that this generalizes to $L = L_n = \lambda_{\max}(\mathbf{M})$ for gradient descent and $L = L_1 = \max_i \{M_{ii}\}$ for coordinate descent.

Now, we will breakdown the cases depending on the block selection procedure.

- **Proximal Gradient Descent:** The first result in the proximal setting was the Iterative Shrinkage Thresholding Algorithm (ISTA), which selects all the coordinates on each iteration. The bound on the proportion function is given by

$$\theta([n], \mathbf{x}) \stackrel{(40)}{=} \frac{\sum_{i=1}^n \lambda_i(\mathbf{x})}{L_n \sum_{j=1}^n \lambda_j(\mathbf{x})} = \frac{1}{L_n} \stackrel{(41)}{=} \frac{1}{\lambda_{\max}(\mathbf{M})}. \quad (42)$$

- **Serial Proximal Coordinate descent:** Suppose $|S| = 1$ and specifically let $S = \{i\}$. Then we have that

$$\theta(\{i\}, \mathbf{x}) \stackrel{(40)}{=} \frac{\lambda_i(\mathbf{x})}{L_1 \sum_{j=1}^n \lambda_j(\mathbf{x})}, \quad (43)$$

as the size of the selected sets are upper bounded by 1. The procedure leading to the choice of i distinguishes between various serial approaches.

- **Uniform probabilities:** Assume we choose the coordinate i uniformly at random at each iteration. Then we can bound the expectation of the proportion function as

$$\begin{aligned} \mathbf{E}[\theta(\{i\}, \mathbf{x}) \mid \mathbf{x}] &\stackrel{(43)}{=} \mathbf{E} \left[\frac{\lambda_i(\mathbf{x})}{L_1 \sum_{j=1}^n \lambda_j(\mathbf{x})} \right] = \frac{\frac{1}{n} \sum_{i=1}^n \lambda_i(\mathbf{x})}{L_1 \sum_{j=1}^n \lambda_j(\mathbf{x})} = \frac{1}{nL_1} \\ &\stackrel{(41)}{=} \frac{1}{n \max_i \{M_{ii}\}}. \end{aligned} \quad (44)$$

- **Greedy choice:** Another approach is to pick the coordinate i , which maximizes the proportion function in (43), which was analyzed in [16]. The best bound independent on \mathbf{x} coincides with the above bound for uniform sampling, using the fact that a mean of some quantity is less than its maximum

$$\begin{aligned} \theta(\{i\}, \mathbf{x}) &\stackrel{(43)}{=} \frac{\max_i \{\lambda_i(\mathbf{x})\}}{L_1 \sum_{j=1}^n \lambda_j(\mathbf{x})} \geq \frac{\frac{1}{n} \sum_{i=1}^n \lambda_i(\mathbf{x})}{L_1 \sum_{j=1}^n \lambda_j(\mathbf{x})} = \frac{1}{nL_1} \\ &\stackrel{(41)}{=} \frac{1}{n \max_i \{M_{ii}\}}. \end{aligned} \quad (45)$$

Similarly as in the smooth case, observe that the quantity $\max_i \{\lambda_i(\mathbf{x})\}$ is potentially up to n times larger than $\frac{1}{n} \sum_{i=1}^n \lambda_i(\mathbf{x})$ and therefore the bound (45) can be potentially n times larger in some cases, which usually results in better empirical results.

- **Minibatch Proximal Coordinate Descent:** Suppose $|S| = \tau$, where τ is given, which implies that we can use $L = L_\tau$ in the bounds. We introduce two options for the block selection procedure.

- **τ -nice sampling:** Only one variant of a sampling for this setup was considered before [13], and that is a uniform choice of τ coordinates without repetition. In expectation, each coordinate has a chance of τ/n to be picked, which is used in the bound to get

$$\begin{aligned} \mathbf{E}[\theta(S, \mathbf{x}) \mid \mathbf{x}] &\stackrel{(40)}{=} \mathbf{E} \left[\frac{\sum_{i \in S} \lambda_i(\mathbf{x})}{L_\tau \sum_{j=1}^n \lambda_j(\mathbf{x})} \right] = \frac{\frac{\tau}{n} \sum_{i=1}^n \lambda_i(\mathbf{x})}{L_\tau \sum_{j=1}^n \lambda_j(\mathbf{x})} = \frac{\tau}{nL_\tau} \\ &\stackrel{(41)}{=} \frac{\tau}{n \max_{S:|S|=\tau} \{\lambda_{\max}(\mathbf{M}_S)\}}. \end{aligned} \quad (46)$$

To our best knowledge, this bound is new, as the previous bound considered L_n instead of L_τ . As $L_\tau \leq L_n$, the new bound is better.

- **Greedy minibatches:** Similarly as in the smooth case, we introduce a new selection rule – greedy minibatches. Specifically, the corresponding set S^g is given by

$$S^g := \arg \max_{S:|S|=\tau} \left\{ \sum_{i \in S} \lambda_i(\mathbf{x}) \right\}. \quad (47)$$

Note that for $\tau = 1$ we recover the greedy coordinate descent. To give global bounds independent of \mathbf{x} for this strategy, we again use the fact that maximum is an upper bound for the mean, to get

$$\begin{aligned} \theta(S^g, \mathbf{x}) &\stackrel{(40)}{=} \frac{\sum_{i \in S^g} \lambda_i(\mathbf{x})}{L_\tau \sum_{j=1}^n \lambda_j(\mathbf{x})} \geq \frac{\frac{\tau}{n} \sum_{i=1}^n \lambda_i(\mathbf{x})}{L_\tau \sum_{j=1}^n \lambda_j(\mathbf{x})} = \frac{\tau}{nL_\tau} \\ &\stackrel{(41)}{=} \frac{\tau}{n \max_{S:|S|=\tau} \{\lambda_{\max}(\mathbf{M}_S)\}}. \end{aligned} \quad (48)$$

Note that the the above bound can be potentially very pessimistic, as

$$1 \geq \frac{\sum_{i=1}^n \lambda_i(\mathbf{x})}{\sum_{i \in S^g} \lambda_i(\mathbf{x})} \geq \frac{\tau}{n}$$

by using $\sum_{i \in S^g} \lambda_i(\mathbf{x}) \geq \frac{\tau}{n} \sum_{i=1}^n \lambda_i(\mathbf{x}) \geq \frac{\tau}{n} \sum_{i \in S^g} \lambda_i(\mathbf{x})$. Therefore, the bound (48) can be up to n/τ times better in certain cases.

5 Strongly Polyak-Łojasiewicz Functions

In this section, we reinvent the strongly PL functions using the proximal forcing function (21), and develop the corresponding convergence rates. Also, we show how to recover the known results in this setting by applying our theory.

5.1 Strongly PL functions

Definition 14 (Strongly PL functions: composite case). We say that F is a strongly PL function there exists a scalar $\mu > 0$ satisfying

$$\mu(\mathbf{x}) \geq \mu \quad (49)$$

for all $\mathbf{x} \in \text{dom}(g)/\mathcal{X}^*$. The collection of all functions F satisfying inequality (49) will be denoted $\mathcal{S}_{PL}^g(\mu)$, and we say that F is strongly PL with parameter μ .

Recall that in the smooth case we said that a function $f \in \mathcal{S}_{PL}(\mu)$, if f satisfied the condition (2), which can be observed to be equivalent to (49) for smooth functions. Therefore we have that $\mathcal{S}_{PL}(\mu) \subset \mathcal{S}_{PL}^g(\mu)$.

The above definition is not new, it was originally introduced in a slightly different form by Karimi et al. [3].

5.2 Strongly convex functions are strongly PL

Let $\lambda \geq 0$. Function $F : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ is said to be λ -strongly convex, if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and all $\beta \in [0, 1]$ we have

$$F(\beta\mathbf{x} + (1 - \beta)\mathbf{y}) \leq \beta F(\mathbf{x}) + (1 - \beta)F(\mathbf{y}) - \frac{\lambda\beta(1 - \beta)}{2} \|\mathbf{x} - \mathbf{y}\|^2. \quad (50)$$

If F is 0-strongly convex, we refer to it simply as *convex*. Consider a differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. If f is λ -strongly convex for $\lambda \geq 0$, then

$$f(\mathbf{x} + \mathbf{h}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{\lambda}{2} \|\mathbf{h}\|^2, \quad \mathbf{x}, \mathbf{h} \in \mathbb{R}^n. \quad (51)$$

We will now show that if F is strongly convex, then $F \in \mathcal{S}_{PL}^g(\rho)$ for some specific ρ and \mathcal{A} . This means that the class of strongly PL (composite) functions contains the class of strongly convex (composite) functions.

Theorem 15. Assume F is λ_F -strongly convex with $\lambda_F > 0$, and f is λ_f -strongly convex with $\lambda_f \geq 0$. Then

$$\mu(\mathbf{x}) \geq \mu := \min \left\{ \frac{L}{2}, \frac{L\lambda_F}{\lambda_F - \lambda_f + L} \right\} \quad (52)$$

for all $\mathbf{x} \in \text{dom}(g)/\mathcal{X}^*$, and hence $F \in \mathcal{S}_{PL}^g(\mu)$.

Proof. See Section B.2. □

In the above theorem we do not enforce separability assumption on g .

5.3 Convergence

We have the following convergence result, establishing convergence to a global minimizer.

Theorem 16. *Invoke Assumption 10. Further, assume $F \in \mathcal{S}_{PL}^g(\mu)$ (i.e., F is strongly PL with parameter μ), and let*

$$\mu_k := \frac{\mu \mathbf{E}[\xi(\mathbf{x}^k) \cdot \theta(S_k, \mathbf{x}^k)]}{\mathbf{E}[\xi(\mathbf{x}^k)]}. \quad (53)$$

Then

$$\sum_{k=0}^{K-1} \mu_k \geq \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon. \quad (54)$$

Proof. See Section B.3. □

In order to get concrete complexity results from the above theorem, we need to estimate the speed of growth of $\sum_{k=0}^{K-1} \mu_k$ in K . There no universal way to do this, which is why we state the above result the way we do. Instead, in each situation this needs to be estimated separately. Typically, this will be done by lower bounding μ_k for each k separately. Let us illustrate this using a couple examples.

If the blocks are selected *deterministically*, then the expectations in (53) do not play any role, and we have $\mu_k = \mu \theta(S_k, \mathbf{x}^k)$. If we have a global lower bound of the form

$$\theta(S_k, \mathbf{x}^k) \geq c > 0$$

readily available, then $\sum_{k=0}^{K-1} \mu_k \geq \mu c K$, which implies the rate

$$K \geq \frac{1}{c\mu} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon \quad (55)$$

or, equivalently, $\xi(\mathbf{x}^K) \leq \xi(\mathbf{x}^0) \cdot e^{-c\mu K}$. More generally, convergence is established whenever we can lower bound $\theta(S_k, \mathbf{x}^k) \geq c_k > 0$, where the constants c_k sum up to infinity.

If the blocks are selected *stochastically*, then the sequence of iterates \mathbf{x}^k is also stochastic. In such cases, it is often possible to come up with a bound for the expectation of $\theta(S_k, \mathbf{x}^k)$ conditioned on \mathbf{x}^k :

$$\mathbf{E}[\theta(S_k, \mathbf{x}^k) \mid \mathbf{x}^k] \geq c > 0. \quad (56)$$

If this is the case, we claim that μ_k can be lower bounded by μc , and hence Theorem 16 implies the same rate as before: given by (55), only with a bound on the expectation $\mathbf{E}[\xi(\mathbf{x}^k)] \leq \epsilon$ on the right-hand side. More generally, if $\mathbf{E}[\theta(S_k, \mathbf{x}^k) \mid \mathbf{x}^k] \geq c_k \geq 0$, then $\mu_k \geq \mu c_k$, and convergence is guaranteed as long as $\sum_k c_k = \infty$.

Let us now return to the claim. Indeed, the bound $\mu_k \geq \mu c$ follows from (56) by applying Lemma 25 (see Appendix) with $X := \mathbf{x}^k, Y := \theta_k$ and $f(\mathbf{x}) := \xi(\mathbf{x})$.

5.4 Applications of Theorem 16

We now showcase the use of Theorem 16 on selected algorithms which arise as special cases of our generic method (Algorithm 18).

Proximal gradient descent. If for all k we choose $S_k = [n]$ with probability 1, Algorithm 18 reduces to (proximal) gradient descent. In both smooth and non-smooth case we have $\theta([n], \mathbf{x}^k) \geq 1/\lambda_{\max}(\mathbf{M})$ (see (28) and (42)). Substituting into (55), we get the rate

$$K \geq \frac{L}{\mu} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon,$$

in both the smooth and non-smooth case. These results were previously obtained for strongly PL functions in [3].

Randomized coordinate descent with uniform probabilities. Randomized coordinate descent, analyzed in [11, 21], arises as special case of Algorithm 18 by choosing $S_k = \{i_k\}$, where i_k is an index chosen from $[n]$ uniformly at random, and independently of the history of the method.

Note that a bound on $\mathbf{E}[\theta(S_k, \mathbf{x}^k | \mathbf{x}^k)]$ is readily available in Section 4, specifically in (30) for the smooth case and (44) in the non-smooth case, and it takes the form

$$\mathbf{E}[\theta(S_k, \mathbf{x}^k | \mathbf{x}^k)] \geq c := \frac{1}{n \max_i \{M_{ii}\}}.$$

As we have seen in the discussion immediately following Theorem 16, this implies the bound $\mu_k \geq \mu c$. Applying Theorem 16, we conclude that

$$K \geq \frac{n \max\{M_{ii}\}}{\mu} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon,$$

which is the same bound as given in [3] for strongly PL functions in both smooth and non-smooth case.

Randomized coordinate descent with importance sampling We now allow for specific nonuniform probabilities: probability of choosing $S_k = \{i\}$ is proportional to M_{ii} (see (31)). In view of (32), we get $\mathbf{E}[\theta(S_k, \mathbf{x}^k | \mathbf{x}^k)] \geq c := 1/\sum_{i=1}^n M_{ii}$ for smooth functions. This leads to the complexity result

$$K \geq \frac{\sum_{i=1}^n M_{ii}}{\mu} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon,$$

which is *is a new result* for strongly PL functions. However, in the special case of strongly-convex functions, this result is known [11, 21, 12].

Minibatch coordinate descent. Assume S_k is a subset of $[n]$ of cardinality τ , chosen uniformly at random. This leads to (strandard) minibatch coordinate descent. In view of (35), we have the bound $\mathbf{E}[\theta(S, \mathbf{x} | \mathbf{x})] \geq \lambda_{\min} \left(\mathbf{E} \left[\mathbf{M}_{[S]}^{-1} \right] \right)$ in the smooth case. Substituting into (55), we get the rate

$$K \geq \frac{1}{\mu \lambda_{\min}(\mathbf{E}[\mathbf{M}_S^{-1}])} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon.$$

This result is novel for the strongly PL case, but it was already established before for strongly convex functions in [20].

In the non-smooth case we can use the bound (46) given by $\mathbf{E}[\theta(S, \mathbf{x}) \mid \mathbf{x}] \geq \tau/nL_\tau$, where L_τ is given by (41). It follows that we have the guarantee

$$K \geq \frac{nL_\tau}{\mu\tau} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon,$$

which is novel for strongly PL and also for strongly convex functions, as L_τ is less or equal to the standard L_n [13].

Greedy coordinate descent. Assume a single coordinate $S = \{i\}$ is chosen using the rule stated in (33), i.e., the choice maximizes the proportion function on the given iteration. In the smooth case we have the bound (34), i.e., $\theta(S_k, \mathbf{x}^k) \geq c := 1/\sum_{i=1}^n M_{ii}$. This leads to the rate

$$K \geq \frac{\sum_{i=1}^n M_{ii}}{\mu} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon.$$

Note that this is identical to the rate of randomized coordinate descent with importance sampling, with the exception that we have $\xi(\mathbf{x}^K)$ instead of $\mathbf{E}[\xi(\mathbf{x}^K)]$ due to the deterministic nature of the method. Again, this result was already established in [3]. In the special case of strongly convex functions, this was first established by Nesterov [11].

As for the proximal case, we have a bound (45) given by $\theta(S, \mathbf{x}^k) \geq c := 1/n \max_i \{M_{ii}\}$ which leads to a similar rate

$$K \geq \frac{n \max_i \{M_{ii}\}}{\mu} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon.$$

This rate is identical to the rate of randomized coordinate descent, except for the expectation on the right-hand side. This result was also established in [3], while the strongly convex version is due [16].

Greedy minibatches. Assume the smooth and and that we choose the set of coordinates S according to the rule described in (37), which is

$$S = S(\mathbf{x}) := \arg \max_{S: |S|=\tau} \left\{ (\nabla_S f(\mathbf{x}))^\top (\mathbf{M}_S)^{-1} \nabla_S f(\mathbf{x}) \right\}.$$

Assuming that $|S| = \tau$, we can see that this rule maximizes the proportion function for the given iteration. The corresponding lower bound takes the form (38) which leads to the rate

$$K \geq \frac{1}{\lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1}])\mu} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon.$$

Similarly as in the serial case, this bound is identical to uniform minibatches, with the only exception being the dropped expectation on the optimality gap $\xi(\mathbf{x})$, due to this algorithm being deterministic.

In the non-smooth case we have the bound on the proportion function given by (48) which leads to the rate

$$K \geq \frac{nL_\tau}{\mu\tau} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon.$$

Both of the above bounds are novel, as greedy minibatches is a novel sampling approach.

6 Weakly Polyak-Łojasiewicz Functions

In this section we introduce a generalized definition of Weakly PL functions, which specifies to Definition 2 for a specific choice of parameters, and also covers the proximal case. We show that proximal convex functions can be analyzed using this framework and we give a general convergence rate guarantee for this class. Lastly, we specify our theory to several known setups, showcasing the generality of our definition.

6.1 Weakly PL functions

Definition 17 (Weakly PL functions: general case). We say that F is a weakly PL function, if there exists a scalar function $\rho : \mathbb{R}^n \rightarrow \mathbb{R}^+$, such that

$$\mu(\mathbf{x}) \geq \rho(\mathbf{x}^0) \cdot \xi(\mathbf{x}) > 0, \quad (57)$$

for all $\mathbf{x}^0, \mathbf{x} \in \text{dom}(g)/\mathcal{X}^*$ such that $\xi(\mathbf{x}) \leq \xi(\mathbf{x}^0)$. The collection of all functions F satisfying inequality (57) will be denoted $\mathcal{W}_{PL}^g(\rho)$, and we say that F is weakly PL with parameter ρ .

Recall that in the smooth case we said that a function $f \in \mathcal{W}_{PL}(\mu)$, if f satisfied the condition (11)

$$\mu(\mathbf{x}) \geq \frac{\mu\xi(\mathbf{x})}{2\|\mathbf{x} - \mathbf{x}^*\|^2}, \quad \mathbf{x} \in \mathbb{R}^n/\mathcal{X}^*.$$

In the proof of convergence of these methods, we further bounded the right-hand side of the above expression by $\mu\xi(\mathbf{x})/2R^2$, where we defined $R := \max_{\mathbf{x} \in \mathcal{L}_0} \|\mathbf{x} - \mathbf{x}^*\| < +\infty$, with $\mathcal{L}_0 := \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$. To see the above smooth case in our general framework defined in (57), we can use $\rho(\mathbf{x}^0) = \frac{1}{2R^2}$, which satisfies our assumption, as R does not depend on \mathbf{x} .

6.2 Weakly convex functions are weakly PL

We will now show that if F is convex, then $F \in \mathcal{W}_{PL}^g(\rho)$ for some specific ρ . This means that weakly PL functions generalize weakly convex functions also in the composite setting.

Theorem 18. *Assume f and g are convex, and let \mathbf{x}^* be a global minimizer of F . Then*

$$\mu(\mathbf{x}) \geq \xi(\mathbf{x}) \cdot \min \left\{ \frac{1}{2\xi(\mathbf{x})}, \frac{1}{2L\|\mathbf{x} - \mathbf{x}^*\|^2} \right\} \quad (58)$$

for all $\mathbf{x} \in \text{dom}(g)/\mathcal{X}^*$. Also, F is a weakly PL function with the parameter ρ given by

$$\rho(\mathbf{x}^0) = \min \left\{ \frac{L}{2\xi(\mathbf{x}^0)}, \frac{1}{2R^2} \right\}, \quad (59)$$

where

$$R := \max_{\mathbf{x} \in \mathbb{R}^n: f(\mathbf{x}) \leq f(\mathbf{x}^0)} \|\mathbf{x} - \mathbf{x}^*\| < +\infty. \quad (60)$$

Proof. See Section B.4. □

Note that in the smooth case we have $\xi(\mathbf{x}) \leq \frac{L}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$, therefore

$$\min \left\{ \frac{1}{2\xi(\mathbf{x})}, \frac{1}{2L\|\mathbf{x} - \mathbf{x}^*\|^2} \right\} = \frac{1}{2L\|\mathbf{x} - \mathbf{x}^*\|^2},$$

which also leads to $\rho(\mathbf{x}^0) = 1/(2R^2)$ in the smooth case.

6.3 Convergence

For this class, we can show the following convergence result.

Theorem 19. *Invoke Assumption 10. Further, assume $F \in \mathcal{W}_{PL}^g(\rho)$ (i.e., F is weakly PL with parameter ρ). Let*

$$\mu_k := \rho(\mathbf{x}^0) \frac{\mathbf{E}[(\xi(\mathbf{x}^k))^2 \cdot \theta(S_k, \mathbf{x}^k)]}{(\mathbf{E}[\xi(\mathbf{x}^k)])^2}. \quad (61)$$

Then

$$\sum_{k=0}^{K-1} \mu_k \geq \frac{1}{\epsilon} \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon. \quad (62)$$

Proof. See Section B.5. □

Similarly as in the previous section, we need to bound the quantity $\sum_{k=0}^K \mu_k$ in K to get a complexity result. The standard theory is developed by bounding each of μ_k separately, although this is apparently not the optimal way.

In the case that the blocks are selected deterministically, then the expectations in (61) do not play any role and we have $\mu_k = \rho(\mathbf{x}^0)\theta(S_k, \mathbf{x}^k)$. Additionally, if we have a global lower bound $\theta(S_k, \mathbf{x}^k) \geq c > 0$, then $\sum_{k=0}^{K-1} \mu_k \geq \rho(\mathbf{x}^0)cK$, which implies the rate

$$K \geq \frac{1}{\rho(\mathbf{x}^0)c\epsilon} \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon$$

and equivalently $\xi(\mathbf{x}^K) \leq \frac{1}{\rho(\mathbf{x}^0)cK}$.

We get a similar result for stochastic block selection. Specifically, if we have a global bound for the quantity $\mathbf{E}[\theta(S_k, \mathbf{x}^k) \mid \mathbf{x}^k] \geq c > 0$, we can lower bound the values of μ_k in Theorem 19 by $\rho(\mathbf{x}^0)c$. This can be done by first lower bounding the denominator of (61) by $\mathbf{E}[(\xi(\mathbf{x}^k))^2]$ using the trivial bound $\mathbf{E}[X]^2 \leq \mathbf{E}[X^2]$ and further applying Lemma 25 with $X := \mathbf{x}^k, Y := \theta_k, f(\mathbf{x}) = (\xi(\mathbf{x}))^2$.

In general, we can claim convergence if we have a sequence of lower bounds $\{c_k\}_{k=0}^{\infty}$ for $\theta(S_k, \mathbf{x}^k) \geq c_k$ in the deterministic case or $\mathbf{E}[\theta(S_k, \mathbf{x}^k) \mid \mathbf{x}^k] \geq c_k$ in the stochastic case, and additionally $\sum_{k=0}^{\infty} c_k = \infty$.

6.4 Applications of Theorem 19

We now showcase the use of Theorem 19 on several methods which arise as special cases of our generic method (Algorithm 18). All the results for general \mathcal{W}_{PL}^g functions are novel, as the notion itself is novel. In most cases, we recover known theory by specializing the results to convex objectives.

Proximal gradient descent. In both smooth and non-smooth case we have $\theta([n], \mathbf{x}^k) \geq c := 1/\lambda_{\max}(\mathbf{M})$ (see (28) and (42)). Substituting into (62), we get the rate

$$K \geq \frac{\lambda_{\max}(\mathbf{M})}{\rho(\mathbf{x}^0)\epsilon} \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon,$$

in both the smooth and non-smooth case. This result is novel for Weakly PL functions, but it was established before for convex functions in [9].

Randomized coordinate descent with uniform probabilities. The bound on the quantity $\mathbf{E}[\theta(S_k, \mathbf{x}^k | \mathbf{x}^k)]$ is readily available in Section 4, specifically in (30) for the smooth case and (44) in the non-smooth case, and it takes the form $\mathbf{E}[\theta(S_k, \mathbf{x}^k | \mathbf{x}^k)] \geq c := \frac{1}{n \max_i \{M_{ii}\}}$. As discussed right after Theorem 19, this implies the bound $\mu_k \geq \mu c$. Applying Theorem 19, we conclude that

$$K \geq \frac{n \max\{M_{ii}\}}{\rho(\mathbf{x}^0)\epsilon} \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon,$$

which is novel for weakly PL functions, but it is well known for convex objectives [11].

Randomized coordinate descent with importance sampling In view of (32), we get the lower bound $\mathbf{E}[\theta(S_k, \mathbf{x}^k | \mathbf{x}^k)] \geq c := 1/\sum_{i=1}^n M_{ii}$ for smooth functions. This leads to the complexity result

$$K \geq \frac{\sum_{i=1}^n M_{ii}}{\rho(\mathbf{x}^0)\epsilon} \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon,$$

which is a novel result for weakly PL functions. However, in the special case of convex functions, this result is known [11].

Minibatch coordinate descent. In view of (35), we have the lower bound $\mathbf{E}[\theta(S, \mathbf{x} | \mathbf{x})] \geq \lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1}])$ in the smooth case. Substituting into (62), we get the rate

$$K \geq \frac{1}{\rho(\mathbf{x}^0)\lambda_{\min}(\mathbf{E}[\mathbf{M}_S^{-1}])\epsilon} \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon.$$

This result is novel for the weakly PL case and to the best of our knowledge also for the convex case, as the results in [20] only consider strongly convex objectives.

In the non-smooth case we can use the bound (46) given by $\mathbf{E}[\theta(S, \mathbf{x} | \mathbf{x})] \geq \tau/nL_\tau$, where L_τ is given by (41). It follows that we have the guarantee

$$K \geq \frac{nL_\tau}{\rho(\mathbf{x}^0)\tau\epsilon} \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon,$$

which is novel for weakly PL and to the best of our knowledge also for weakly convex functions, as L_τ is less or equal to the standard L_n [13].

Greedy coordinate descent. In the smooth case we have the bound (34), i.e., $\theta(S_k, \mathbf{x}^k) \geq c := 1/\sum_{i=1}^n M_{ii}$. This leads to the rate

$$K \geq \frac{\sum_{i=1}^n M_{ii}}{\rho(\mathbf{x}^0)\epsilon} \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon.$$

This is identical to the above rate of randomized coordinate descent with importance sampling, with a dropped expectation on $\xi(\mathbf{x}^0)$. This result is new for weakly PL functions and in the special case of weakly convex functions, it was first established by Nesterov [11].

In the proximal case, we have a bound (45) given by $\theta(S, \mathbf{x}^k) \geq c := 1/n \max_i \{M_{ii}\}$ which leads to the rate

$$K \geq \frac{n \max_i \{M_{ii}\}}{\rho(\mathbf{x}^0)\epsilon} \Rightarrow \xi(\mathbf{x}^K) \leq \epsilon.$$

This rate is identical to the rate of randomized coordinate descent, except for the expectation on the right-hand side. This result is novel for weakly PL and to the best of our knowledge it is also novel for the special case of convex functions.

Greedy minibatches. In the smooth case we have the bound (38) which leads to the rate

$$K \geq \frac{1}{\rho(\mathbf{x}^0)\lambda_{\min}(\mathbf{E}[\mathbf{M}_S^{-1}])\epsilon} \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon.$$

In the non-smooth case we have the bound on the proportion function given by (48) which leads to the rate

$$K \geq \frac{nL_\tau}{\rho(\mathbf{x}^0)\tau\epsilon} \Rightarrow \mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon.$$

Both of the above results are shared with uniform minibatch coordinate descent, except that we are bounding the optimality gap itself instead of its expectation.

Both of the above rates are novel, as greedy minibatches constitute a novel sampling approach.

7 General Nonconvex Functions

In this section we establish a generic convergence result applicable to general nonconvex functions. This is done at the expense of losing global optimality: we will show that either $\lambda(\mathbf{x}^k)$ gets small, or that $F(\mathbf{x}^k)$ is close to the global minimum $F(\mathbf{x}^*)$. Recall that in the smooth case ($g = 0$) we have $\lambda(\mathbf{x}^k) = \frac{1}{2}\|\nabla f(\mathbf{x}^k)\|^2$.

Theorem 20. *Invoke Assumption 10. Let $\epsilon > 0$ be fixed. Further, let*

$$\mu_k := \frac{\mathbf{E}[\xi(\mathbf{x}^k)\theta(S_k, \mathbf{x}^k)]}{\mathbf{E}[\xi(\mathbf{x}^k)]}. \quad (63)$$

If the inequality

$$\frac{\xi(\mathbf{x}^0)}{\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right) \leq \sum_{k=0}^{K-1} \mu_k, \quad (64)$$

holds, then at least one of the following conclusions holds:

- (i) $\lambda(\mathbf{x}^k) < \epsilon$ for at least one $k \in \{0, \dots, K-1\}$,
- (ii) $\mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon$.

Proof. See Section B.6. □

In the smooth case ($g = 0$), condition (i) reduces to $\frac{1}{2}\|\nabla f(\mathbf{x}^k)\|^2 \leq \epsilon$ for at least one $k \in \{0, \dots, K-1\}$. Also note that the same strategies for bounding μ_k as those outlined in 5.3 apply here. To sum it up, if we have a global bound $\theta(S_k, \mathbf{x}^k) \geq c > 0$ in the deterministic case, then it follows from Theorem 20 that

$$K \geq \frac{\xi(\mathbf{x}^0)}{c\epsilon} \log\left(\frac{\xi(\mathbf{x}^0)}{\epsilon}\right) \Rightarrow \left((\xi(\mathbf{x}^K) \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right).$$

Similarly, if we have a bound $\mathbf{E}[\theta(S_k, \mathbf{x}^k) \mid \mathbf{x}^k] \geq c > 0$ in the stochastic case, we get the same result as above with an expectation over the optimality gap $\mathbf{E}[\xi(\mathbf{x}^K)]$, as the optimality gap becomes a random variable.

7.1 Applications of Theorem 20

In this part we apply the results from Theorem 20 to several known methods to acquire local convergence guarantees. To the best of our knowledge, all the results in this section are novel.

Proximal gradient descent. In both smooth and non-smooth case we have $\theta([n], \mathbf{x}^k) \geq c := 1/\lambda_{\max}(\mathbf{M})$ (see (28) and (42)). Substituting into (64), we get the rate

$$K \geq \frac{\lambda_{\max}(\mathbf{M})\xi(\mathbf{x}^0)}{\epsilon} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \left((\xi(\mathbf{x}^K) \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right)$$

in both the smooth and non-smooth case.

Randomized coordinate descent with uniform probabilities. In both smooth and non-smooth case we have the bound $\mathbf{E}[\theta(S_k, \mathbf{x}^k) \mid \mathbf{x}^k] \geq c := \frac{1}{n \max_i \{M_{ii}\}}$ (see (30) and (44)). This implies the rate

$$K \geq \frac{\max_i \{M_{ii}\} \xi(\mathbf{x}^0)}{\epsilon} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \left((\mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right)$$

for both smooth and non-smooth functions.

Randomized coordinate descent with importance sampling In view of (32), we have that $\mathbf{E}[\theta(S_k, \mathbf{x}^k \mid \mathbf{x}^k)] \geq c := 1/\sum_{i=1}^n M_{ii}$ for smooth functions. This leads to the complexity result

$$K \geq \frac{\xi(\mathbf{x}^0) \sum_{i=1}^n M_{ii}}{\epsilon} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \left((\mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right).$$

Minibatch coordinate descent. We have the bound $\mathbf{E}[\theta(S, \mathbf{x}) \mid \mathbf{x}] \geq \lambda_{\min} \left(\mathbf{E} \left[\mathbf{M}_{[S]}^{-1} \right] \right)$ in the smooth case from (35). Substituting into (64), we get the rate

$$K \geq \frac{\xi(\mathbf{x}^0)}{\epsilon \lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1}])} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \left((\mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right)$$

for smooth objectives. In the non-smooth case we can use the bound (46) given by $\mathbf{E}[\theta(S, \mathbf{x}) \mid \mathbf{x}] \geq \tau/nL_\tau$, where L_τ is given by (41). It follows that we get the rate

$$K \geq \frac{\xi(\mathbf{x}^0)nL_\tau}{\tau\epsilon} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \left((\mathbf{E}[\xi(\mathbf{x}^K)] \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right)$$

for non-smooth functions.

Greedy coordinate descent. In the smooth case we have the bound (34), i.e., $\theta(S_k, \mathbf{x}^k) \geq c := 1/\sum_{i=1}^n M_{ii}$. This leads to the rate

$$K \geq \frac{\xi(\mathbf{x}^0) \sum_{i=1}^n M_{ii}}{\epsilon} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \left((\xi(\mathbf{x}^K) \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right)$$

for smooth functions. In the proximal case, we have a bound (45) given by $\theta(S, \mathbf{x}^k) \geq c := 1/n \max_i \{M_{ii}\}$ which leads to the rate

$$K \geq \frac{\max_i \{M_{ii}\} \xi(\mathbf{x}^0)}{\epsilon} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \left((\xi(\mathbf{x}^K) \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right)$$

for non-smooth functions.

Greedy minibatches. In the smooth case we have the bound (38) which leads to the rate

$$K \geq \frac{\xi(\mathbf{x}^0)}{\epsilon \lambda_{\min}(\mathbf{E}[\mathbf{M}_{[S]}^{-1}])} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \left((\xi(\mathbf{x}^K) \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right)$$

In the non-smooth case we have the bound on the proportion function given by (48) which leads to the rate

$$K \geq \frac{\xi(\mathbf{x}^0) n L \tau}{\tau \epsilon} \log \left(\frac{\xi(\mathbf{x}^0)}{\epsilon} \right) \Rightarrow \left((\xi(\mathbf{x}^K) \leq \epsilon) \vee (\exists k \in [K] : \lambda(\mathbf{x}^k) \leq \epsilon) \right)$$

Uniform minibatch coordinate descent has bounds of the same form, except that the expectation is missing due to the deterministic nature of the method.

Also, both of the above rate are new, as greedy minibatches was introduced as a novel sampling approach.

8 Experiments

In this section we show results of sample numerical experiments. We will focus on showcasing the theory of the general non-convex optimization methods introduced in Section 7.

8.1 Setup

For our experiments, we consider a function defined as in (13) with f defined as

$$f(\mathbf{x}) := \frac{1}{2m} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \frac{1}{m} \cos(\langle \mathbf{c}, \mathbf{x} \rangle) \quad (65)$$

with $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$, and the function g defined as

$$g(\mathbf{x}) := \lambda \|\mathbf{x}\|_1. \quad (66)$$

While it is not motivated by any specific problem, it is clearly non-convex and we can easily control it to observe the behavior of the proposed method.

8.2 Global convergence of serial coordinate descent

In the first part, we consider the setup defined in the above section using (65) and (66) with $m = 1000$ and $n = 100$. We generate $\mathbf{A} \in \mathbb{R}^{m \times n}$ as a random matrix with fixed singular values linearly spaced between $\frac{1}{m}$ and 1. The vector $\mathbf{b} \in \mathbb{R}^m$ is set to $\mathbf{b} = \mathbf{A}\mathbf{y}$ for a vector $\mathbf{y} \in \mathbb{R}^n$ randomly generated from a normalized Gaussian distribution. Similarly, \mathbf{c} is also randomly generated from

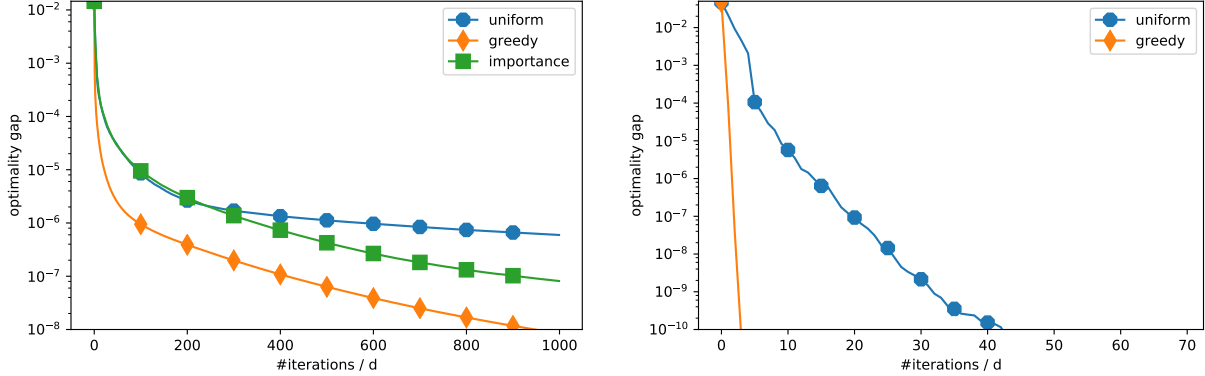


Figure 1: Plots of the convergences of various sampling methods for the smooth (left) and non-smooth (right) experiment.

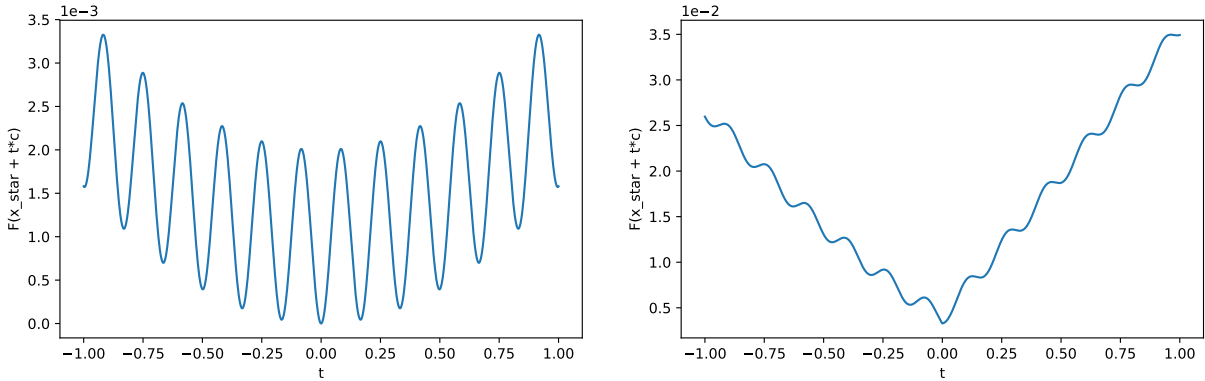


Figure 2: Plots of the functions in a 1-dimensional slice around the optimum for the smooth (left) and non-smooth (right) experiment.

a normalized Gaussian. We consider two different problems, based on the value of λ . We have a smooth problem for $\lambda = 0$ and a non-smooth problem for $\lambda = \frac{1}{2m}$.

We measure the performance of the various coordinate descent approaches described in Section 7.1. The convergence behaviors corresponding to the smooth and non-smooth setup can be found in Figure 1, on the left and right plots, respectively.

To make sure that the functions being optimized are indeed non-convex, we plotted a 1-dimensional slice of the function being optimized around its optimum. These plots can be found on Figure 2, for the smooth and non-smooth version respectively.

8.3 Local convergence of the gradient

The theory of the non-convex case in Theorem 20 does not always guarantee convergence to the global optimum, but it at least guarantees a convergence of the magnitude of the gradient. To showcase this scenario, we focused on a 1-dimensional instance of the smooth problem defined in (65). We set $\mathbf{A} = [1]$, $\mathbf{b} = \frac{\pi}{c}$ and we have chosen c such that the function f has a flat inflection

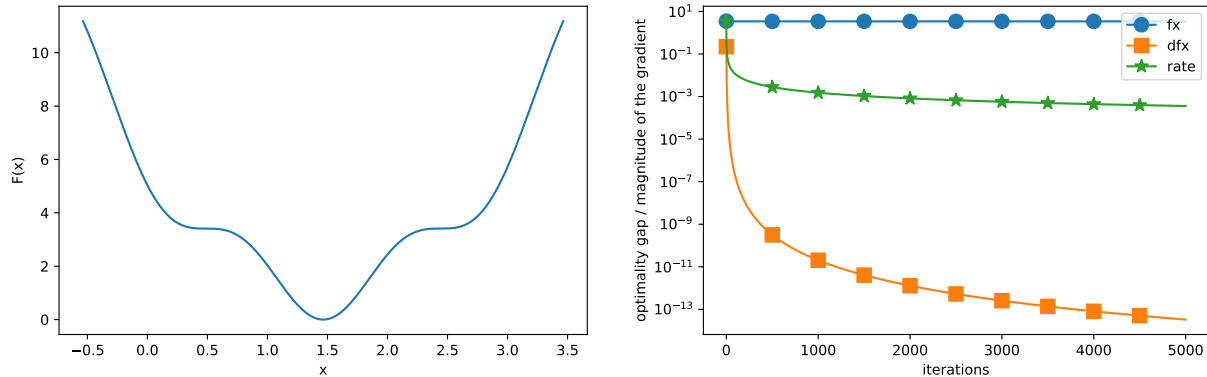


Figure 3: Plot of a one-dimensional function with a plateau being optimized (left) and the convergence of the magnitude of the gradient descent with its corresponding rate (right).

point ($c \approx 2.15$). It is trivial to show, that the optimal value is 0 and it is achieved for $x = \frac{\pi}{c} \approx 1.46$. The shape of the function around the optimum can be found on the left plot of Figure 3.

We used a 1-dimensional gradient descent method for the convergence and we reported on three quantities: The function suboptimality (fx), the magnitude of the gradient (dfx) and the rate predicted by the theory ($rate$). Theory states that either the function suboptimality or the magnitude of the gradient has to be below the predicted rate, which is what we observe in the right plot of Figure 3 as well. Note that the theory focuses on worst case bounds, which is possibly the case why the difference between the rate and the magnitudes is so huge.

References

- [1] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.
- [2] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.
- [3] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.
- [4] Soham De, Abhay Yadav, David Jacobs, and Tom Goldstein. Big batch SGD: Automated inference using adaptive batch sizes. *arXiv preprint arXiv:1610.05792*, 2016.
- [5] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Nonconvex finite-sum optimization via SCSG methods. *arXiv preprint arXiv:1706.09156*, 2017.
- [6] Bin Gao, Xin Liu, Xiaojun Chen, and Ya-xiang Yuan. On the lojasiewicz exponent of the quadratic sphere constrained optimization problem. *arXiv preprint arXiv:1611.08781*, 2016.

- [7] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016.
- [8] Augustin Cauchy. Méthode générale pour la résolution des systemes déquations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.
- [9] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- [10] Dennis Leventhal and Adrian S Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.
- [11] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.
- [12] Peter Richtárik and Martin Takáč. On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, 10(6):1233–1243, 2016.
- [13] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016.
- [14] Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling II: Expected separable overapproximation. *Optimization Methods and Software*, 31(5):858–884, 2016.
- [15] Dominik Csiba and Peter Richtárik. Importance sampling for minibatches. *arXiv preprint arXiv:1602.02283*, 2016.
- [16] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *International Conference on Machine Learning*, pages 1632–1641, 2015.
- [17] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.
- [18] Roy Frostig, Rong Ge, Sham Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *ICML*, pages 2540–2548, 2015.
- [19] Dominik Csiba, Zheng Qu, and Peter Richtárik. Stochastic dual coordinate ascent with adaptive probabilities. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 674–683, 2015.
- [20] Zheng Qu, Peter Richtárik, Martin Takáč, and Olivier Fercoq. Sdna: stochastic dual newton ascent for empirical risk minimization. In *International Conference on Machine Learning*, pages 1823–1832, 2016.
- [21] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

A Some basic properties of WPL functions

Here we establish some basic properties of smooth WPL functions.

Proposition 21. *Assume $f \in \mathcal{W}_{PL}(\mu)$ for some $\mu > 0$. Then*

$$(i) \mathcal{X}^* = \{\mathbf{x} \in \mathbb{R}^n : \nabla f(\mathbf{x}) = 0\}$$

$$(ii) \mathcal{X}^* = \{\mathbf{x} \in \mathbb{R}^n : \xi(\mathbf{x}) = 0\}$$

(iii) *If $\nabla f(\mathbf{x})$ is continuous, then \mathcal{X}^* is convex.*

(iv) *Assume $\nabla f(\mathbf{x})$ is continuous. Then f is non-decreasing on all rays emanating from \mathbf{x}^* . That is, $\psi(t) := f(\mathbf{x}^* + t(\mathbf{x} - \mathbf{x}^*))$ is non-decreasing on $t \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.*

Proposition 22. *Assume that f has continuous gradient. If there exists a constant $c > 0$ such that*

$$\int_0^1 \|\nabla f(\mathbf{x}^* + t(\mathbf{x} - \mathbf{x}^*))\| dt \leq c \|\nabla f(\mathbf{x})\|, \quad \mathbf{x} \in \mathbb{R}^n,$$

then $f \in \mathcal{W}_{PL}(\frac{1}{c^2})$.

Proof. Using the fundamental theorem of calculus, Cauchy-Schwartz inequality, and then applying the assumption, we get

$$\begin{aligned} \xi(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}^*) &= \int_0^1 \langle \nabla f(\mathbf{x}^* + t(\mathbf{x} - \mathbf{x}^*)), \mathbf{x} - \mathbf{x}^* \rangle dt \\ &\leq \int_0^1 \|\nabla f(\mathbf{x}^* + t(\mathbf{x} - \mathbf{x}^*))\| \cdot \|\mathbf{x} - \mathbf{x}^*\| dt \\ &\leq c \|\nabla f(\mathbf{x})\| \|\mathbf{x} - \mathbf{x}^*\|. \end{aligned}$$

□

Let us shed light on the above result. If f is convex, then the directional derivative $\psi(t) := \langle \nabla f(\mathbf{x}^* + t(\mathbf{x} - \mathbf{x}^*)), \mathbf{x} - \mathbf{x}^* \rangle$ is an increasing function of t , and hence can be bounded above on $[0, 1]$ by $\psi(1) = \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \leq \|\nabla f(\mathbf{x})\| \|\mathbf{x} - \mathbf{x}^*\|$. It follows that $f \in \mathcal{W}_{PL}(1)$, which we already know.

Theorem 23. *The following hold:*

1. *If $\mu_1 \geq \mu_2$, then $\mathcal{W}_{PL}(\mu_1) \subseteq \mathcal{W}_{PL}(\mu_2)$.*
2. *If f is convex, then $f \in \mathcal{W}_{PL}(\mu)$ for all $\mu \leq 1$.*
3. *If $f \in \mathcal{W}_{PL}(\mu)$, then $af + b \in \mathcal{W}_{PL}(\mu)$ for all $a \geq 0$ and $b \in \mathbb{R}$.*
4. *If $f \in \mathcal{C}^1(L)$, then $\mathcal{S}_{PL}(\mu) \subseteq \mathcal{W}_{PL}(4\mu/L)$.*
5. *Let $f \in \mathcal{C}^1(L)$, fix $\mathbf{x}^* \in \mathcal{X}^*$, and assume that there exists a constant $c > 0$ such that $\|\mathbf{x} - \mathbf{x}^*\| \leq c \|\nabla f(\mathbf{x})\|$ for all $\mathbf{x} \in \mathbb{R}^n$. Then $f \in \mathcal{W}_{PL}(\frac{4}{L^2 c^2})$.*

6. Assume that there exists a constant $c > 0$ such that $\|\nabla f(\mathbf{x})\| \leq c$ for all $\mathbf{x} \in \mathbb{R}^n$. If $f \in \mathcal{W}_{PL}(\mu)$, then $\xi(\mathbf{x}) \leq \frac{c}{\sqrt{\mu}}\|\mathbf{x} - \mathbf{x}^*\|$ for all $\mathbf{x} \in \mathbb{R}^n$. That is, f is Lipschitz on each ray emanating from \mathbf{x}^* with Lipschitz constant $\frac{c}{\sqrt{\mu}}$.

7. Assume $f \in \mathcal{W}_{PL}(\mu)$ with $\mu > 0$. If

$$\frac{\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle}{\|\nabla f(\mathbf{x})\| \|\mathbf{x} - \mathbf{x}^*\|} \geq \frac{1}{\sqrt{\mu}},$$

then f satisfies the restricted convexity property:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle, \quad \mathbf{x} \in \mathbb{R}^n.$$

Proof. 1. Obvious.

2. This was established in the introduction for $\mu = 1$. It only remains to apply 1) to conclude that 2) holds for all $\mu \leq 1$.

3. Obvious.

4. If $f \in \mathcal{S}_{PL}(\mu)$, then

$$\frac{\|\nabla f(\mathbf{x})\|^2 \cdot \|\mathbf{x} - \mathbf{x}^*\|^2}{\xi^2(\mathbf{x})} \stackrel{(2)}{\geq} \frac{2\mu\|\mathbf{x} - \mathbf{x}^*\|^2}{\xi(\mathbf{x})} \stackrel{(4)}{\geq} \frac{4\mu}{L}.$$

5. We have $\xi(\mathbf{x}) \leq \frac{L}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$ using smoothness. Combining this with the assumption $\|\mathbf{x} - \mathbf{x}^*\| \leq c\|\nabla f(\mathbf{x})\|$ from the claim we will prove $f \in \mathcal{W}_{PL}(\frac{4}{L^2c^2})$ from the definition (7)

$$\frac{2}{Lc}\xi(\mathbf{x}) \leq \frac{1}{c}\|\mathbf{x} - \mathbf{x}^*\|^2 \leq \|\nabla f(\mathbf{x})\| \cdot \|\mathbf{x} - \mathbf{x}^*\|.$$

6. Directly using the definition of $f \in \mathcal{W}_{PL}(\mu)$ with the assumption $\|\nabla f(\mathbf{x})\| \leq c$ we get

$$\xi(\mathbf{x}) \stackrel{(7)}{\leq} \frac{1}{\sqrt{\mu}}\|\nabla f(\mathbf{x})\|\|\mathbf{x} - \mathbf{x}^*\| \leq \frac{c}{\sqrt{\mu}}\|\mathbf{x} - \mathbf{x}^*\|.$$

7. Combining the definition of $f \in \mathcal{W}_{PL}(\mu)$ (7) with the assumption from the claim we get that

$$\sqrt{\mu}\xi(\mathbf{x}) \stackrel{(7)}{\leq} \|\nabla f(\mathbf{x})\| \cdot \|\mathbf{x} - \mathbf{x}^*\| \leq \sqrt{\mu}\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle.$$

Dividing both sides by $\sqrt{\mu}$ and adding $f(\mathbf{x}^*)$ we get the restricted convexity. □

B Proofs

B.1 Proof of Lemma 13

Proof. From the definition of Algorithm 1 we have that $\mathbf{x}^+ = \mathbf{x} + \mathbf{u}_{[S]}^*$, where

$$\mathbf{u}^* := \arg \min_{\mathbf{u} \in \mathbb{R}^n} \left\{ \langle \nabla_{[S]} f(\mathbf{x}), \mathbf{u} \rangle + \frac{1}{2} \mathbf{u}^\top \mathbf{M}_{[S]} \mathbf{u} + \sum_{i \in S} [g_i(x_i + u_i) - g_i(x_i)] \right\} \quad (67)$$

It follows that

$$\begin{aligned} \xi(\mathbf{x}^+) &= \xi(\mathbf{x} + \mathbf{u}^*) \\ &\stackrel{(19)}{=} F(\mathbf{x} + \mathbf{u}^*) - F(\mathbf{x}^*) \\ &\stackrel{(13)}{=} f(\mathbf{x} + \mathbf{u}_{[S]}^*) + g(\mathbf{x} + \mathbf{u}_{[S]}^*) - F(\mathbf{x}^*) \\ &\stackrel{(14)}{\leq} f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{u}_{[S]}^* \rangle + \frac{1}{2} (\mathbf{u}_{[S]}^*)^\top \mathbf{M} \mathbf{u}_{[S]}^* + g(\mathbf{x} + \mathbf{u}_{[S]}^*) - F(\mathbf{x}^*) \\ &\stackrel{(13)}{=} F(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{u}_{[S]}^* \rangle + \frac{1}{2} (\mathbf{u}_{[S]}^*)^\top \mathbf{M} \mathbf{u}_{[S]}^* + g(\mathbf{x} + \mathbf{u}_{[S]}^*) - g(\mathbf{x}) - F(\mathbf{x}^*) \\ &\stackrel{(19)}{=} \xi(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{u}_{[S]}^* \rangle + \frac{1}{2} (\mathbf{u}_{[S]}^*)^\top \mathbf{M} \mathbf{u}_{[S]}^* + g(\mathbf{x} + \mathbf{u}_{[S]}^*) - g(\mathbf{x}) \\ &\stackrel{(15)}{=} \xi(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{u}_{[S]}^* \rangle + \frac{1}{2} (\mathbf{u}_{[S]}^*)^\top \mathbf{M} \mathbf{u}_{[S]}^* + \sum_{i \in S} [g_i(x_i + u_i^*) - g_i(x_i)] \\ &\stackrel{(67)}{=} \xi(\mathbf{x}) + \min_{\mathbf{u} \in \mathbb{R}^n} \left\{ \langle \nabla_{[S]} f(\mathbf{x}), \mathbf{u} \rangle + \frac{1}{2} \mathbf{u}^\top \mathbf{M}_{[S]} \mathbf{u} + \sum_{i \in S} [g_i(x_i + u_i) - g_i(x_i)] \right\} \\ &\stackrel{(24)}{\leq} \xi(\mathbf{x}) + \theta(S, \mathbf{x}) \cdot \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x} + \mathbf{y}) - g(\mathbf{x}) \right\} \\ &\stackrel{(21)}{=} [1 - \theta(S, \mathbf{x}) \cdot \mu(\mathbf{x})] \cdot \xi(\mathbf{x}) \end{aligned}$$

Note that the inequality in the one-to-last line might happen in the case, when $\mathbf{x} \notin \mathcal{X}$ in the definition of the proportion function in (24). Chaining up the resulting expressions from \mathbf{x}^K all the way back to \mathbf{x}^0 proves the K -step bound. \square

B.2 Proof of Theorem 15

Proof. Using the result of Lemma 24, we have the bound (71). To get the result, we will use the bound $(a + b)^2 \geq 4ab$, which holds for any $a, b > 0$. Specifically, we will use it for $a = \xi(\mathbf{x})$ and $b = \frac{\lambda_F}{2} \|\mathbf{x} - \mathbf{x}^*\|^2$ in the expression in (71) to get

$$\begin{aligned} &-L \cdot \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x} + \mathbf{y}) - g(\mathbf{x}) \right\} \\ &\geq \xi(\mathbf{x}) \cdot \min \left\{ \frac{L}{2}, \frac{L\lambda_F}{\lambda_F - \lambda_f + L} \right\}, \end{aligned}$$

which is the claimed result. \square

B.3 Proof of Theorem 16

Proof. The one-step bound (26) combined with the definition of the class \mathcal{S}_{PL}^g (7) gives

$$\xi(\mathbf{x}^{k+1}) \stackrel{(26)}{\leq} \left(1 - \theta(S_k, \mathbf{x}^k) \cdot \mu(\mathbf{x}^k)\right) \cdot \xi(\mathbf{x}^k) \stackrel{(7)}{\leq} (1 - \theta(S_k, \mathbf{x}^k) \cdot \mu) \cdot \xi(\mathbf{x}^k).$$

Now, by taking full expectation over the whole sampling procedure on both sides and using the definition of μ_k (53) we get

$$\begin{aligned} \mathbf{E}[\xi(\mathbf{x}^{k+1})] &\leq \mathbf{E} \left[\left(1 - \theta(S_k, \mathbf{x}^k) \mu\right) \cdot \xi(\mathbf{x}^k) \right] \\ &= \mathbf{E}[\xi(\mathbf{x}^k)] - \mu \mathbf{E}[\theta(S_k, \mathbf{x}^k) \cdot \xi(\mathbf{x}^k)] \\ &\stackrel{(53)}{=} (1 - \mu_k) \cdot \mathbf{E}[\xi(\mathbf{x}^k)]. \end{aligned}$$

To establish the convergence rate (54), we use the estimate $1 - s \leq e^{-s}$ to get $\mathbf{E}[\xi(\mathbf{x}^{k+1})] \leq e^{-\mu_k} \mathbf{E}[\xi(\mathbf{x}^k)]$, which we can simply chain together repeatedly to get

$$\mathbf{E}[\xi(\mathbf{x}^{k+1})] \leq e^{-\sum_{k=0}^{K-1} \mu_k} \xi(\mathbf{x}^0).$$

Setting the right-hand side less or equal to ϵ and rearranging we finally get (54). \square

B.4 Proof of Theorem 18

Proof. Using the result of Lemma 24, we have the bound (71). Plugging $\lambda_F = \lambda_f = 0$ into the expression in (71) we get

$$\begin{aligned} &-L \cdot \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x} + \mathbf{y}) - g(\mathbf{x}) \right\} \\ &\geq \xi^2(\mathbf{x}) \cdot \min \left\{ \frac{L}{2\xi(\mathbf{x})}, \frac{1}{2\|\mathbf{x} - \mathbf{x}^*\|^2} \right\}, \end{aligned}$$

which is the first part of the claimed result. As for the second part, we can directly bound $\xi(\mathbf{x}) \leq \xi(\mathbf{x}^0)$, as these are the only pairs of \mathbf{x}, \mathbf{x}^0 we need to consider according to Definition 17. Similarly, as the function value is bounded, the quantities $\|\mathbf{x} - \mathbf{x}^*\|$ can be upper bounded by the largest distance in the level set of $f(\mathbf{x}^0)$, which is given by R in (60). Combining these arguments, we get that

$$\mu(\mathbf{x}) \geq \xi(\mathbf{x}) \cdot \min \left\{ \frac{L}{2\xi(\mathbf{x})}, \frac{1}{2\|\mathbf{x} - \mathbf{x}^*\|^2} \right\} \geq \xi(\mathbf{x}) \cdot \min \left\{ \frac{L}{2\xi(\mathbf{x}^0)}, \frac{1}{2R^2} \right\},$$

which is the claimed result. \square

B.5 Proof of Theorem 19

Proof. Combining the one-step bound (26) with the definition of the class \mathcal{W}_{PL}^g (57) we get that

$$\begin{aligned} \xi(\mathbf{x}^{k+1}) &\stackrel{(26)}{\leq} [1 - \theta(S_k, \mathbf{x}^k) \cdot \mu(\mathbf{x}^k)] \cdot \xi(\mathbf{x}^k) \\ &\stackrel{(57)}{\leq} [1 - \theta(S_k, \mathbf{x}^k) \cdot \rho(\mathbf{x}^0) \cdot \xi(\mathbf{x}^k)] \cdot \xi(\mathbf{x}^k) \end{aligned}$$

Now, taking full expectation over the whole sampling process on both sides and using the definition of μ_k (61) we get

$$\begin{aligned}\mathbf{E}[\xi(\mathbf{x}^{k+1})] &\leq \mathbf{E}\left[\left(1 - \theta(S_k, \mathbf{x}^k) \cdot \rho(\mathbf{x}^0) \cdot \xi(\mathbf{x}^k)\right) \cdot \xi(\mathbf{x}^k)\right] \\ &= \mathbf{E}[\xi(\mathbf{x}^k)] - \rho(\mathbf{x}^0)\mathbf{E}[\theta(S_k, \mathbf{x}^k) \cdot (\xi(\mathbf{x}^k))^2] \\ &\stackrel{(61)}{=} \left(1 - \mu_k \mathbf{E}[\xi(\mathbf{x}^k)]\right) \cdot \mathbf{E}[\xi(\mathbf{x}^k)].\end{aligned}$$

Observe, that $\{\mu_k\}_{k=0}^K$ and $\{\mathbf{E}[\xi(\mathbf{x}^k)]\}_{k=0}^K$ are both positive scalars, therefore we can use Lemma 4 to get the bound

$$\mathbf{E}[\xi(\mathbf{x}^K)] \leq \frac{\xi(\mathbf{x}^0)}{1 + \xi(\mathbf{x}^0) \sum_{k=0}^{K-1} \mu_k}.$$

Putting the right-hand side less than ϵ and rearranging leads to the claimed result (62). \square

B.6 Proof of Theorem 20

Proof. If part (i) from the claim holds, we are done. Now assume on the contrary, that (i) does not hold, i.e.,

$$-L \cdot \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ \langle \nabla f(\mathbf{x}^k), \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x}^k + \mathbf{y}) - g(\mathbf{x}^k) \right\} \geq \epsilon \quad (68)$$

for all $k \in \{0, \dots, K-1\}$. It follows, that

$$\mu(\mathbf{x}^k) \stackrel{(22)}{=} \frac{-L \cdot \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ \langle \nabla f(\mathbf{x}^k), \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x}^k + \mathbf{y}) - g(\mathbf{x}^k) \right\}}{\xi(\mathbf{x}^k)} \stackrel{(68)}{\geq} \frac{\epsilon}{\xi(\mathbf{x}^k)}.$$

for all $k \in \{0, \dots, K-1\}$. Using the result from Lemma 13, we have that $\xi(\mathbf{x}^k) \leq \xi(\mathbf{x}^0)$ for all k , which we can use to further bound

$$\mu(\mathbf{x}^k) \geq \frac{\epsilon}{\xi(\mathbf{x}^0)}. \quad (69)$$

Using (26) combined with the above result (69) we get

$$\begin{aligned}\xi(\mathbf{x}^K) &\stackrel{(26)}{\leq} [1 - \mu(\mathbf{x}^{K-1}) \cdot \theta(S_{K-1}, \mathbf{x}^{K-1})] \xi(\mathbf{x}^{K-1}) \\ &\stackrel{(69)}{\leq} \left[1 - \frac{\epsilon \theta(S_{K-1}, \mathbf{x}^{K-1})}{\xi(\mathbf{x}^0)}\right] \xi(\mathbf{x}^{K-1})\end{aligned}$$

Taking the expectation over the whole sampling procedure on both sides and using the definition of μ_k in (63) we get

$$\begin{aligned}\mathbf{E}[\xi(\mathbf{x}^K)] &\leq \mathbf{E}\left[\left(1 - \frac{\epsilon \theta(S_{K-1}, \mathbf{x}^{K-1})}{\xi(\mathbf{x}^0)}\right) \xi(\mathbf{x}^{K-1})\right] \\ &= \mathbf{E}[\xi(\mathbf{x}^{K-1})] - \frac{\epsilon}{\xi(\mathbf{x}^0)} \mathbf{E}[\theta(S_{K-1}, \mathbf{x}^{K-1}) \cdot \xi(\mathbf{x}^{K-1})] \\ &\stackrel{(63)}{=} \left(1 - \frac{\epsilon \mu_{K-1}}{\xi(\mathbf{x}^0)}\right) \mathbf{E}[\xi(\mathbf{x}^{K-1})]\end{aligned} \quad (70)$$

Combining the above inequality (70) with $(1 - z) \leq \exp(-z)$ repeatedly, we get

$$\begin{aligned}
\mathbf{E}[\xi(\mathbf{x}^K)] &\stackrel{(70)}{\leq} \left(1 - \frac{\epsilon\mu_{K-1}}{\xi(\mathbf{x}^0)}\right) \mathbf{E}[\xi(\mathbf{x}^{K-1})] \\
&\leq \exp\left(-\frac{\epsilon\mu_{K-1}}{\xi(\mathbf{x}^0)}\right) \mathbf{E}[\xi(\mathbf{x}^{K-1})] \\
&\stackrel{(70)}{\leq} \dots \\
&\stackrel{(70)}{\leq} \exp\left(-\frac{\epsilon\sum_{k=0}^{K-1}\mu_k}{\xi(\mathbf{x}^0)}\right) \xi(\mathbf{x}^0) \\
&\stackrel{(64)}{\leq} \epsilon,
\end{aligned}$$

where the last line follows from comparing the logarithms of both sides. This proves (ii). \square

C Technical Lemmas

C.1 Lemma 24

Lemma 24. *Let f be λ_f -strongly convex (51) with $\lambda_f \geq 0$ and F be λ_F -strongly convex (50) with $\lambda_F \geq 0$. Then*

$$\begin{aligned}
&-\min_{\mathbf{y} \in \mathbb{R}^n} \left\{ \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x} + \mathbf{y}) - g(\mathbf{x}) \right\} \\
&\geq \min \left\{ \frac{1}{2} \xi(\mathbf{x}), \frac{\left(\xi(\mathbf{x}) + \frac{\lambda_F}{2} \|\mathbf{x} - \mathbf{x}^*\|^2\right)^2}{2(\lambda_F - \lambda_f + L) \|\mathbf{x} - \mathbf{x}^*\|^2} \right\}. \tag{71}
\end{aligned}$$

Proof. Let

$$\beta = \min \left\{ 1, \frac{\xi(\mathbf{x}) + \frac{\lambda_F}{2} \|\mathbf{x} - \mathbf{x}^*\|^2}{(\lambda_F - \lambda_f + L) \|\mathbf{x} - \mathbf{x}^*\|^2} \right\}. \tag{72}$$

Observe, that $\beta = 1$ implies, that

$$(\lambda_F - \lambda_f + L) \|\mathbf{x} - \mathbf{x}^*\|^2 \leq \xi(\mathbf{x}) + \frac{\lambda_F}{2} \|\mathbf{x} - \mathbf{x}^*\|^2,$$

from which it follows that

$$\frac{\lambda_f - L}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \geq -\frac{1}{2} \xi(\mathbf{x}) + \frac{\lambda_F}{4} \|\mathbf{x} - \mathbf{x}^*\|^2 > -\frac{1}{2} \xi(\mathbf{x}) \tag{73}$$

Now, it follows that

$$\begin{aligned}
& - \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x} + \mathbf{y}) - g(\mathbf{x}) \right\} \\
& \stackrel{(13)}{=} F(\mathbf{x}) - \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x} + \mathbf{y}) \right\} \\
& \stackrel{(51)}{\geq} F(\mathbf{x}) - \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ f(\mathbf{x} + \mathbf{y}) - \frac{\lambda_f}{2} \|\mathbf{y}\|^2 + \frac{L}{2} \|\mathbf{y}\|^2 + g(\mathbf{x} + \mathbf{y}) \right\} \\
& \stackrel{(13)}{=} F(\mathbf{x}) - \min_{\mathbf{y} \in \mathbb{R}^n} \left\{ F(\mathbf{x} + \mathbf{y}) + \frac{L - \lambda_f}{2} \|\mathbf{y}\|^2 \right\} \quad (\text{let } \mathbf{y} = \beta(\mathbf{x}^* - \mathbf{x})) \\
& \geq F(\mathbf{x}) - F(\beta\mathbf{x}^* + (1 - \beta)\mathbf{x}) - \frac{\beta^2(L - \lambda_f)}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \\
& \stackrel{(50)}{\geq} F(\mathbf{x}) - \beta F(\mathbf{x}^*) - (1 - \beta)F(\mathbf{x}) + \left(\frac{\lambda_F \beta(1 - \beta)}{2} - \frac{\beta^2(L - \lambda_f)}{2} \right) \|\mathbf{x} - \mathbf{x}^*\|^2 \\
& \stackrel{(19)}{=} \beta \left(\xi(\mathbf{x}) + \frac{\lambda_F}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right) - \beta^2 \frac{(\lambda_F - \lambda_f + L)}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \\
& \stackrel{(72)}{=} \min \left\{ \xi(\mathbf{x}) + \frac{\lambda_f - L}{2} \|\mathbf{x} - \mathbf{x}^*\|^2, \frac{\left(\xi(\mathbf{x}) + \frac{\lambda_F}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right)^2}{2(\lambda_F - \lambda_f + L) \|\mathbf{x} - \mathbf{x}^*\|^2} \right\} \\
& \stackrel{(73)}{>} \min \left\{ \frac{1}{2} \xi(\mathbf{x}), \frac{\left(\xi(\mathbf{x}) + \frac{\lambda_F}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 \right)^2}{2(\lambda_F - \lambda_f + L) \|\mathbf{x} - \mathbf{x}^*\|^2} \right\}.
\end{aligned}$$

□

C.2 Lemma 25

Lemma 25. *Let f be a given function and let X, Y be random variables such that $\mathbf{Prob}(f(X) \geq 0) = 1$ and $\mathbf{Prob}(\mathbf{E}[Y | X] \geq 0) = 1$. Additionally, let $c > 0$ be a scalar such that*

$$c \leq \mathbf{E}[Y | X] \tag{74}$$

Then it holds that

$$\mathbf{E}[f(X)Y] \geq c\mathbf{E}[f(X)]. \tag{75}$$

Proof. Using the tower property $\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]]$ combined with assumption (74) we get

$$\mathbf{E}[f(X)Y] = \mathbf{E}[\mathbf{E}[f(X)Y | X]] = \mathbf{E}[f(X)\mathbf{E}[Y | X]] \stackrel{(74)}{\geq} \mathbf{E}[f(X)c] = c\mathbf{E}[f(X)],$$

which concludes the proof. □

D Notation Glossary

Notation	Description
\mathbb{R}	the set of real numbers
\mathbb{R}_+	the set of positive real numbers
$\bar{\mathbb{R}}$	the set $\mathbb{R} \cup \{+\infty\}$
\mathbf{x}	a vector
x_i	the i -th entry of the vector \mathbf{x}
\mathbf{X}	a matrix
$\mathbf{X}_{:i}$	the i -th row of the matrix \mathbf{X}
$\mathbf{X}_{:j}$	the j -th column of the matrix \mathbf{X}
X_{ij}	the entry at the i -th row and j -th column of the matrix \mathbf{X}
$[n]$	a shorthand for the set $\{1, \dots, n\}$, usually containing the coordinates of the space
\mathbf{x}_S	a $ S $ -dimensional vector containing entries of \mathbf{x} with indices in S
$\mathbf{x}_{[S]}$	the vector \mathbf{x} with the entries with indices outside S zeroed out
\mathbf{X}_S	a $ S \times S $ submatrix of \mathbf{X} containing only rows and columns with indices in S
$\mathbf{X}_{[S]}$	the matrix \mathbf{X} with all entries on columns or rows outside of S zeroed out
$\mathbf{X}_{[S]}^{-1}$	the $n \times n$ matrix containing $(\mathbf{X}_S)^{-1}$ at the rows and columns indicated by S .
f	a smooth function from \mathbb{R}^n to \mathbb{R} (see Def. 8)
g	a separable (9) and possibly non-smooth function from \mathbb{R}^n to \mathbb{R}
F	the objective function from \mathbb{R}^n to \mathbb{R} , defined as $f + g$ (13)
$\nabla_S f(\mathbf{x})$	a shorthand for $(\nabla f(\mathbf{x}))_S$
$\nabla_{[S]} f(\mathbf{x})$	a shorthand for $(\nabla f(\mathbf{x}))_{[S]}$
$\xi(\mathbf{x})$	the optimality gap defined in (19)
$\lambda(\mathbf{x})$	the auxiliary function defined in (20)
$\mu(\mathbf{x})$	the forcing function defined in (21)
$\theta(S, \mathbf{x})$	the proportion function defined in (24)
\mathcal{X}^*	the set of global minimizers of F
\mathcal{X}	the set of vectors with nonzero $\lambda(\mathbf{x})$ (23)
$\lambda_{\min}(\mathbf{M})$	the smallest eigenvalue of a square matrix \mathbf{M}
$\lambda_{\max}(\mathbf{M})$	the largest eigenvalue of a square matrix \mathbf{M}
$\mathbf{M}, L\mathbf{I}$	the matrices defining \mathbf{M} -smoothness of f (14) for smooth and non-smooth cases
L_τ	the smoothness parameter defined as $\max_{S: S =\tau} \{\lambda_{\max}(\mathbf{M}_S)\}$
λ_F, λ_f	strong convexity parameters of F and f , respectively
$U_S(\mathbf{x}, \mathbf{u})$	the quadratic upper bound function at a given point \mathbf{x} , defined in (16)

Table 3: Notation Glossary.