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, BC, 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 coordinate descent and stochastic gradient for many non-strongly-convex (and some non-convex) functions. We further propose a generalization that applies to proximal-gradient methods for non-smooth optimization, leading to simple proofs of linear convergence for support vector machines and L1-regularized least squares without additional assumptions.

Keywords: Gradient descent · Coordinate descent · Stochastic gradient · Variance-reduction · Boosting · Support vector machines · L1-regularization

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 [28]. 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.

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46128-1_50) contains supplementary material, which is available to authorized users.

P. Frasconi et al. (Eds.): ECML PKDD 2016, Part I, LNAI 9851, pp. 795–811, 2016.
DOI: 10.1007/978-3-319-46128-1_50

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 [22], but four other recently-considered conditions are *essential strong convexity* (ESC) [20], *weak strong convexity* (WSC) [25], the *restricted secant inequality* (RSI) [45], and the *quadratic growth* (QG) condition [2]. Some of these conditions have different names in the special case of convex functions: a convex function satisfying RSI is said to satisfy *restricted strong convexity* (RSC) [45] while a convex function satisfying QG is said to satisfy *optimal strong convexity* (OSC) [19] or (confusingly) WSC [23]. 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 [31], 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 [21]. 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 (Sect. 2.3). In Sect. 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 closely-related 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. This leads to a simple analysis showing linear convergence of methods for training support vector machines. It also implies that we obtain a linear convergence rate for ℓ_1-regularized least squares problems, showing that the extra conditions previously assumed to derive linear converge rates in this setting are in fact not needed.

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 [31]. 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 of the inequality

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

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 Sect. 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 Sect. 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 [19,25,44]. In Appendix 2.1, 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

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

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, 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*. Invexity was first introduced by Hanson in 1981 [12], and has been used in the context of learning output kernels [8]. Craven and Glover [7] 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. Indeed, 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 2.3). 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 2.3, 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 [9] to analyze the Frank-Wolfe algorithm. Further, inspired by the resilient backpropagation (RPROP) algorithm of Riedmiller and Braun [32], in Appendix 3 we also give the first convergence rate analysis for sign-based gradient descent methods.

3.1 Randomized Coordinate Descent

Nesterov [29] 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

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

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 [29], 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. [30] 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. [30] 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,

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

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 [24] 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 (SG) 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 $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \le C^2$ 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*

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

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,

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

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 $\mathbb{E}[\nabla f_{i_k}(x_k)] = f'(x_k)$ 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 [27]. 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 [4], 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 minimizing *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 [33]. 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 [16] does not have this drawback. In Appendix 3.4, we give a new analysis of the SVRG method that shows that it achieves a linear convergence rate under the PL inequality. Similar results for finite-sum methods under the PL inequality recently appeared in the works of Reddi et al. [36,37]. Garber and Hazan [10] have also given a related result in the context of an improved algorithm for principal component analysis (PCA), showing that the f_i do not need to be convex in order to achieve a linear convergence rate. However, their result still assumes that f is SC while our analysis only assumes the PL inequality is satisfied.

4 Proximal-Gradient Generalization

Attouch and Bolte [3] consider a generalization of the PL inequality due to Kurdyak to give conditions under which the classic proximal-point algorithm achieves a linear convergence rate for non-smooth problems (called the KL inequality). However, in practice proximal-*gradient* methods are more relevant to many machine learning problems. While the KL inequality has been used to show local linear convergence of proximal-gradient methods [6,18], in this section we propose a different generalization of the PL inequality that yields a simple global linear convergence analysis.

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

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

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* $1/L$*,*

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

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 function 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. □

We note that the condition $\mu \leq L$ is implicit in the definition of the proximal-PL inequality, but this is not restrictive since we can simply set μ to a smaller value to satisfy this. While other conditions have been proposed to show linear convergence rates of proximal-gradient methods without SC [17,44], their analyses tend to be much more complicated than the above while, as we discuss in the next section, the proximal-PL inequality includes the standard scenarios where these 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 Appendices 4.1, 4.2, and 4.4.

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 [34], 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. In Appendices 4.2 and 4.4 we show that L1-regularized least squares and the support vector machine dual (respectively) fall into this category, and we discuss these two notable cases further below.

We expect that it is possible to show the proximal-PL inequality holds in other cases where the proximal-gradient achieves a linear convergence rate like the case of group L1-regularization [40] and nuclear-norm regularization [14].

4.2 Least Squares with L1-Regularization

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

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

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 sublinear convergence rates. There subsequently have been a variety of works showing that linear convergence rates can be achieved under additional assumptions. For example, Gu et al. [11] 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 [43] also assume the RIP property and show linear convergence using a homotopy method that slowly decreases the value of λ. Agarwal et al. [1] give a linear convergence rate under a *modified restricted strong convexity* and *modified restricted smoothness* assumption. In Appendix 4.2 we show that *any* L1-regularized least squares problem satisfies the QG property if we use a descent method and thus by convexity also satisfies the proximal-PL inequality. Thus, Theorem 5 implies a global linear convergence rate for these problems without making additional assumptions or making any modifications to the algorithm. A similar result recently appeared in the work of Necoara and Clipici [26] under a generalized EB, but with a much more complicated analysis.

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

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

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 4.3 and although it is more complicated than the proofs of Theorems 4 and 5, it is still simpler than existing proofs for proximal coordinate descent under SC [39]. It is also possible to analyze stochastic proximal-gradient algorithms, and indeed Reddi et al. use the proximal-PL inequality to analyze finite-sum methods in the proximal stochastic case [38].

4.4 Support Vector Machines

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

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

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 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 matrix M and constant U. This 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, since Hush et al. [15] show that we can obtain an ϵ-accurate solution to the primal problem with an $O(\epsilon^2)$-accurate solution to the dual problem, this also implies a linear convergence rate for stochastic dual coordinate ascent on the primal problem. Global linear convergence rates for SVMs have also been shown by others [23,41,42], but again we note that these works lead to much 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 section 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 [35]. While Shalev-Shwartz and Zhang [35] 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. Throughout the paper, we have also pointed out how our analysis implies new 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 and Martin Takáč for valuable discussions. 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.

References

1. Agarwal, A., Negahban, S.N., Wainwright, M.J.: Fast global convergence rates of gradient methods for high-dimensional statistical recovery. Ann. Statist. **40**, 2452–2482 (2012)
2. Anitescu, M.: Degenerate nonlinear programming with a quadratic growth condition. SIAM J. Optim. **10**, 1116–1135 (2000)
3. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. Ser. B **116**, 5–16 (2009)
4. Bach, F., Moulines, E.: Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In: Advances in Neural Information Processing Systems (NIPS), pp. 773–791 (2013)
5. Ben-Israel, A., Mond, B.: What is invexity? J. Austral. Math. Soc. **28**, 1–9 (1986)
6. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From Error Bounds to the Complexity of First-Order Descent Methods for Convex Functions. arXiv:1510.08234 (2015)
7. Craven, B.D., Glover, B.M.: Invex functions and duality. J. Austral. Math. Soc. **39**, 1–20 (1985)
8. Dinuzzo, F., Ong, C.S., Gehler, P., Pillonetto, G.: Learning output kernels with block coordinate descent. In: Proceedings of the 28th ICML, pp. 49–56 (2011)
9. Garber, D., Hazan, E.: Faster rates for the Frank-Wolfe method over strongly-convex sets. In: Proceedings of the 32nd ICML, pp. 541–549 (2015)
10. Garber, D., Hazan, E.: Faster and Simple PCA via Convex Optimization. arXiv:1509.05647v4 (2015)
11. Gu, M., Lim, L.-H., Wu, C.J.: ParNes: a rapidly convergent algorithm for accurate recovery of sparse and approximately sparse signals. Numer. Algor. **64**, 321–347 (2013)
12. Hanson, M.A.: On sufficiency of the Kuhn-Tucker conditions. J. Math. Anal. Appl. **80**, 545–550 (1981)
13. Hoffman, A.J.: On approximate solutions of systems of linear inequalities. J. Res. Nat. Bur. Stand. **49**, 263–265 (1952)
14. Hou, K., Zhou, Z., So, A.M.-C., Luo, Z.-Q.: On the linear convergence of the proximal gradient method for trace norm regularization. In: Advances in Neural Information Processing Systems (NIPS), pp. 710–718 (2013)
15. Hush, D., Kelly, P., Scovel, C., Steinwart, I.: QP algorithms with guaranteed accuracy and run time for support vector machines. J. Mach. Learn. Res. **7**, 733–769 (2006)
16. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in Neural Information Processing Systems (NIPS), pp. 315–323 (2013)
17. Kadkhodaie, M. Sanjabi, M., Luo, Z.-Q.: On the Linear Convergence of the Approximate Proximal Splitting Method for Non-Smooth Convex Optimization. arXiv:1404.5350v1 (2014)
18. Li, G., Pong, T.K.: Calculus of the Exponent of Kurdyka-Łojasiewicz Inequality and its Applications to Linear Convergence of First-Order Methods. arXiv:1602.02915v1 (2016)
19. Liu, J., Wright, S.J.: Asynchronous stochastic coordinate descent: parallelism and convergence properties. SIAM J. Optim. **25**, 351–376 (2015)
20. Liu, J., Wright, S.J., Ré, C., Bittorf, V., Sridhar, S.: An Asynchronous Parallel Stochastic Coordinate Descent Algorithm. arXiv:1311.1873v3 (2014)

21. Łojasiewicz, S.: A Topological Property of Real Analytic Subsets (in French). Coll. du CNRS, Les équations aux dérivées partielles, vol. 117, pp. 87–89 (1963)
22. Luo, Z.-Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**, 157–178 (1993)
23. Ma, C., Tappenden, T., Takáč, M.: Linear Convergence of the Randomized Feasible Descent Method Under the Weak Strong Convexity Assumption. arXiv:1506.02530 (2015)
24. Meir, R., Rätsch, G.: An introduction to boosting and leveraging. In: Mendelson, S., Smola, A.J. (eds.) Advanced Lectures on Machine Learning. LNCS (LNAI), vol. 2600, pp. 118–183. Springer, Heidelberg (2003). doi:10.1007/3-540-36434-X_4
25. Necoara, I., Nesterov, Y., Glineur, F.: Linear Convergence of First Order Methods for Non-Strongly Convex Optimization. arXiv:1504.06298v3 (2015)
26. Necoara, I., Clipici, D.: Parallel random coordinate descent method for composite minimization: convergence analysis and error bounds. SIAM J. Optim. **26**, 197–226 (2016)
27. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**, 1574–1609 (2009)
28. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Dordrecht (2004)
29. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**, 341–362 (2012)
30. Nutini, J., Schmidt, M., Laradji, I.H., Friedlander, M., Koepke, H.: Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In: Proceedings of the 32nd ICML, pp. 1632–1641 (2015)
31. Polyak, B.T.: Gradient methods for minimizing functionals. Zh. Vychisl. Mat. Mat. Fiz. **3**, 643–653 (1963). (in Russian)
32. Riedmiller, M., Braun, H.: RPROP - a fast adaptive learning algorithm. In: Proceedings of ISCIS VII (1992)
33. Roux, N.L., Schmidt, M., Bach, F.R.: A stochastic gradient method with an exponential convergence rate for finite training sets. In: Advances in Neural Information Processing Systems (NIPS), pp. 2672–2680 (2012)
34. Schmidt, M., Roux, N.L., Bach, F.R.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems (NIPS), pp. 1458–1466 (2011)
35. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. J. Mach. Learn. Res. **14**, 567–599 (2013)
36. Reddi, S.J., Sra, S., Poczos, B., Smola, A.: Fast Incremental Method for Nonconvex Optimization. arXiv:1603.06159 (2016)
37. Reddi, S.J., Hefny, A., Sra, S., Poczos, B., Smola, A.: Stochastic Variance Reduction for Nonconvex Optimization. arXiv:1603.06160 (2016)
38. Reddi, S.J., Sra, S., Poczos, B., Smola, A.,: Fast Stochastic Methods for Nonsmooth Nonconvex Optimization. arXiv:1605.06900 (2016)
39. Richtárik, P., Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Math. Program. Ser. A **144**, 1–38 (2014)
40. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. Ser. B **125**, 263–295 (2010)
41. Tseng, P., Yun, S.: Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. J. Optim. Theory Appl. **140**, 513–535 (2009)

42. Wang, P.-W., Lin, C.-J.: Iteration complexity of feasible descent methods for convex optimization. J. Mach. Learn. Res. **15**, 1523–1548 (2014)
43. Xiao, L., Zhang, T.: A proximal-gradient homotopy method for the sparse least-squares problem. SIAM J. Optim. **23**, 1062–1091 (2013)
44. Zhang, H.: The Restricted Strong Convexity Revisited: Analysis of Equivalence to Error Bound and Quadratic Growth. arXiv:1511.01635 (2015)
45. Zhang, H., Yin, W.: Gradient Methods for Convex Minimization: Better Rates Under Weaker Conditions. arXiv:1303.4645v2 (2013)