Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition

Hamed Karimi, Julie Nutini, and Mark Schmidt

Department of Computer Science, University of British Columbia
Vancouver, British Columbia, Canada
{hamedkarim@gmail.com},{jnutini,schmidtm}@cs.ubc.ca

Abstract. In 1963, Polyak proposed a simple condition that is sufficient to show a global linear convergence rate for gradient descent. This condition is a special case of the Łojasiewicz inequality proposed in the same year, and it does not require strong convexity (or even convexity). In this work, we show that this much-older Polyak-Łojasiewicz (PL) inequality is actually weaker than the main conditions that have been explored to show linear convergence rates without strong convexity over the last 25 years. We also use the PL inequality to give new analyses of randomized and greedy coordinate descent methods, sign-based gradient descent methods, and stochastic gradient methods in the classic setting (with decreasing or constant step-sizes) as well as the variance-reduced setting. We further propose a generalization that applies to proximal-gradient methods for non-smooth optimization, leading to simple proofs of linear convergence of these methods. Along the way, we give simple convergence results for a wide variety of problems in machine learning: least squares, logistic regression, boosting, resilient backpropagation, L1-regularization, support vector machines, stochastic dual coordinate ascent, and stochastic variance-reduced gradient methods.

1 Introduction

Fitting most machine learning models involves solving some sort of optimization problem. Gradient descent, and variants of it like coordinate descent and stochastic gradient, are the workhorse tools used by the field to solve very large instances of these problems. In this work we consider the basic problem of minimizing a smooth function and the convergence rate of gradient descent methods. It is well-known that if f is strongly-convex, then gradient descent achieves a global linear convergence rate for this problem [Nesterov, 2004]. However, many of the fundamental models in machine learning like least squares and logistic regression yield objective functions that are convex but not strongly-convex. Further, if f is only convex, then gradient descent only achieves a sub-linear rate.

This situation has motivated a variety of alternatives to strong convexity (SC) in the literature, in order to show that we can obtain linear convergence rates for problems like least squares and logistic regression. One of the oldest of these conditions is the *error bounds* (EB) of Luo and Tseng [1993], but four other recently-considered conditions are *essential strong convexity* (ESC) [Liu et al., 2014], *weak strong convexity* (WSC) [Necoara et al., 2015], the *restricted secant inequality* (RSI) [Zhang and Yin, 2013], and the *quadratic growth* (QG) condition [Anitescu, 2000]. Some of these conditions have different names in the special case of convex functions. For example, a convex function satisfying RSI is said to satisfy *restricted strong convexity* (RSC) [Zhang and Yin, 2013]. Names describing convex functions satisfying QG include *optimal strong convexity* (OSC) [Liu and Wright, 2015], *semi-strong convexity* (SSC) [Gong and Ye, 2014], and (confusingly) WSC [Ma et al., 2015]. The proofs of linear convergence under all of these relaxations are typically not straightforward, and it is rarely discussed how these conditions relate to each other.

In this work, we consider a much older condition that we refer to as the Polyak-Łojasiewicz (PL) inequality. This inequality was originally introduced by Polyak [1963], who showed that it is a sufficient condition for gradient descent to achieve a linear convergence rate. We describe it as the PL inequality because it is also a special case of the inequality introduced in the same year by Łojasiewicz [1963]. We review the PL inequality in the next section and how it leads to a trivial proof of the linear convergence rate of gradient descent. Next, in terms of showing a global linear convergence rate to the optimal solution, we show that the PL inequality is *weaker* than all of the more recent conditions discussed in the previous paragraph. This suggests that we can replace the long and complicated proofs under any of the conditions above with simpler proofs based on the PL inequality. Subsequently, we show how

this result implies gradient descent achieves linear rates for standard problems in machine learning like least squares and logistic regression that are not necessarily SC, and even for some non-convex problems (Section 2.3). In Section 3 we use the PL inequality to give new convergence rates for randomized and greedy coordinate descent (implying a new convergence rate for certain variants of boosting), sign-based gradient descent methods, and stochastic gradient methods in either the classical or variance-reduced setting. Next we turn to the problem of minimizing the sum of a smooth function and a simple non-smooth function. We propose a generalization of the PL inequality that allows us to show linear convergence rates for proximal-gradient methods without SC. In this setting, the new condition is equivalent to the well-known Kurdyka-Łojasiewicz (KL) condition which has been used to show linear convergence of proximal-gradient methods for certain problems like support vector machines and ℓ_1-regularized least squares [Bolte et al., 2015]. But this new alternate generalization of the PL inequality leads to shorter and simpler proofs in these cases.

2 Polyak-Łojasiewicz Inequality

We first focus on the basic unconstrained optimization problem

$$\operatorname*{argmin}_{x\in\mathbb{R}^d} f(x), \tag{1}$$

and we assume that the first derivative of f is L-Lipschitz continuous. This means that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}||y-x||^2, \tag{2}$$

for all x and y. For twice-differentiable objectives this assumption means that the eigenvalues of $\nabla^2 f(x)$ are bounded above by some L, which is typically a reasonable assumption. We also assume that the optimization problem has a non-empty solution set $\mathcal{X}^*$, and we use f^* to denote the corresponding optimal function value. We will say that a function satisfies the PL inequality if the following holds for some $\mu > 0$,

$$\frac{1}{2}||\nabla f(x)||^2 \geq \mu(f(x) - f^*), \quad \forall\, x. \tag{3}$$

This inequality simply requires that the gradient grows faster than a quadratic function as we move away from the optimal function value. Note that this inequality implies that every stationary point is a global minimum. But unlike SC, it does not imply that there is a unique solution. Linear convergence of gradient descent under these assumptions was first proved by Polyak [1963]. Below we give a simple proof of this result when using a step-size of $1/L$.

Theorem 1. *Consider problem* (1)*, where f has an L-Lipschitz continuous gradient (2), a non-empty solution set $\mathcal{X}^*$, and satisfies the PL inequality (3). Then the gradient method with a step-size of $1/L$,*

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \tag{4}$$

has a global linear convergence rate,

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Proof. By using update rule (4) in the Lipschitz inequality condition (2) we have

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}||\nabla f(x_k)||^2.$$

Now by using the PL inequality (3) we get

$$f(x_{k+1}) - f(x_k) \leq -\frac{\mu}{L}(f(x_k) - f^*).$$

Re-arranging and subtracting f^* from both sides gives us $f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)(f(x_k) - f^*)$. Applying this inequality recursively gives the result. □

Note that the above result also holds if we use the optimal step-size at each iteration, because under this choice we have

$$f(x_{k+1}) = \min_{\alpha}\{f(x_k - \alpha\nabla f(x_k))\} \leq f\left(x_k - \frac{1}{L}\nabla f(x_k)\right).$$

A beautiful aspect of this proof is its simplicity; in fact it is *simpler* than the proof of the same fact under the usual SC assumption. It is certainly simpler than typical proofs which rely on the other conditions mentioned in Section 1. Further, it is worth noting that the proof does *not* assume convexity of f. Thus, this is one of the few general results we have for global linear convergence on non-convex problems.

2.1 Relationships Between Conditions

As mentioned in the Section 1, several other assumptions have been explored over the last 25 years in order to show that gradient descent achieves a linear convergence rate. These typically assume that f is convex, and lead to more complicated proofs than the one above. However, it is rarely discussed how the conditions relate to each other. Indeed, all of the relationships that have been explored have only been in the context of convex functions [Bolte et al., 2015, Liu and Wright, 2015, Necoara et al., 2015, Zhang, 2015]. In Appendix A, we give the precise definitions of all conditions and also prove the result below giving relationships between the conditions.

Theorem 2. *For a function f with a Lipschitz-continuous gradient, the following implications hold:*

$$(SC) \rightarrow (ESC) \rightarrow (WSC) \rightarrow (RSI) \rightarrow (EB) \equiv (PL) \rightarrow (QG).$$

If we further assume that f is convex then we have

$$(RSI) \equiv (EB) \equiv (PL) \equiv (QG).$$

Note the equivalence between EB and PL is a special case of a more general result by Bolte et al. [2015, Theorem 5], while Zhang [2016] independently also recently gave the relationships between RSI, EB, PL, and QG.[1] This result shows that QG is the weakest assumption among those considered. However, QG allows non-global local minima so it is not enough to guarantee that gradient descent finds a global minimizer. This means that, among those considered above, *PL and the equivalent EB are the most general conditions* that allow linear convergence to a global minimizer. Note that in the convex case QG is called OSC or SSC, but the result above shows that in the convex case it is also equivalent to EB and PL (as well as RSI which is known as RSC in this case).

2.2 Invex and Non-Convex Functions

While the PL inequality does not imply convexity of f, it does imply the weaker condition of *invexity*. A function is invex if it is differentiable and there exists a vector valued function η such that for any x and y in $\mathbb{R}^n$, the following inequality holds

$$f(y) \geq f(x) + \nabla f(x)^T \eta(x, y).$$

We obtain convex functions as the special case where $\eta(x, y) = y - x$.

Invexity was first introduced by Hanson [1981], and has been used in the context of learning output kernels [Dinuzzo et al., 2011]. Craven and Glover [1985] show that a smooth f is invex if and only if every stationary point of f is a global minimum. Since the PL inequality implies that all stationary points are global minimizers, functions satisfying the PL inequality must be invex. It is easy to see this by noting that at any stationary point $\bar{x}$ we have $\nabla f(\bar{x}) = 0$, so we have

$$0 = \frac{1}{2}\|\nabla f(\bar{x})\|^2 \geq \mu(f(x) - f^*) \geq 0,$$

where the last inequality holds because $\mu > 0$ and $f(x) \geq f^*$ for all x. This implies that $f(\bar{x}) = f^*$ and thus any stationary point must be a global minimum.

[1] Drusvyatskiy and Lewis [2016] is a recent work discussing the relationships among many of these conditions for non-smooth functions.

Theorem 2 shows that all of the previous conditions (except QG) imply invexity. The function $f(x) = x^2 + 3\sin^2(x)$ is an example of an invex but non-convex function satisfying the PL inequality (with $\mu = 1/32$). Thus, Theorem 1 implies gradient descent obtains a global linear convergence rate on this function.

Unfortunately, many complicated models have non-optimal stationary points. For example, typical deep feed-forward neural networks have sub-optimal stationary points and are thus not invex. A classic way to analyze functions like this is to consider a *global convergence phase* and a *local convergence phase*. The global convergence phase is the time spent to get "close" to a local minimum, and then once we are "close" to a local minimum the local convergence phase characterizes the convergence rate of the method. Usually, the local convergence phase starts to apply once we are locally SC around the minimizer. But this means that the local convergence phase may be arbitrarily small: for example, for $f(x) = x^2 + 3\sin^2(x)$ the local convergence rate would not even apply over the interval $x \in [-1, 1]$. If we instead defined the local convergence phase in terms of locally satisfying the PL inequality, then we see that it can be *much* larger ($x \in \mathbb{R}$ for this example).

2.3 Relevant Problems

If f is μ-SC, then it also satisfies the PL inequality with the same μ (see Appendix B). Further, by Theorem 2, f satisfies the PL inequality if it satisfies any of ESC, WSC, RSI, or EB (while for convex f, QG is also sufficient). Although it is hard to precisely characterize the general class of functions for which the PL inequality is satisfied, we note one important special case below.

Strongly-convex composed with linear: This is the case where f has the form $f(x) = g(Ax)$ for some σ-SC function g and some matrix A. In Appendix B, we show that this class of functions satisfies the PL inequality, and we note that this form frequently arises in machine learning. For example, least squares problems have the form

$$f(x) = \|Ax - b\|^2,$$

and by noting that $g(z) \triangleq \|z - b\|^2$ is SC we see that least squares falls into this category. Indeed, this class includes all convex quadratic functions.

In the case of logistic regression we have

$$f(x) = \sum_{i=1}^{n} \log(1 + \exp(b_i a_i^T x)).$$

This can be written in the form $g(Ax)$, where g is strictly convex but not SC. In cases like this where g is only strictly convex, the PL inequality will still be satisfied over any compact set. Thus, if the iterations of gradient descent remain bounded, the linear convergence result still applies. It is reasonable to assume that the iterates remain bounded when the set of solutions is finite, since each step must decrease the objective function. Thus, for practical purposes, we can relax the above condition to "strictly-convex composed with linear" and the PL inequality implies a linear convergence rate for logistic regression.

3 Convergence of Huge-Scale Methods

In this section, we use the PL inequality to analyze several variants of two of the most widely-used techniques for handling large-scale machine learning problems: coordinate descent and stochastic gradient methods. In particular, the PL inequality yields very simple analyses of these methods that apply to more general classes of functions than previously analyzed. We also note that the PL inequality has recently been used by Garber and Hazan [2015a] to analyze the Frank-Wolfe algorithm. Further, inspired by the resilient backpropagation (RPROP) algorithm of Riedmiller and Braun [1992], in Appendix C we also give a convergence rate analysis for a sign-based gradient descent method.

3.1 Randomized Coordinate Descent

Nesterov [2012] shows that randomized coordinate descent achieves a faster convergence rate than gradient descent for problems where we have d variables and it is d times cheaper to update one coordinate than it is to compute the entire gradient. The expected linear convergence rates in this previous work

rely on SC, but in this section we show that randomized coordinate descent achieves an expected linear convergence rate if we only assume that the PL inequality holds.

To analyze coordinate descent methods, we assume that the gradient is coordinate-wise Lipschitz continuous, meaning that for any x and y we have

$$f(x + \alpha e_i) \leq f(x) + \alpha \nabla_i f(x) + \frac{L}{2}\alpha^2, \quad \forall \alpha \in \mathbb{R}, \quad \forall x \in \mathbb{R}^d, \tag{5}$$

for any coordinate i, and where e_i is the ith unit vector.

Theorem 3. *Consider problem* (1)*, where f has a coordinate-wise L-Lipschitz continuous gradient (5), a non-empty solution set $\mathcal{X}^*$, and satisfies the PL inequality (3). Consider the coordinate descent method with a step-size of $1/L$,*

$$x_{k+1} = x_k - \frac{1}{L}\nabla_{i_k} f(x_k) e_{i_k}. \tag{6}$$

If we choose the variable to update i_k uniformly at random, then the algorithm has an expected linear convergence rate of

$$\mathbb{E}[f(x_k) - f^*] \leq \left(1 - \frac{\mu}{dL}\right)^k [f(x_0) - f^*].$$

Proof. By using the update rule (6) in the Lipschitz condition (5) we have

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}|\nabla_{i_k} f(x_k)|^2.$$

By taking the expectation of both sides with respect to i_k we have

$$\begin{aligned}
\mathbb{E}\left[f(x_{k+1})\right] &\leq f(x_k) - \frac{1}{2L}\mathbb{E}\left[|\nabla_{i_k} f(x_k)|^2\right] \\
&= f(x_k) - \frac{1}{2L}\sum_i \frac{1}{d}|\nabla_i f(x_k)|^2 \\
&= f(x_k) - \frac{1}{2dL}||\nabla f(x_k)||^2.
\end{aligned}$$

By using the PL inequality (3) and subtracting f^* from both sides, we get

$$\mathbb{E}[f(x_{k+1}) - f^*] \leq \left(1 - \frac{\mu}{dL}\right)[f(x_k) - f^*].$$

Applying this recursively and using iterated expectations yields the result. □

As before, instead of using $1/L$ we could perform exact coordinate optimization and the result would still hold. If we have a Lipschitz constant L_i for each coordinate and sample proportional to the L_i as suggested by Nesterov [2012], then the above argument (using a step-size of $1/L_{i_k}$) can be used to show that we obtain a faster rate of

$$\mathbb{E}[f(x_k) - f^*] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^k [f(x_0) - f^*],$$

where $\bar{L} = \frac{1}{d}\sum_{j=1}^d L_j$.

3.2 Greedy Coordinate Descent

Nutini et al. [2015] have recently analyzed coordinate descent under the greedy Gauss-Southwell (GS) rule, and argued that this rule may be suitable for problems with a large degree of sparsity. The GS rule chooses i_k according to the rule $i_k = \text{argmax}_j |\nabla_j f(x_k)|$. Using the fact that

$$\max_i |\nabla_i f(x_k)| \geq \frac{1}{d}\sum_{i=1}^d |\nabla_i f(x_k)|,$$

it is straightforward to show that the GS rule satisfies the rate above for the randomized method.

However, Nutini et al. [2015] show that a faster convergence rate can be obtained for the GS rule by measuring SC in the 1-norm. Since the PL inequality is defined on the dual (gradient) space, in order to derive an analogous result we could measure the PL inequality in the ∞-norm,

$$\frac{1}{2}\|\nabla f(x)\|_\infty^2 \geq \mu_1(f(x) - f^*).$$

Because of the equivalence between norms, this is not introducing any additional assumptions beyond that the PL inequality is satisfied. Further, if f is μ_1-SC in the 1-norm, then it satisfies the PL inequality in the ∞-norm with the same constant μ_1. By using that $|\nabla_{i_k} f(x_k)| = \|\nabla f(x_k)\|_\infty$ when the GS rule is used, the above argument can be used to show that coordinate descent with the GS rule achieves a convergence rate of

$$f(x_k) - f^* \leq \left(1 - \frac{\mu_1}{L}\right)^k [f(x_0) - f^*],$$

when the function satisfies the PL inequality in the ∞-norm with a constant of μ_1. By the equivalence between norms we have that $\mu/d \leq \mu_1$, so this is faster than the rate with random selection.

Meir and Rätsch [2003] show that we can view some variants of boosting algorithms as implementations of coordinate descent with the GS rule. They use the error bound property to argue that these methods achieve a linear convergence rate, but this property does not lead to an explicit rate. Our simple result above thus provides the first explicit convergence rate for these variants of boosting.

3.3 Stochastic Gradient Methods

Stochastic gradient (SG) methods apply to the general stochastic optimization problem

$$\underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \; f(x) = \mathbb{E}[f_i(x)], \tag{7}$$

where the expectation is taken with respect to i. These methods are typically used to optimize finite sums,

$$f(x) = \frac{1}{n}\sum_{i}^{n} f_i(x). \tag{8}$$

Here, each f_i typically represents the fit of a model on an individual training example. SG methods are suitable for cases where the number of training examples n is so large that it is infeasible to compute the gradient of all n examples more than a few times.

Stochastic gradient methods use the iteration

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k), \tag{9}$$

where α_k is the step size and i_k is a sample from the distribution over i so that $\mathbb{E}[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$. Below, we analyze the convergence rate of stochastic gradient methods under standard assumptions on f, and under both a decreasing and a constant step-size scheme.

Theorem 4. *Consider problem* (7)*. Assume that each f has an L-Lipschitz continuous gradient (2), f has a non-empty solution set $\mathcal{X}^*$, f satisfies the PL inequality (3), and $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq C^2$ for all x_k and some C. If we use the SG algorithm* (9) *with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$, then we get a convergence rate of*

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{LC^2}{2k\mu^2}.$$

If instead we use a constant $\alpha_k = \alpha < \frac{1}{2\mu}$, then we obtain a linear convergence rate up to a solution level that is proportional to α,

$$\mathbb{E}[f(x_k) - f^*] \leq (1 - 2\mu\alpha)^k [f(x_0) - f^*] + \frac{LC^2\alpha}{4\mu}.$$

Proof. By using the update rule (9) inside the Lipschitz condition (2), we have

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle f'(x_k), \nabla f_{i_k}(x_k) \rangle + \frac{L\alpha_k^2}{2} ||\nabla f_{i_k}(x_k)||^2.$$

Taking the expectation of both sides with respect to i_k we have

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})] &\leq f(x_k) - \alpha_k \langle \nabla f(x_k), \mathbb{E}\left[\nabla f_{i_k}(x_k)\right] \rangle + \frac{L\alpha_k^2}{2} \mathbb{E}[\|\nabla f_i(x_k)\|^2] \\
&\leq f(x_k) - \alpha_k ||f'(x_k)||^2 + \frac{LC^2\alpha_k^2}{2} \\
&\leq f(x_k) - 2\mu\alpha_k (f(x_k) - f^*) + \frac{LC^2\alpha_k^2}{2},
\end{aligned}$$

where the second line uses that $\mathbb{E}[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$ and $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq C^2$, and the third line uses the PL inequality. Subtracting f^* from both sides yields:

$$\mathbb{E}[f(x_{k+1}) - f^*] \leq (1 - 2\alpha_k\mu)[f(x_k) - f^*] + \frac{LC^2\alpha_k^2}{2}. \tag{10}$$

Decreasing step size: With $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ in (10) we obtain

$$\mathbb{E}[f(x_{k+1}) - f^*] \leq \tfrac{k^2}{(k+1)^2}[f(x_k) - f^*] + \tfrac{LC^2(2k+1)^2|}{8\mu^2(k+1)^4}.$$

Multiplying both sides by $(k+1)^2$ and letting $\delta_f(k) \equiv k^2\mathbb{E}[f(x_k) - f^*]$ we get

$$\begin{aligned}
\delta_f(k+1) &\leq \delta_f(k) + \frac{LC^2(2k+1)^2}{8\mu^2(k+1)^2} \\
&\leq \delta_f(k) + \frac{LC^2}{2\mu^2},
\end{aligned}$$

where the second line follows from $\frac{2k+1}{k+1} < 2$. Summing up this inequality from $k = 0$ to k and using the fact that $\delta_f(0) = 0$ we get

$$\begin{aligned}
&\delta_f(k+1) \leq \delta_f(0) + \tfrac{LC^2}{2\mu^2} \textstyle\sum_{i=0}^{k} 1 \leq \tfrac{LC^2(k+1)}{2\mu^2} \\
\Rightarrow \qquad &(k+1)^2\mathbb{E}[f(x_{k+1}) - f^*] \leq \tfrac{LC^2(k+1)}{2\mu^2}
\end{aligned}$$

which gives the stated rate.

Constant step size: Choosing $\alpha_k = \alpha$ for any $\alpha < 1/2\mu$ and applying (10) recursively yields

$$\begin{aligned}
\mathbb{E}[f(x_{k+1}) - f^*] &\leq (1 - 2\alpha\mu)^k[f(x_0) - f^*] + \frac{LC^2\alpha^2}{2} \sum_{i=0}^{k} (1 - 2\alpha\mu)^i \\
&\leq (1 - 2\alpha\mu)^k[f(x_0) - f^*] + \frac{LC^2\alpha^2}{2} \sum_{i=0}^{\infty} (1 - 2\alpha\mu)^i \\
&= (1 - 2\alpha\mu)^k[f(x_0) - f^*] + \frac{LC^2\alpha}{4\mu},
\end{aligned}$$

where the last line uses that $\alpha < 1/2\mu$ and the limit of the geometric series. □

The $O(1/k)$ rate for a decreasing step size matches the convergence rate of stochastic gradient methods under SC [Nemirovski et al., 2009]. It was recently shown using a non-trivial analysis that a stochastic Newton method could achieve an $O(1/k)$ rate for least squares problems [Bach and Moulines, 2013], but our result above shows that the basic stochastic gradient method already achieves this property (although the constants are worse than for this Newton-like method). Further, our result does not rely on convexity. Note that if we are happy with a solution of fixed accuracy, then the result with a constant step-size is perhaps the more useful strategy in practice: it supports the often-used empirical strategy of using a constant size for a long time, then halving the step-size if the algorithm appears to have stalled (the above result indicates that halving the step-size will at least halve the sub-optimality).

3.4 Finite Sum Methods

In the setting of (8) where we are minimizing a *finite* sums, it has recently been shown that there are methods that have the low iteration cost of stochastic gradient methods but that still have linear convergence rates for SC functions [Le Roux et al., 2012]. While the first methods that achieved this remarkable property required a *memory* of previous gradient values, the stochastic variance-reduced gradient (SVRG) method of Johnson and Zhang [2013] does not have this drawback. Gong and Ye [2014] show that SVRG has a linear convergence rate without SC under the weaker assumption of QG plus convexity (where QG is equivalent to PL). We review how the analysis of Johnson and Zhang [2013] can be easily modified to give a similar result in Appendix D. A related result appears in Garber and Hazan [2015b], who assume that f is SC but do not assume that the individual functions are convex. More recent analyses by Reddi et al. [2016a,b] have considered these types of methods under the PL inequality without convexity assumptions.

4 Proximal-Gradient Generalization

A generalization of the PL inequality for non-smooth optimization is the KL inequality [Kurdyka, 1998, Bolte et al., 2008]. The KL inequality has been used to analyze the convergence of the classic proximal-point algorithm [Attouch and Bolte, 2009] as well as a variety of other optimization methods [Attouch et al., 2013]. In machine learning, a popular generalization of gradient descent is proximal-gradient methods. Bolte et al. [2015] show that the proximal-gradient method has a linear convergence rate for functions satisfying the KL inequality, while Li and Pong [2016] give a related result. The set of problems satisfying the KL inequality notably includes problems like support vector machines and ℓ_1-regularized least squares, implying that the algorithm has a linear convergence rate for these problems. In this section we propose a different generalization of the PL inequality that leads to a simpler linear convergence rate analysis for the proximal-gradient method as well as its coordinate-wise variant.

Proximal-gradient methods apply to problems of the form

$$\operatorname*{argmin}_{x\in\mathbb{R}^d} F(x) = f(x) + g(x), \tag{11}$$

where f is a differentiable function with an L-Lipschitz continuous gradient and g is a simple but potentially non-smooth convex function. Typical examples of simple functions g include a scaled ℓ_1-norm of the parameter vectors, $g(x) = \lambda\|x\|_1$, and indicator functions that are zero if x lies in a simple convex set and are infinity otherwise. In order to analyze proximal-gradient algorithms, a natural (though not particularly intuitive) generalization of the PL inequality is that there exists a $\mu > 0$ satisfying

$$\frac{1}{2}\mathcal{D}_g(x, L) \geq \mu(F(x) - F^*), \tag{12}$$

where

$$\mathcal{D}_g(x, \alpha) \equiv -2\alpha \min_y \left[\langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}||y - x||^2 + g(y) - g(x)\right]. \tag{13}$$

We call this the *proximal-PL* inequality, and we note that if g is constant (or linear) then it reduces to the standard PL inequality. Below we show that this inequality is sufficient for the proximal-gradient method to achieve a global linear convergence rate.

Theorem 5. *Consider problem* (11)*, where f has an L-Lipschitz continuous gradient (2), F has a non-empty solution set $\mathcal{X}^*$, g is convex, and F satisfies the proximal-PL inequality* (12)*. Then the proximal-gradient method with a step-size of $1/L$,*

$$x_{k+1} = \operatorname*{argmin}_y \left[\langle \nabla f(x_k), y - x_k\rangle + \frac{L}{2}||y - x_k||^2 + g(y) - g(x_k)\right] \tag{14}$$

converges linearly to the optimal value F^*,

$$F(x_k) - F^* \leq \left(1 - \frac{\mu}{L}\right)^k [F(x_0) - F^*].$$

Proof. By using Lipschitz continuity of the gradient of f we have

$$\begin{aligned}
F(x_{k+1}) &= f(x_{k+1}) + g(x_k) + g(x_{k+1}) - g(x_k) \\
&\leq F(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}||x_{k+1} - x_k||^2 + g(x_{k+1}) - g(x_k) \\
&\leq F(x_k) - \frac{1}{2L}\mathcal{D}_g(x_k, L) \\
&\leq F(x_k) - \frac{\mu}{L}[F(x_k) - F^*],
\end{aligned}$$

which uses the definition of x_{k+1} and $\mathcal{D}_g$ followed by the proximal-PL inequality (12). This subsequently implies that

$$F(x_{k+1}) - F^* \leq \left(1 - \frac{\mu}{L}\right)[F(x_k) - F^*], \tag{15}$$

which applied recursively gives the result. □

While other conditions have been proposed to show linear convergence rates of proximal-gradient methods without SC [Kadkhodaie et al., 2014, Bolte et al., 2015, Zhang, 2015, Li and Pong, 2016], their analyses tend to be more complicated than the above. Further, in Appendix G we show that the proximal-PL condition is in fact equivalent to the KL condition, which itself is known to be equivalent to a proximal-gradient variant on the EB condition [Bolte et al., 2015]. Thus, the proximal-PL inequality includes the standard scenarios where existing conditions apply.

4.1 Relevant Problems

As with the PL inequality, we now list several important function classes that satisfy the proximal-PL inequality (12). We give proofs that these classes satisfy the inequality in Appendix F and G.

1. The inequality is satisfied if f satisfies the PL inequality and g is constant. Thus, the above result generalizes Theorem 1.
2. The inequality is satisfied if f is SC. This is the usual assumption used to show a linear convergence rate for the proximal-gradient algorithm [Schmidt et al., 2011], although we note that the above analysis is much simpler than standard arguments.
3. The inequality is satisfied if f has the form $f(x) = h(Ax)$ for a SC function h and a matrix A, while g is an indicator function for a polyhedral set.
4. The inequality is satisfied if F is convex and satisfies the QG property.
5. The inequality is satisfied if F satisfies the proximal-EB condition or the KL inequality.

By the equivalence shown in Appendix G, the proximal-PL inequality also holds for other problems where a linear convergence rate has been show like group L1-regularization [Tseng, 2010], sparse group L1-regularization [Zhang et al., 2013], nuclear-norm regularization [Hou et al., 2013], and other classes of functions [Zhou and So, 2015, Drusvyatskiy and Lewis, 2016].

4.2 Least Squares with L1-Regularization

Perhaps the most interesting example of problem (11) is the ℓ_1-regularized least squares problem,

$$\underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1,$$

where $\lambda > 0$ is the regularization parameter. This problem has been studied extensively in machine learning, signal processing, and statistics. This problem structure seems well-suited to using proximal-gradient methods, but the first works analyzing proximal-gradient methods for this problem only showed sub-linear convergence rates [Beck and Teboulle, 2009]. Subsequent works show that linear convergence rates can be achieved under additional assumptions. For example, Gu et al. [2013] prove that their algorithm achieves a linear convergence rate if A satisfies a *restricted isometry property* (RIP) and the solution is sufficiently sparse. Xiao and Zhang [2013] also assume the RIP property and show linear convergence using a homotopy method that slowly decreases the value of λ. Agarwal et al. [2012] give a

linear convergence rate under a *modified restricted strong convexity* and *modified restricted smoothness* assumption. But these problems have also been shown to satisfy proximal variants of the KL and EB conditions [Tseng, 2010, Bolte et al., 2015, Necoara and Clipici, 2016], and Bolte et al. [2015] in particular analyzes the proximal-gradient method under KL while giving explicit bounds on the constant. This means *any* L1-regularized least squares problem also satisfies the proximal-PL inequality. Thus, Theorem 5 gives a simple proof of global linear convergence for these problems without making additional assumptions or making any modifications to the algorithm.

4.3 Proximal Coordinate Descent

It is also possible to adapt our results on coordinate descent and proximal-gradient methods in order to give a linear convergence rate for coordinate-wise proximal-gradient methods for problem (11). To do this, we require the extra assumption that g is a separable function. This means that $g(x) = \sum_i g_i(x_i)$ for a set of univariate functions g_i. The update rule for the coordinate-wise proximal-gradient method is

$$x_{k+1} = \underset{\alpha}{\operatorname{argmin}} \left[\alpha \nabla_{i_k} f(x_k) + \frac{L}{2}\alpha^2 + g_{i_k}(x_{i_k} + \alpha) - g_{i_k}(x_{i_k})\right], \tag{16}$$

We state the convergence rate result below.

Theorem 6. *Assume the setup of Theorem 5 and that g is a separable function $g(x) = \sum_i g_i(x_i)$, where each g_i is convex. Then the coordinate-wise proximal-gradient update rule (16) achieves a convergence rate*

$$\mathbb{E}[F(x_k) - F^*] \leq \left(1 - \frac{\mu}{dL}\right)^k [F(x_0) - F^*], \tag{17}$$

when i_k is selected uniformly at random.

The proof is given in Appendix H and although it is more complicated than the proofs of Theorems 4 and 5, it is arguably still simpler than existing proofs for proximal coordinate descent under SC [Richtárik and Takáč, 2014], KL [Attouch et al., 2013], or QG [Zhang, 2016]. It is also possible to analyze stochastic proximal-gradient algorithms, and indeed Reddi et al. [2016c] use the proximal-PL inequality to analyze finite-sum methods in the proximal stochastic case.

4.4 Support Vector Machines

Another important model problem that arises in machine learning is support vector machines,

$$\underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \frac{\lambda}{2} x^T x + \sum_{i=1}^{n} \max(0, 1 - b_i x^T a_i). \tag{18}$$

where (a_i, b_i) are the labelled training set with $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, 1\}$. We often solve this problem by performing coordinate optimization on its Fenchel dual, which has the form

$$\min_{\bar{w}} f(\bar{w}) = \frac{1}{2}\bar{w}^T M \bar{w} - \sum \bar{w}_i, \quad \bar{w}_i \in [0, U], \tag{19}$$

for a particular positive semi-definite matrix M and constant U. This convex function satisfies the QG property and thus Theorem 6 implies that coordinate optimization achieves a linear convergence rate in terms of optimizing the dual objective. Further, note that Hush et al. [2006] show that we can obtain an ϵ-accurate solution to the primal problem with an $O(\epsilon^2)$-accurate solution to the dual problem. Thus this result also implies we can obtain a linear convergence rate on the primal problem by showing that stochastic dual coordinate ascent has a linear convergence rate on the dual problem. Global linear convergence rates for SVMs have also been shown by others [Tseng and Yun, 2009, Wang and Lin, 2014, Ma et al., 2015], but again we note that these works lead to more complicated analyses. Although the constants in these convergence rate may be quite bad (depending on the smallest non-zero singular value of the Gram matrix), we note that the existing sublinear rates still apply in the early iterations while, as the algorithm begins to identify support vectors, the constants improve (depending on the smallest non-zero singular value of the block of the Gram matrix corresponding to the support vectors).

The result of the previous section is not only restricted to SVMs. Indeed, the result of the previous subsection implies a linear convergence rate for many ℓ_2-regularized linear prediction problems, the framework considered in the stochastic dual coordinate ascent (SDCA) work of Shalev-Shwartz and Zhang [2013]. While Shalev-Shwartz and Zhang [2013] show that this is true when the primal is smooth, our result gives linear rates in many cases where the primal is non-smooth.

5 Discussion

We believe that this work provides a unifying and simplifying view of a variety of optimization and convergence rate issues in machine learning. Indeed, we have shown that many of the assumptions used to achieve linear convergence rates can be replaced by the PL inequality and its proximal generalization. While we have focused on sufficient conditions for linear convergence, another recent work has turned to the question of necessary conditions for convergence [Zhang, 2016]. Further, while we've focused on non-accelerated methods, Zhang [2016] has recently analyzed Nesterov's accelerated gradient method without strong convexity. We also note that, while we have focused on first-order methods, Nesterov and Polyak [2006] have used the PL inequality to analyze a second-order Newton-style method with cubic regularization. They also consider a generalization of the inequality under the name *gradient-dominated* functions.

Throughout the paper, we have pointed out how our analyses imply convergence rates for a variety of machine learning models and algorithms. Some of these were previously known, typically under stronger assumptions or with more complicated proofs, but many of these are novel. Note that we have not provided any experimental results in this work, since the main contributions of this work are showing that existing algorithms actually work better on standard problems than we previously thought. We expect that going forward efficiency will no longer be decided by the issue of whether functions are SC, but rather by whether they satisfy a variant of the PL inequality.

Acknowledgments. We would like to thank Simon LaCoste-Julien, Martin Takáč, Ruoyu Sun, Hui Zhang, and Dmitriy Drusvyatskiy for valuable discussions. We would like to thank Ting Kei Pong and Zirui Zhou for pointing out an error in the first version of this paper, to Ting Kei Pong for discussions that lead to the addition of Appendix G, to Jérôme Bolte for an informative discussion about the KL inequality and pointing us to related results that we had missed, to Liam Madden and Stephen Becker for pointing out an error (and the fix) in our "PL implies QG" proof, and to Boris Polyak for providing an English translation of his original work. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-06068-2015). Julie Nutini is funded by a UBC Four Year Doctoral Fellowship (4YF) and Hamed Karimi is support by a Mathematics of Information Technology and Complex Systems (MITACS) Elevate Fellowship.

Appendix A Relationships Between Conditions

We start by stating the different conditions. All of these definitions involve some constant $\mu > 0$ (which may not be the same across conditions), and we'll use the convention that x_p is the projection of x onto the solution set $\mathcal{X}^*$.

1. **Strong Convexity** (SC): For all x and y we have

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

2. **Essential Strong Convexity** (ESC): For all x and y such that $x_p = y_p$ we have

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

3. **Weak Strong Convexity** (WSC): For all x we have

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

4. **Restricted Secant Inequality** (RSI): For all x we have

$$\langle \nabla f(x), x - x_p \rangle \geq \mu \|x_p - x\|^2.$$

   If the function f is also convex it is called **restricted strong convexity** (RSC).
5. **Error Bound** (EB): For all x we have

$$\|\nabla f(x)\| \geq \mu \|x_p - x\|.$$

6. **Polyak-Łojasiewicz** (PL): For all x we have

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*).$$

7. **Quadratic Growth** (QG): For all x we have

$$f(x) - f^* \geq \frac{\mu}{2}\|x_p - x\|^2.$$

   If the function f is also convex it is called **optimal strong convexity** (OSC) or **semi-strong convexity** or sometimes WSC (but we'll reserve the expression WSC for the definition above).

Below we prove a subset of the implications in Theorem 2. The remaining relationships in Theorem 2 follow from these results and transitivity.

- **SC** $\rightarrow$ **ESC**: The SC assumption implies that the ESC inequality is satisfied for all x and y, so it is also satisfied under the constraint $x_p = y_p$.
- **ESC** $\rightarrow$ **WSC**: Take $y = x_p$ in the ESC inequality (which clearly has the same projection as x) to get WSC with the same μ as a special case.
- **WSC** $\rightarrow$ **RSI**: Re-arrange the WSC inequality to

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

  Since $f(x) - f^* \geq 0$, we have RSI with $\frac{\mu}{2}$.
- **RSI** $\rightarrow$ **EB**: Using Cauchy-Schwartz on the RSI we have

$$\|\nabla f(x)\|\|x - x_p\| \geq \langle \nabla f(x), x - x_p \rangle \geq \mu \|x_p - x\|^2,$$

  and dividing both sides by $\|x - x_p\|$ (assuming $x \neq x_p$) gives EB with the same μ (while EB clearly holds if $x = x_p$).
- **EB** $\rightarrow$ **PL**: By Lipschitz continuity we have

$$f(x) \leq f(x_p) + \langle \nabla f(x_p), x - x_p \rangle + \frac{L}{2}\|x_p - x\|^2,$$

  and using EB along with $f(x_p) = f^*$ and $\nabla f(x_p) = 0$ we have

$$f(x) - f^* \leq \frac{L}{2}\|x_p - x\|^2 \leq \frac{L}{2\mu}\|\nabla f(x)\|^2,$$

  which is the PL inequality with constant $\frac{\mu}{L}$.
- **PL** $\rightarrow$ **EB**: Below we show that PL implies QG with the same constant. Using this result in PL we get

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*) \geq \frac{\mu^2}{2}\|x - x_p\|^2,$$

  which implies that EB holds with the same constant.
- **QG + Convex** $\rightarrow$ **RSI**: By convexity we have

$$f(x_p) \geq f(x) + \langle \nabla f(x), x_p - x \rangle.$$

  Re-arranging and using QG we get

$$\langle \nabla f(x), x - x_p \rangle \geq f(x) - f^* \geq \frac{\mu}{2}\|x_p - x\|^2,$$

  which is RSI with constant $\frac{\mu}{2}$.

– **PL $\rightarrow$ QG**: Our argument that this implication holds is similar to the argument used in related works [Bolte et al., 2015, Zhang, 2015] Define the function

$$g(x) = \sqrt{f(x) - f^*}.$$

If we assume that f satisfies the PL inequality then for any $x \notin \mathcal{X}^*$ we have

$$\|\nabla g(x)\|^2 = \left\| \frac{1}{2\sqrt{f(x) - f^*}} \nabla f(x) \right\|^2 = \frac{\|\nabla f(x)\|^2}{4(f(x) - f^*)} \geq \frac{\mu}{2},$$

or that

$$\|\nabla g(x)\| \geq \sqrt{\frac{\mu}{2}}. \tag{20}$$

By the definition of g, to show QG it is sufficient to show that

$$g(x) \geq \sqrt{\frac{\mu}{2}} \|x - x_p\|. \tag{21}$$

As f is assumed to satisfy the PL inequality we have that f is an invex function and thus by definition g is a positive invex function ($g(x) \geq 0$) with a closed optimal solution set $\mathcal{X}^*$ such that for all $y \in \mathcal{X}^*$, $g(y) = 0$. For any point $x_0 \notin \mathcal{X}^*$, consider solving the following differential equation:

$$\begin{aligned} \frac{dx(t)}{dt} &= -\nabla g(x(t)) \\ x(t = 0) &= x_0, \end{aligned} \tag{22}$$

for $x(t) \notin \mathcal{X}^*$. (This is a flow orbit starting at x_0 and flowing along the gradient of g.) By (20), ∇g is bounded from below, and as g is a positive invex function g is also bounded from below. Thus, by moving along the path defined by (22) we are sufficiently reducing the function and will eventually reach the optimal set. Thus there exists a T such that $x(T) \in \mathcal{X}^*$ (and at this point the differential equation ceases to be defined). We can show this by using the steps

$$\begin{aligned} g(x_0) - g(x_t) &= \int_{x_t}^{x_0} \langle \nabla g(x), dx \rangle && \text{(gradient theorem for line integrals)} \\ &= -\int_{x_0}^{x_t} \langle \nabla g(x), dx \rangle && \text{(flipping integral bounds)} \\ &= -\int_0^T \langle \nabla g(x(t)), \frac{dx(t)}{dt} \rangle \, dt && \text{(reparameterization)} \\ (*) \qquad &= \int_0^T \|\nabla g(x(t))\|^2 \, dt && \text{(from (22))} \\ &\geq \int_0^T \frac{\mu}{2} dt && \text{(from (20))} \\ &= \frac{\mu}{2} T. \end{aligned}$$

As $g(x_t) \geq 0$, this shows we need to have $T \leq 2g(x_0)/\mu$, so there must be a T with $x(T) \in \mathcal{X}^*$. The *length* of the orbit $x(t)$ starting at x_0, which we'll denote by $\mathcal{L}(x_0)$, is given by

$$\mathcal{L}(x_0) = \int_0^T \|dx(t)/dt\| dt = \int_0^T \|\nabla g(x(t))\| \, dt \geq \|x_0 - x_p\|, \tag{23}$$

where x_p is the projection of x_0 onto $\mathcal{X}^*$ and the inequality follows because the orbit is a path from x_0 to a point in $\mathcal{X}^*$ (and thus it must be at least as long as the projection distance).

Starting from the line marked $(*)$ above we have

$$\begin{aligned} g(x_0) - g(x_T) &= \int_0^T \|\nabla g(x(t))\|^2 \, dt \\ &\geq \sqrt{\frac{\mu}{2}} \int_0^T \|\nabla g(x(t))\| \, dt \qquad \text{(by the PL inequality variation in (20))} \\ &\geq \sqrt{\frac{\mu}{2}} \|x_0 - x_p\|. \qquad \text{(by (23))} \end{aligned}$$

As $g(x_T) = 0$, this yields our result (21), or equivalently

$$f(x) - f^* \geq \frac{\mu}{2}\|x - x_p\|^2,$$

which is QG with the same constant.

Appendix B Relevant Problems

Strongly-convex:
By minimizing both sides of the SC inequality with respect to y we get

$$f(x^*) \geq f(x) - \frac{1}{2\mu}||\nabla f(x)||^2,$$

which implies the PL inequality holds with the same value μ. Thus, Theorem 1 exactly matches the known rate for gradient descent with a step-size of $1/L$ for a μ-SC function.

Strongly-convex composed with linear:
To show that this class of functions satisfies the PL inequality, we first define $f(x) := g(Ax)$ for a σ-strongly convex function g. For arbitrary x and y, we define $u := Ax$ and $v := Ay$. By the strong convexity of g, we have

$$g(v) \geq g(u) + \nabla g(u)^T(v - u) + \frac{\sigma}{2}\|v - u\|^2.$$

By our definitions of u and v, we get

$$g(Ay) \geq g(Ax) + \nabla g(Ax)^T(Ay - Ax) + \frac{\sigma}{2}\|Ay - Ax\|^2,$$

where we can write the middle term as $(A^T\nabla g(Ax))^T(y - x)$. By the definition of f and its gradient being $\nabla f(x) = A^T\nabla g(Ax)$ by the multivariate chain rule, we obtain

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

Using x_p to denote the projection of x onto the optimal solution set $\mathcal{X}^*$, we have

$$\begin{aligned} f(x_p) &\geq f(x) + \langle \nabla f(x), x_p - x\rangle + \frac{\sigma}{2}||A(x_p - x)||^2 \\ &\geq f(x) + \langle \nabla f(x), x_p - x\rangle + \frac{\sigma\theta(A)}{2}||x_p - x||^2 \\ &\geq f(x) + \min_y \left[\langle \nabla f(x), y - x\rangle + \frac{\sigma\theta(A)}{2}||y - x||^2\right] \\ &= f(x) - \frac{1}{2\theta(A)\sigma}||\nabla f(x)||^2. \end{aligned}$$

In the second line we use that $\mathcal{X}^*$ is polyhedral, and use the theorem of Hoffman [1952] to obtain a bound in terms of $\theta(A)$ (the smallest non-zero singular value of A). This derivation implies that the PL inequality is satisfied with $\mu = \sigma\theta(A)$.

Appendix C Sign-Based Gradient Methods

The learning heuristic RPROP (Resilient backPROPagation) is a classic iterative method used for supervised learning problems in feedforward neural networks [Riedmiller and Braun, 1992]. The general update for some vector of step sizes $\alpha_k \in \mathbb{R}^d$ is given by

$$x^{k+1} = x^k - \alpha^k \circ \text{sign}(\nabla f(x^k)),$$

where the $\circ$ operator indicates coordinate-wise multiplication. Although this method has been used for many years in the machine learning community, we are not aware of any previous convergence rate analysis of such a method. Here we give a convergence rate when the individual step-sizes α_i^k are chosen proportional to $1/\sqrt{L_i}$, where the L_i are constants such that the gradient is 1-Lipschitz continuous in the norm defined by

$$\|z\|_{L^{-1}[1]} \triangleq \sum_i \frac{1}{\sqrt{L_i}}|z_i|.$$

Formally, we assume that the L_i are set so that for all x and y we have

$$\|\nabla f(y) - \nabla f(x)\|_{L^{-1}[1]} \leq \|y - x\|_{L[\infty]},$$

and where the dual norm of the $\|\cdot\|_{L^{-1}[1]}$ norm above is given by the $\|\cdot\|_{L[\infty]}$ norm,

$$\|z\|_{L[\infty]} \triangleq \max_i \sqrt{L_i}|z_i|.$$

We note that such L_i always exist if the gradient is Lipschitz continuous, so this is not adding any assumptions on the function f. The particular choice of the step-sizes α_i^k that we will analyze is

$$\alpha_i^k = \frac{\|\nabla f(x^k)\|_{L^{-1}[1]}}{\sqrt{L_i}},$$

which yields a linear convergence rate for problems where the PL inequality is satisfied.

The coordinate-wise iteration update under this choice of α_i^k is given by

$$x_i^{k+1} = x_i^k - \frac{\|\nabla f(x^k)\|_{L^{-1}[1]}}{\sqrt{L_i}}\text{sign}(\nabla_i f(x^k)).$$

Defining a diagonal matrix Λ with $1/\sqrt{L_i}$ along the diagonal, the update can be written as

$$x^{k+1} = x^k - \|\nabla f(x^k)\|_{L^{-1}[1]}\Lambda \circ \text{sign}(\nabla f(x^k)).$$

Consider the function $g(\tau) = f(x + \tau(y - x))$ with $\tau \in \mathbb{R}$. Then

$$\begin{aligned}
f(y) - f(x) - \langle \nabla f(x), y - x\rangle &= g(1) - g(0) - \langle \nabla f(x), y - x\rangle \\
&= \int_0^1 \frac{dg}{d\tau}(\tau) - \langle \nabla f(x), y - x\rangle \; d\tau \\
&= \int_0^1 \langle \nabla f(x + \tau(y - x)), y - x\rangle - \langle \nabla f(x), y - x\rangle \; d\tau \\
&= \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x\rangle \; d\tau \\
&\leq \int_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_{L^{-1}[1]}\|y - x\|_{L[\infty]} \; d\tau \\
&\leq \int_0^1 \tau\|y - x\|_{L[\infty]}^2 \; d\tau \\
&= \tau^2 \frac{1}{2}\|y - x\|_{L[\infty]}^2\Big|_0^1 \\
&= \frac{1}{2}\|y - x\|_{L[\infty]}^2 \\
&= \frac{1}{2}|y - x\|_{L[\infty]}^2.
\end{aligned}$$

where the second inequality uses the Lipschitz assumption, and in the first inequality we've used the Cauchy-Schwarz inequality and that the dual norm of the $L^{-1}[1]$ norm is the $L[\infty]$ norm. The above gives an upper bound on the function in terms of this $L[\infty]$-norm,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\|y - x\|^2_{L[\infty]}.$$

Plugging in our iteration update we have

$$\begin{aligned}
&f(x^{k+1}) \\
&\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2}\|x^{k+1} - x^k\|^2_{L[\infty]} \\
&= f(x^k) - \|\nabla f(x^k)\|_{L^{-1}[1]} \langle \nabla f(x^k), \Lambda \circ \text{sign}(\nabla f(x^k)) \rangle + \frac{\|\nabla f(x^k)\|^2_{L^{-1}[1]}}{2} \|\Lambda \circ \text{sign}(\nabla f(x^k))\|^2_{L[\infty]} \\
&= f(x^k) - \|\nabla f(x^k)\|^2_{L^{-1}[1]} + \frac{\|\nabla f(x^k)\|^2_{L^{-1}[1]}}{2} \left(\max_i \frac{1}{\sqrt{L_i}} \sqrt{L_i} |\text{sign}(\nabla_i f(x^k))| \right)^2 \\
&= f(x^k) - \frac{1}{2}\|\nabla f(x^k)\|^2_{L^{-1}[1]}.
\end{aligned}$$

Subtracting f^* from both sides yields

$$f(x^{k+1}) - f(x^*) \leq f(x^k) - f(x^*) - \frac{1}{2}\|\nabla f(x^k)\|^2_{L^{-1}[1]}.$$

Applying the PL inequality with respect to the $L^{-1}[1]$-norm (which, if the PL inequality is satisfied, holds for some $\mu_{L[\infty]}$ by the equivalence between norms),

$$\frac{1}{2}\|\nabla f(x^k)\|^2_{L^{-1}[1]} \geq \mu_{L[\infty]} \left(f(x^k) - f^* \right),$$

we have

$$f(x^{k+1}) - f(x^*) \leq \left(1 - \mu_{L[\infty]}\right) \left(f(x^k) - f(x^*) \right).$$

Appendix D Linear Convergence Rate of SVRG Method

In this section, we look at the SVRG method for the finite-sum optimization problem,

$$f(w) = \frac{1}{n} \sum_i f_i(w). \tag{24}$$

To minimize functions of this form, the SVRG algorithm of Johnson and Zhang [2013] uses iterations of the form

$$x_t = x_{t-1} - \alpha[\nabla f_{i_t}(x_{t-1}) - f_{i_t}(x^s) + \mu^s], \tag{25}$$

where i_t is chosen uniformly from $\{1, 2, \ldots, n\}$ and we assume the step-size satisfies $\alpha < 2/L$. In this algorithm we start with some x^0 and initially set $\mu^0 = \nabla f(x^0)$ and $x_0 = x^0$, but after every m steps we set x^{s+1} to a random x_t for $t \in \{ms + 1, \ldots, m(s + 1)\}$, then replace μ^s with $\nabla f(x^s)$ and x^t with x^{s+1}. Analogous to Johnson and Zhang [2013] for the SC case, we now show tnat SVRG has a linear convergence rate if each f_i is a convex function with a Lipschitz-continuous gradient and f satisfies the PL inequality.

Following the same argument as Johnson and Zhang [2013], for any solution x^* the assumptions on the f_i mean that the "outer" SVRG iterations x^s satisfy

$$2\alpha(1 - 2L\alpha)m\mathbb{E}[f(x^s) - f^*] \leq \mathbb{E}[\|x^{s-1} - x^*\|^2] + 4L\alpha^2 m\mathbb{E}[f(x^{s-1}) - f^*].$$

Choosing the particular x^* that is the projection of x^{s-1} onto the solution set and using QG (which is equivalent to PL in this convex setting) we have

$$2\alpha(1 - 2L\alpha)m\mathbb{E}[f(x^s) - f^*] \leq \frac{2}{\mu}\mathbb{E}[f(x^{s-1}) - f^*] + 4L\alpha^2 m\mathbb{E}[f(x^{s-1}) - f^*].$$

Dividing both sides by $2\alpha(1-2L\alpha)m$ we get

$$\mathbb{E}[f(x^s) - f^*] \leq \frac{1}{1-2\alpha L}\left(\frac{1}{m\mu\alpha} + 2L\alpha\right)\mathbb{E}[f(x^{s-1}) - f^*],$$

which is a linear convergence rate for sufficiently large m and sufficiently small α.

Appendix E Proximal-PL Lemma

In this section we give a useful property of the function $\mathcal{D}_g$.

Lemma 1. *For any differentiable function f and any convex function g, given $\mu_2 \geq \mu_1 > 0$ we have*

$$\mathcal{D}_g(x, \mu_2) \geq \mathcal{D}_g(x, \mu_1).$$

We'll prove Lemma 1 as a corollary of a related result. We first restate the definition

$$\mathcal{D}_g(x, \lambda) = -2\lambda \min_y \left[\langle \nabla f(x), y - x\rangle + \frac{\lambda}{2}||y-x||^2 + g(y) - g(x)\right], \tag{26}$$

and we note that we require $\lambda > 0$. By completing the square, we have

$$\begin{aligned}\mathcal{D}_g(x,\lambda) &= -\min_y \left[-\|\nabla f(x)\|^2 + \|\nabla f(x)\|^2 + 2\lambda\langle \nabla f(x), y-x\rangle + \lambda^2||y-x||^2 + 2\lambda(g(y)-g(x))\right] \\ &= ||\nabla f(x)||^2 - \min_y \left[||\lambda(y-x) + \nabla f(x)||^2 + 2\lambda(g(y)-g(x))\right].\end{aligned}$$

Notice that if $g = 0$, then $\mathcal{D}_g(x,\lambda) = ||\nabla f(x)||^2$ and the proximal-PL inequality reduces to the PL inequality. We'll the define the *proximal residual* function as the second part of the above equality,

$$\mathcal{R}_g(\lambda, x, a) \triangleq \min_y \left[||\lambda(y-x) + a||^2 + 2\lambda(g(y) - g(x)\right]. \tag{27}$$

Lemma 2. *If g is convex then for any x and a, and for $0 < \lambda_1 \leq \lambda_2$ we have*

$$\mathcal{R}_g(\lambda_1, x, a) \geq \mathcal{R}_g(\lambda_2, x, a). \tag{28}$$

Proof. Without loss of generality, assume $x = 0$. Then we have

$$\begin{aligned}\mathcal{R}_g(\lambda, a) &= \min_y \left[||\lambda y + a||^2 + 2\lambda(g(y) - g(0)\right] \\ &= \min_{\bar{y}} \left[||\bar{y} + a||^2 + 2\lambda(g(\bar{y}/\lambda) - g(0)\right],\end{aligned} \tag{29}$$

where in the second line we used a changed of variables $\bar{y} = \lambda y$ (note that we are minimizing over the whole space of $\mathbb{R}^n$). By the convexity of g, for any $\alpha \in [0,1]$ and $z \in \mathbb{R}^n$ we have

$$\begin{aligned} & g(\alpha z) \leq \alpha g(z) + (1-\alpha)g(0) \\ \iff \quad & g(\alpha z) - g(0) \leq \alpha(g(z) - g(0)).\end{aligned} \tag{30}$$

By using $0 < \lambda_1/\lambda_2 \leq 1$ and using the choices $\alpha = \frac{\lambda_1}{\lambda_2}$ and $z = \bar{y}/\lambda_1$ we have

$$\begin{aligned} & g(\bar{y}/\lambda_2) - g(0) \leq \frac{\lambda_1}{\lambda_2}(g(\bar{y}/\lambda_1) - g(0)) \\ \iff \quad & \lambda_2(g(\bar{y}/\lambda_2) - g(0)) \leq \lambda_1(g(\bar{y}/\lambda_1) - g(0)),\end{aligned} \tag{31}$$

Adding $||\bar{y} + a||^2$ to both sides, we get

$$||\bar{y} + a||^2 + \lambda_2(g(\bar{y}/\lambda_2) - g(0)) \leq ||\bar{y} + a||^2 + \lambda_1(g(\bar{y}/\lambda_1) - g(0)). \tag{32}$$

Taking the minimum over both sides with respect to $\bar{y}$ yields Lemma 2 due to (29). □

Corollary 1. *For any differentiable function f and convex function g, given $\lambda_1 \leq \lambda_2$, we have*

$$\mathcal{D}_g(x, \lambda_2) \geq \mathcal{D}_g(x, \lambda_1). \tag{33}$$

By using $\mathcal{D}_g(x,\lambda) = ||\nabla f(x)||^2 - \mathcal{R}_g(\lambda, x, \nabla f(x))$, Corollary 1 is exactly Lemma 1.

Appendix F Relevant Problems

In this section we prove that the three classes of functions listed in Section 4.1 satisfy the proximal-PL inequality condition. Note that while we prove these hold for $\mathcal{D}_g(x,\lambda)$ for $\lambda \leq L$, by Lemma 1 above they also hold for $\mathcal{D}_g(x,L)$.

1. *$f(x)$, where f satisfies the PL inequality (g is constant)*:
   As g is assumed to be constant, we have $g(y) - g(x) = 0$ and the left-hand side of the proximal-PL inequality simplifies to

$$\begin{aligned}\mathcal{D}_g(x,\mu) &= -2\mu \min_y \left\{ \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y-x\|^2 \right\} \\ &= -2\mu \left(-\frac{1}{2\mu}\|f(x)\|^2 \right) \\ &= \|\nabla f(x)\|^2,\end{aligned}$$

   Thus, the proximal PL inequality simplifies to f satisfying the PL inequality,

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\left(f(x) - f^*\right),$$

   as we assumed.
2. *$F(x) = f(x) + g(x)$ and f is strongly convex*:
   By the strong convexity of f we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}||y-x||^2, \tag{34}$$

   which leads to

$$F(y) \geq F(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}||y-x||^2 + g(y) - g(x). \tag{35}$$

   Minimizing both sides respect to y,

$$\begin{aligned}F^* &\geq F(x) + \min_y \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}||y-x||^2 + g(y) - g(x) \\ &= F(x) - \frac{1}{2\mu}\mathcal{D}_g(x,\mu).\end{aligned} \tag{36}$$

   Rearranging, we have our result.
3. *$F(x) = f(Ax) + g(x)$ and f is strongly convex, g is the indicator function for a polyhedral set $\mathcal{X}$, and A is a linear transformation*:
   By defining $\tilde{f}(x) = f(Ax)$ and using strong convexity of f, we have

$$\tilde{f}(y) \geq \tilde{f}(x) + \langle \nabla \tilde{f}(x), y - x\rangle + \frac{\mu}{2}||A(y-x)||^2, \tag{37}$$

   which leads to

$$F(y) \geq F(x) + \langle \nabla \tilde{f}(x), y - x\rangle + \frac{\mu}{2}||A(y-x)||^2 + g(y) - g(x). \tag{38}$$

   Since $\mathcal{X}$ is polyhedral, it can be written as a set $\{x : Bx \leq c\}$ for a matrix B and a vector c. As before, assume that x_p is the projection of x onto the optimal solution set $\mathcal{X}^*$ which in this case is

$\{x : Bx \leq c, Ax = z\}$ for some z.

$$\begin{aligned}
F^* = F(x_p) &\geq F(x) + \langle \nabla \tilde{f}(x), x_p - x \rangle + \frac{\mu}{2}||A(x - x_p)||^2 + g(x_p) - g(x) \\
&= F(x) + \langle \nabla \tilde{f}(x), x_p - x \rangle + \frac{\mu}{2}||Ax - z||^2 + g(x_p) - g(x) \\
&= F(x) + \langle \nabla \tilde{f}(x), x_p - x \rangle + \frac{\mu}{2}||\{Ax - z\}_+ + \{-Ax + z\}_+||^2 + g(x_p) - g(x) \\
&= F(x) + \langle \nabla \tilde{f}(x), x_p - x \rangle + \frac{\mu}{2}\left\| \left\{ \begin{bmatrix} A \\ -A \\ B \end{bmatrix} x - \begin{bmatrix} z \\ -z \\ c \end{bmatrix} \right\}_+ \right\|^2 + g(x_p) - g(x) \\
&\geq F(x) + \langle \nabla \tilde{f}(x), x_p - x \rangle + \frac{\mu\theta(A,B)}{2}||x - x_p||^2 + g(x_p) - g(x) \\
&\geq F(x) + \min_y \left[\langle \nabla \tilde{f}(x), y - x \rangle + \frac{\mu\theta(A,B)}{2}||y - x||^2 + g(y) - g(x) \right] \\
&= F(x) - \frac{1}{2\mu\ \theta(A)}\mathcal{D}_g(x, \mu\theta(A,B)).
\end{aligned} \tag{39}$$

where we've used the notation that $\{\cdot\}_+ = \max\{0, \cdot\}$, the fourth equality follows because x was projected onto $\mathcal{X}$ in the previous iteration (so $Bx - c \leq 0$), and the line after that uses Hoffman's bound [Hoffman, 1952].

4. *$F(x) = f(x) + g(x)$, f is convex, and F satisfies the quadratic growth (QG) condition*:
A function F satisfies the QG condition if

$$F(x) - F^* \geq \frac{\mu}{2}||x - x_p||^2. \tag{40}$$

For any $\lambda > 0$ we have,

$$\begin{aligned}
&\min_y \left[\langle \nabla f(x), y - x \rangle + \frac{\lambda}{2}||y - x||^2 + g(y) - g(x) \right] \\
&\leq \langle \nabla f(x), x_p - x \rangle + \frac{\lambda}{2}||x_p - x||^2 + g(x_p) - g(x) \\
&\leq f(x_p) - f(x) + \frac{\lambda}{2}||x_p - x||^2 + g(x_p) - g(x) \\
&= \frac{\lambda}{2}||x_p - x||^2 + F^* - F(x) \\
&\leq \left(1 - \frac{\lambda}{\mu}\right)(F^* - F).
\end{aligned} \tag{41}$$

The third line follows from the convexity of f, and the last inequality uses the QG condition of F. Multiplying both sides by -2λ, we have

$$\mathcal{D}_g(x, \lambda) = -2\lambda \min_y \left[\langle \nabla \tilde{f}(x), y - x \rangle + \frac{\lambda}{2}||y - x||^2 + g(y) - g(x) \right] \geq 2\lambda \left(1 - \frac{\lambda}{\mu}\right)(F(x) - F^*). \tag{42}$$

This is true for any $\lambda > 0$, and by choosing $\lambda = \mu/2$ we have

$$\mathcal{D}_g(x, \mu/2) \geq \frac{\mu}{2}(F(x) - F^*). \tag{43}$$

5. *F satisfies the KL inequality or the proximal-EB inequality*:
In the next section we show that these are equivalent to the proximal-PL inequality.

Appendix G Equivalence of Proximal-PL with KL and EB

The equivalence of the KL condition and the proximal-gradient variant of the Luo-Tseng EB condition is known for convex f, see [Drusvyatskiy and Lewis, 2016, Corollary 3.6] and the proof of [Bolte et al., 2015, Theorem 5]. Here we prove the equivalence of these conditions with the proximal-PL inequality for non-convex f. First we review the definitions of the three conditions:

1. **Proximal-PL**: There exists a $\mu > 0$ such that

$$\frac{1}{2}\mathcal{D}_g(x, L) \geq \mu(F(x) - F_*)$$

where

$$\mathcal{D}_g(x, L) = -2L \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + g(y) - g(x) \right\}.$$

2. **Proximal-EB**: There exists $c > 0$ such that we have

$$\|x - x_p\| \leq c \left\| x - \text{prox}_{\frac{1}{L}g}\left(x - \frac{1}{L}\nabla f(x) \right) \right\|. \tag{44}$$

3. **Kurdyka-Łojasiewicz**: The KL condition with exponent $\frac{1}{2}$ holds if there exist $\tilde{\mu} > 0$ such that

$$\min_{s \in \partial F(x)} \|s\|^2 \geq 2\tilde{\mu}(F(x) - F_*) \tag{45}$$

where $\partial F(x)$ is the Frechet subdifferential. In particular, if $F : H \to \mathcal{R}$ is a real-valued function then we say that $s \in H$ is a Frechet subdifferential of F at $x \in \text{dom } F$ if

$$\liminf_{y \to x, y \neq x} \frac{F(y) - F(x) - \langle s, y - x \rangle}{\|y - x\|^2} \geq 0. \tag{46}$$

Note that for differentiable f the Frechet subdifferential only contains the gradient, $\nabla f(x)$. In our case where $F(x) = f(x) + g(x)$ with a differentiable f and a convex g we have

$$\partial F(x) = \{\nabla f(x) + \xi \mid \xi \in \partial g(x)\}.$$

The KL inequality is an intuitive generalization of the PL inequality since, analogous to the gradient vector in the smooth case, the negation of the quantity $\text{argmin}_{s \in \partial F(x)} \|s\|$ points in the direction of steepest descent [see Bertsekas et al., 2003, Section 8.4]

We first derive an alternative representation of $\mathcal{D}_g(x, L)$ in terms of the so-called forward-backward envelope $F_{\frac{1}{L}}$ of F [see Stella et al., 2016, Definition 2.1]. Indeed,

$$\begin{aligned}
\mathcal{D}_g(x, L) &= -2L \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + g(y) - g(x) \right\} \\
&= -2L \left[\min_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + g(y) \right\} - f(x) - g(x) \right] \\
&= -2L[F_{\frac{1}{L}}(x) - F(x)] = 2L[F(x) - F_{\frac{1}{L}}(x)],
\end{aligned} \tag{47}$$

It follows from the definition of $F_{\frac{1}{L}}(x)$ that we have

$$\begin{aligned}
F_{\frac{1}{L}}(x) - F^* &= \min_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + g(y) \right\} - f(x^*) - g(x^*) \\
&\leq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{L}{2}\|x^* - x\|^2 + g(x^*) - f(x^*) - g(x^*) \\
&= f(x) - f(x^*) + \langle \nabla f(x), x^* - x \rangle + \frac{L}{2}\|x^* - x\|^2 \\
&= f(x) - f(x^*) + \langle \nabla f(x), x^* - x \rangle + \frac{L}{2}\|x^* - x\|^2 \\
&\leq 2L\|x^* - x\|^2,
\end{aligned} \tag{48}$$

where the second line uses that we are taking the minimizer and the last line uses the Lipschitz continuity of ∇f as follows,

$$\begin{aligned}
f(x) - f(y) + \langle \nabla f(x), y - x \rangle &\leq \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|y - x\|^2 + \langle \nabla f(x), y - x \rangle \\
&= \langle \nabla f(y) - \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \\
&\leq \|\nabla f(y) - \nabla f(x)\|\|y - x\| + \frac{L}{2}\|y - x\|^2 \leq \frac{3L}{2}\|y - x\|^2.
\end{aligned} \tag{49}$$

– **Proximal-EB $\rightarrow$ proximal-PL**: we have that

$$\begin{aligned} F(x) - F^* &= F(x) - F_{\frac{1}{L}}(x) + F_{\frac{1}{L}}(x) - F^* \\ &\leq F(x) - F_{\frac{1}{L}}(x) + 2L\|x^* - x\|^2 \\ &\leq F(x) - F_{\frac{1}{L}}(x) + C_0 \left\| x - \text{prox}_{\frac{1}{L}g}\left(x - \frac{1}{L}\nabla f(x)\right)\right\|^2 \\ &\leq C_1(F(x) - F_{\frac{1}{L}}(x)), \end{aligned} \tag{50}$$

for some constants C_0 and C_1, where the second inequality uses the proximal-EB and the last inequality follows from Stella et al. [2016, Proposition 2.2(i)]. Now by using the fact that $F(x) - F_{\frac{1}{L}}(x) = \frac{1}{2L}\mathcal{D}_g(x, L)$, the function satisfies the proximal-PL inequality.

– **Proximal-PL $\rightarrow$ KL**: It's sufficient to prove that $\mathcal{D}_g(x, \mu) \leq \min_{s\in\partial F(x)} \|s\|^2$ for any x and μ. First we observe that for any subgradient $\xi \in \partial g(x)$ we have

$$\begin{aligned} \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 + g(y) - g(x) \geq \\ \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 + \langle \xi, y - x\rangle = \\ \langle \nabla f(x) + \xi, y - x\rangle + \frac{\mu}{2}\|y - x\|^2, \end{aligned} \tag{51}$$

where the inequality follows from the definition of a subgradient. Now by minimizing both sides over y we have

$$\begin{aligned} \min_y \left\{ \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 + g(y) - g(x)\right\} \geq \\ \min_y \left\{ \langle \nabla f(x) + \xi, y - x\rangle + \frac{\mu}{2}\|y - x\|^2\right\} \geq -\frac{1}{2\mu}\|\nabla f(x) + \xi\|^2. \end{aligned} \tag{52}$$

Multiplying both sides with -2μ we get

$$\mathcal{D}_g(x, \mu) \leq \|\nabla f(x) + \xi\|^2 \tag{53}$$

Since this holds for any $\xi \in \partial g(x)$, then it holds for any $\zeta = \nabla f(x) + \xi \in \partial F(x)$, it also holds for the minimium-norm subgradient of F.

– **KL $\rightarrow$ Proximal-EB**: A function $h(x)$ is called "semiconvex" if there exist an $\alpha > 0$ such that $h(x) + \alpha\|x\|^2$ is convex [see Bolte et al., 2010, Definition 10]. Note that Lipschitz-continuity of ∇f implies semi-convexity of f in light of (49) and Bolte et al. [2010, Remark 11(iii)]. It follows from convexity of g that F is semi-convex. From Bolte et al. [2010, Theorem 13], for any $x \in \text{dom}F$ there exist a subgradient curve $\chi_x : [0, \infty] \rightarrow \text{dom}F$ that satisfies

$$\begin{aligned} \dot{\chi_x}(t) &\in -\partial F(\chi_x(t)) \\ \chi_x(0) &= x \\ \frac{d}{dt}F(\chi_x(t)) &= -\|\dot{\chi_x}(t)\|^2, \end{aligned} \tag{54}$$

where $F(\chi_x(t))$ is non-increasing and Lipschitz continuous on $[\eta, \infty]$ for any $\eta > 0$. By using these facts let's define the function $r(t) = \sqrt{F(\chi_x(t)) - F^*}$. It is easy to see that

$$\begin{aligned} \frac{dr(t)}{dt} &= \frac{\dot{F}(\chi_x(t))}{2\sqrt{F(\chi_x(t)) - F^*}} \\ &= -\frac{\|\dot{\chi_x}(t)\|^2}{2\sqrt{F(\chi_x(t)) - F^*}} \\ &\leq -\sqrt{\tilde{\mu}/2}\|\dot{\chi_x}(t)\|, \end{aligned} \tag{55}$$

for the second line we used the definition of subgradient curve and for the third line we used KL inequality condition and the fact that $\dot{\chi_x}(t) \in -\partial F(\chi_x(t))$.

Now we have

$$\begin{aligned} r(T) - r(0) &= \int_0^T \frac{d}{dt} r(t) dt \\ &\leq -\sqrt{\tilde{\mu}/2} \int_0^T \|\dot{\chi_x}(t)\| dt \\ &= -\sqrt{\tilde{\mu}/2} \text{dist}(\chi_x(T), \chi_x(0)), \end{aligned} \tag{56}$$

where we used the bound on the derivative of $r(t)$ above and that the length of the curve connecting any two points is less than the Euclidean distance between them. We're now going to take the limit of $T \to \infty$, while using the facts that $r(\infty) = 0$ (which we prove below) and using $r(0) = \sqrt{F(x) - F^*}$. This gives

$$\sqrt{F(x) - F^*} \geq \sqrt{\tilde{\mu}/2} \text{dist}(x, \mathcal{X}) \tag{57}$$

From this inequality and also KL condition 45, we proved that there exist a $C > 0$ such that

$$\text{dist}(0, \partial F(x)) \geq C \text{dist}(x, \mathcal{X}). \tag{58}$$

Now let's show that $r(\infty) = 0$ or $\chi_x(\infty) \in \mathcal{X}$. From equation 56 we have

$$\begin{aligned} r(T) - r(0) &= \int_0^T \frac{d}{dt} r(t) dt \\ &= -\int_0^T \frac{\|\dot{\chi_x}(t)\|^2}{2\sqrt{F(\chi_x(t)) - F^*}} \\ &\leq -\frac{\tilde{\mu}}{2} \int_0^T \sqrt{F(\chi_x(t)) - F^*} \\ &\leq -\frac{\tilde{\mu}}{2} \int_0^T \sqrt{F(\chi_x(T)) - F^*} \\ &= -\frac{\tilde{\mu}\sqrt{F(\chi_x(T)) - F^*}}{2} T = -\frac{\tilde{\mu}\, T\, r(T)}{2} \end{aligned} \tag{59}$$

where for the first inequality we used the KL property, and for the second inequality we used the fact that $F(\chi_x(t))$ is non-increasing, which means $F(\chi_x(T)) \leq F(\chi_x(t))$ for any $t \in [0, T]$. This inequality gives a bound on $r(T)$,

$$0 \leq r(T) \leq \frac{2r(0)}{2 + \tilde{\mu} T}$$

now by taking the limit of $T \to \infty$, we get $r(T) \to 0$.

Now by using 58 we can show that the proximal-EB condition is satisfied. Let's define $\hat{x} = \text{prox}_{\frac{1}{L} g}\left(x - \frac{1}{L} \nabla f(x)\right)$. From the optimality of $\hat{x}$ we have $-\nabla f(x) - L(\hat{x} - x) \in \partial g(\hat{x})$, using this we get

$$\nabla f(\hat{x}) - \nabla f(x) - L(\hat{x} - x) \in \partial g(\hat{x}) + \nabla f(\hat{x}) = \partial F(\hat{x}).$$

Denoting the particular subgradient of g that achieves this by ξ, we have

$$\begin{aligned} \text{dist}(0, \partial F(\hat{x})) &\leq \|0 - \xi\| \\ &= \|\nabla f(\hat{x}) - \nabla f(x) - L(\hat{x} - x)\| \\ &\leq L\|(\hat{x} - x)\| + \|\nabla f(\hat{x}) - \nabla f(x)\| \\ &\leq 2L\|(\hat{x} - x)\|, \end{aligned} \tag{60}$$

where the second inequality uses $\|a - b\| \leq \|a\| + \|b\|$, and for the last line we used Lipschitz continuity of of ∇f. We finally get the proximal-EB condition using

$$\begin{aligned} \text{dist}(x, \mathcal{X}) &\leq \|x - \hat{x}\| + \text{dist}(\hat{x}, \mathcal{X}) \\ &\leq \|x - \hat{x}\| + C\text{dist}(0, \partial F(\hat{x}) \\ &\leq \|x - \hat{x}\| + 2CL\|(\hat{x} - x)\| \\ &= (1 + 2CL)\|x - \hat{x}\|, \end{aligned} \tag{61}$$

where the first inequality follows from the triangle inequality, the second line uses 58, and the third inequality uses 60.

Appendix H Proximal Coordinate Descent

Here we show linear convergence of randomized coordinate descent for $F(x) = f(x)+g(x)$ assuming that F satisfies the proximal PL inequality, ∇f is coordinate-wise Lipschitz continuous, and g is a separable convex function ($g(x) = \sum_i g_i(x_i)$).

From coordinate-wise Lipschitz continuity of ∇f and separability of g, we have

$$F(x + y_i e_i) - F(x) \leq y_i \nabla_i f(x) + \frac{L}{2} y_i^2 + g_i(x_i + y_i) - g(x_i). \tag{62}$$

Given a coordinate i the coordinate descent step chooses y_i to minimize this upper bound on the improvement in F,

$$y_i = \underset{t_i}{\operatorname{argmin}} \left\{ t_i \nabla_i f(x) + \frac{L}{2} t_i^2 + g_i(x_i + t_i) - g(x_i). \right\}$$

We next use an argument similar to Richtárik and Takáč [2014] to relate the expected improvement (with random selection of the coordinates) to the function $\mathcal{D}_g$,

$$\begin{aligned}
\mathbb{E}\left\{ \min_{t_i} t_i \nabla_i f(x) + \frac{L}{2} t_i^2 + g_i(x_i + t_i) - g_i(x_i) \right\} &= \frac{1}{n} \sum_i \min_{t_i} t_i \nabla_i f(x) + \frac{L}{2} t_i^2 + g_i(x_i + t_i) - g_i(x_i) \\
&= \frac{1}{n} \min_{t_1, \cdots, t_n} \sum_i t_i \nabla_i f(x) + \frac{L}{2} t_i^2 + g_i(x_i + t_i) - g_i(x_i) \\
&= \frac{1}{n} \min_{y \equiv x + (t_1, \cdots, t_n)} \langle \nabla f(x), y - x \rangle + \frac{L}{2} ||y - x||^2 + g(y) - g(x) \\
&= -\frac{1}{2Ln} \mathcal{D}_g(L, x).
\end{aligned}$$

(Note that separability allows us to exchange the summation and minimization operators.) By using this and taking the expectation of (62) we get

$$\mathbb{E}\left[F(x_{k+1})\right] \leq F(x_k) - \frac{1}{2Ln} \mathcal{D}_g(L, x). \tag{63}$$

Subtracting F^* from both sides and applying the proximal-PL inequality yields a linear convergence rate of $\left(1 - \frac{\mu}{nL}\right)$.

References

A. Agarwal, S. N. Negahban, and M. J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, pages 2452–2482, 2012.

M. Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM J. Optim.*, pages 1116–1135, 2000.

H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program., Ser. B*, pages 5–16, 2009.

H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. *NIPS*, pages 773–791, 2013.

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

D. Bertsekas, A. Nedic, and A. Ozdaglar. *Convex Analysis and Optimization.* Athena Scientific, 2003.

J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterization of lojasiewicz inequalities and applications. *arXiv preprint arXiv:0802.0826*, 2008.

J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *arXiv:1510.08234*, 2015.

B. D. Craven and B. M. Glover. Invex functions and duality. *J. Austral. Math. Soc. (Series A)*, pages 1–20, 1985.

F. Dinuzzo, C. S. Ong, P. Gehler, and G. Pillonetto. Learning output kernels with block coordinate descent. *ICML*, pages 49–56, 2011.

D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *arXiv preprint arXiv:1602.06661*, 2016.

D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. *ICML*, pages 541–549, 2015a.

D. Garber and E. Hazan. Faster and simple PCA via convex optimization. *arXiv:1509.05647v4*, 2015b.

P. Gong and J. Ye. Linear convergence of variance-reduced stochastic gradient without strong convexity. *arXiv:1406.1102*, 2014.

M. Gu, L.-H. Lim, and C. J. Wu. ParNes: A rapidly convergent algorithm for accurate recovery of sparse and approximately sparse signals. *Numer. Algor.*, pages 321–347, 2013.

M. A. Hanson. On sufficiency of the Kuhn-Tucker conditions. *J. Math. Anal. Appl.*, pages 545–550, 1981.

A. J. Hoffman. On approximate solutions of systems of linear inequalities. *J. Res. Nat. Bur. Stand.*, pages 263–265, 1952.

K. Hou, Z. Zhou, A. M.-C. So, and Z.-Q. Luo. On the linear convergence of the proximal gradient method for trace norm regularization. *NIPS*, pages 710–718, 2013.

D. Hush, P. Kelly, C. Scovel, and I. Steinwart. QP algorithms with guaranteed accuracy and run time for support vector machines. *J. Mach. Learn. Res.*, pages 733–769, 2006.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *NIPS*, pages 315–323, 2013.

M. Kadkhodaie, M. Sanjabi, and Z.-Q. Luo. On the linear convergence of the approximate proximal splitting method for non-smooth convex optimization. *arXiv:1404.5350v1*, 2014.

K. Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–784. Chartres: L'Institut, 1950-, 1998.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *NIPS*, pages 2672–2680, 2012.

G. Li and T. K. Pong. Calculus of the exponent of Kurdyka-Lojasiewicz inequality and its applications to linear convergence of first-order methods. *arXiv:1602.02915v1*, 2016.

J. Liu and S. J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM J. Optim.*, pages 351–376, 2015.

J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *arXiv:1311.1873v3*, 2014.

S. Łojasiewicz. A topological property of real analytic subsets (in French). *Coll. du CNRS, Les équations aux dérivées partielles*, pages 87–89, 1963.

Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Ann. Oper. Res.*, pages 157–178, 1993.

C. Ma, T. Tappenden, and M. Takáč. Linear convergence of the randomized feasible descent method under the weak strong convexity assumption. *arXiv:1506.02530*, 2015.

R. Meir and G. Rätsch. *An Introduction to Boosting and Leveraging*, pages 118–183. Springer, Heidelberg, 2003.

I. Necoara and D. Clipici. Parallel random coordinate descent method for composite minimization: Convergence analysis and error bounds. *SIAM J. Optim.*, pages 197–226, 2016.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *arXiv:1504.06298v3*, 2015.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, pages 1574–1609, 2009.

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

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, pages 341–362, 2012.

Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

J. Nutini, M. Schmidt, I. H. Laradji, M. Friedlander, and H. Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. *ICML*, pages 1632–1641, 2015.

B. T. Polyak. Gradient methods for minimizing functionals (in Russian). *Zh. Vychisl. Mat. Mat. Fiz.*, pages 643–653, 1963.

S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. *arXiv:1603.06160*, 2016a.

S. J. Reddi, S. Sra, B. Poczos, and A. Smola. Fast incremental method for nonconvex optimization. *arXiv:1603.06159*, 2016b.

S. J. Reddi, S. Sra, B. Poczos, and A. Smola. Fast stochastic methods for nonsmooth nonconvex optimization. *arXiv:1605.06900*, 2016c.

P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program., Ser. A*, pages 1–38, 2014.

M. Riedmiller and H. Braun. RPROP - A fast adaptive learning algorithm. *In: Proc. of ISCIS VII*, 1992.

M. Schmidt, N. L. Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *NIPS*, pages 1458–1466, 2011.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *J. Mach. Learn. Res.*, pages 567–599, 2013.

L. Stella, A. Themelis, and P. Patrinos. Forward-backward quasi-newton methods for nonsmooth optimization problems. *arXiv preprint arXiv:1604.08096*, 2016.

P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Math. Program., Ser. B*, pages 263–295, 2010.

P. Tseng and S. Yun. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *J. Optim. Theory Appl.*, pages 513–535, 2009.

P.-W. Wang and C.-J. Lin. Iteration complexity of feasible descent methods for convex optimization. *J. Mach. Learn. Res.*, pages 1523–1548, 2014.

L. Xiao and T. Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM J. Optim.*, pages 1062–1091, 2013.

H. Zhang. The restricted strong convexity revisited: Analysis of equivalence to error bound and quadratic growth. *arXiv:1511.01635*, 2015.

H. Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *arXiv:1606.00269v3*, 2016.

H. Zhang and W. Yin. Gradient methods for convex minimization: Better rates under weaker conditions. *arXiv:1303.4645v2*, 2013.

H. Zhang, J. Jiang, and Z.-Q. Luo. On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. *J. Oper. Res. Soc. China*, 1(2):163–186, 2013.

Z. Zhou and A. M.-C. So. A unified approach to error bounds for structured convex optimization problems. *arXiv:1512.03518*, 2015.