Linear Convergence of Proximal-Gradient Methods under the Polyak-Łojasiewicz Condition

Hamed Karimi^{1,2} **and Mark Schmidt**^{1}
^{1}University of British Columbia, ^{2}1QBit Information Technologies

Abstract

In 1963, Polyak proposed a simple condition that is sufficient to show that gradient descent has a global linear convergence rate. 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 four 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, as well as stochastic gradient methods with decreasing or constant step sizes. We then consider a natural generalization of the inequality that applies to proximal-gradient methods for non-smooth optimization, and show this means other conditions that have been proposed to achieve linear convergence for ℓ_1-regularized least squares are unnecessary. Along the way, we give new convergence results for a wide variety of problems in machine learning: least squares, logistic regression, boosting, L1-regularization, support vector machines, stochastic dual coordinate ascent, and stochastic variance-reduced gradient methods.

1 Introduction

In this work we consider the basic problem of minimizing a smooth function and consider 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 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 most well-known examples is the *error bounds* of Luo and Tseng [1993]. Three recently-considered relaxations are *essential strong convexity* [Liu et al., 2013], *optimal strong convexity* [Liu and Wright, 2015], and *restricted strong convexity* [Zhang and Yin, 2013]. The proofs of linear convergence under these relaxations are often not straightforward, and it is rarely discussed whether any of these four conditions are stronger or weaker than the others.

In this work, we consider a much older condition that we refer to as the Polyak-Łojasiewicz (PL) inequality. This inequality was 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, we show that f satisfying the PL inequality is actually a *weaker condition* than all four of the more recent conditions discussed in the previous paragraph. In Section 2.2 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). 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 this scheme without strong-convexity. This result implies that we obtain a linear convergence rate for ℓ_1-regularized least squares problems, showing that conditions that have been assumed to derive linear converge rates in this setting are in fact not needed. Finally, we consider coordinate optimization methods in this setting, and show that the generalized PL inequality gives the first global linear convergence rate for training support vector machines (among other models).

2 Polyak-Łojasiewicz Inequality

We 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. We also assume 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 it holds for some $\mu > 0$ that

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

for all x. Note that this inequality implies that every stationary point is a global minimum, but unlike strong-convexity 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 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

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

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. □

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 linear convergence on non-convex problems.

2.1 Weaker Condition and Relevant problems

As mentioned in the introduction, there are many assumptions that can be made in order to show that gradient descent achieves a linear convergence rate. These other assumptions all require the function f to be convex, and lead to more complicated proofs than the one above. Although we omit the proofs due to space restrictions, for convex functions we can show that the PL inequality is weaker than all four of the other conditions that are typically used to show linear convergence without strong-convexity: the *error bound property* [Luo and Tseng, 1993], *optimal strong convexity* [Liu and Wright, 2015], *essential strong convexity* [Liu et al., 2013], and *restricted strong convexity* [Zhang and Yin, 2013].

It is not easy to characterize the class of functions for which the PL inequality is satisfied, we note that μ-strongly convex functions satisfy the PL inequality with the same value of μ. Further, if we have a function of the $f(x) = g(Ax)$ for a matrix A and a strongly-convex function g, we can also show that it satisfies the PL inequality. Thus, the PL inequality is satisfied for problems like least squares.

2.2 Randomized Coordinate Descent

Nesterov [2012] shows that for strongly convex functions 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. 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

$$f(x + \alpha e_i) \leq f(x) + \alpha \nabla_i f(x) + \frac{L}{2}\alpha^2 \tag{5}$$

for any coordinate i and for any real number α.

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

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

where e_i is the ith unit vector. If we choose the variable to update i_k uniformly at random then the algorithm has an expected linear convergence rate of

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

In the extended version, we also analyze greedy coordinate descent methods (including variants of boosting), stochastic gradient methods, and stochastic variance-reduced gradient methods.

3 Proximal-Gradient Generalization

Attouch and Bolte [2009] consider a generalization of the 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. In this section, we consider a different generalization of the PL inequality that is relevant to the class of proximal-*gradient* methods. In particular, consider problems of the form

$$\operatorname*{argmin}_{x \in \mathbb{R}^d} F(x) = f(x) + g(x) \tag{7}$$

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 ∞ otherwise. Although we could apply proximal-point algorithms to this problem, their inner minimization steps are typically not practical. However, proximal-gradient methods are well-suited to solving problems with this structure.

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

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

$$\text{for } \mathcal{D}_g(x,\mu) \equiv -2\mu \min_y [\langle \nabla f(x), y - x\rangle + \tfrac{\mu}{2}||y-x||^2 + g(y) - g(x)]. \tag{8}$$

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 give a linear convergence result under this assumption.

Theorem 3. *Consider problem* (7)*, 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* (8)*. Then the proximal-gradient method with a step-size of* $1/L$*,*

$$x_{k+1} \quad = \quad \underset{y}{\mathit{argmin}}[\langle \nabla f(x_k), y - x_k\rangle + \frac{L}{2}||y - x_k||^2 + g(y) - g(x_k)] \tag{9}$$

converges linearly to the optimal value F^*

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

Further, if g is fully separable function $g(x) = \sum_i g_i(x_i)$, then randomized proximal-coordinate descent also converges linearly.

$$x_{k+1} \quad = \quad \mathit{argmin}_\alpha [\alpha \nabla_{i_k} f(x_k) + \frac{L}{2}\alpha^2 + g_{i_k}(x_{i_k} + \alpha) - g_{i_k}(x_{i_k})], \tag{10}$$

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

3.1 Relevant Problems

As with the PL inequality, we now list several important function classes that will satisfy the proximal-PL inequality (8).

1. The inequality is satisfied if f is strongly convex, which is a usual assumption to show a linear convergence rate for the proximal-gradient algorithm [Schmidt et al., 2011].
2. The inequality is satisfied if f has the form $f(x) = g(Ax)$ for a strongly-convex function g and a matrix A. This includes ℓ_1-regularized least squares problems as a special case. Thus, for these problems we do not need other properties/algorithms like the restricted isometry property, homotopy methods, or manifold identification.
3. If $F = f + g$ has the optimal strong convexity property [Liu and Wright, 2015], then F satisfies the inequality.

3.2 Support Vector Machines

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

$$\underset{x \in \mathbb{R}^d}{\text{argmin}} \frac{\lambda}{2} x^T x + \sum_{i}^{m} \max(0, 1 - b_i w^T a_i). \tag{11}$$

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

$$\min_y f(t)y \quad = \quad \frac{1}{2}y^T M y - \sum y_i \qquad\qquad y_i \in [0, U], \tag{12}$$

for a particular matrix M and constant U. This problem satisfies the proximal-PL inequality so the result of the previous section applies. Thus, coordinate optimization achieves a linear convergence rate in terms of optimizing the dual objective. Further, 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, we have shown that a global linear convergence rate can be achieved by a stochastic algorithm for training SVMs.

However, the result of the previous section is not only restricted to SVMs. Indeed, the result of this previous section implies a linear convergence rate a wide range of ℓ_2-regularized linear prediction problems, as 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 implies that it holds even in many non-smooth cases like SVMs.

References

H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

D. Hush, P. Kelly, C. Scovel, and I. Steinwart. Qp algorithms with guaranteed accuracy and run time for support vector machines. *The Journal of Machine Learning Research*, 7:733–769, 2006.

J. Liu and S. J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.

J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *arXiv preprint arXiv:1311.1873*, 2013.

S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *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. *Annals of Operations Research*, 46(1):157–178, 1993.

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 Journal on Optimization*, 22(2):341–362, 2012.

B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

M. Schmidt, N. L. Roux, and F. R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

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