Linear Convergence under the Polyak-Łojasiewicz Inequality

Hamed Karimi, Julie Nutini, Mark Schmidt

University of British Columbia

Linear of Convergence of Gradient-Based Methods

- Fitting most machine learning models involves optimization.
- Most common algorithm is gradient descent (GD) and variants:
  - Stochastic gradient, quasi-Newton, coordinate descent, and so on.
- Standard global convergence rate result for GD:

Smoothness + Strong-Convexity $\Rightarrow$ Linear Convergence

  - Error on iteration t is $O(\rho^t)$.
- But even simple models are often not strongly-convex.
  - Least squares, logistic regression, etc.
- **This talk**: How much can we relax strong-convexity?

Smoothness + ~~Strong-Convexity~~ ??? $\Rightarrow$ Linear Convergence

Polyak-Łojasiewicz (PL) Inequality

- Polyak [1963] showed linear convergence of GD assuming

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

that gradient grows as quadratic function of sub-optimality.

- Holds for SC problems, but also problems of the form

$$f(x) = g(Ax), \quad \text{for strongly-convex } g.$$

- Includes least squares, logistic regression (on compact set), etc.
- A special case of the Łojasiewicz' inequality [1963].
  - We'll call this the Polyak-Łojasiewicz (PL) inequality.
- Using the PL inequality we can show

Smoothness + **PL Inequality** ~~Strong-Convexity~~ $\Rightarrow$ Linear Convergence

Linear Convergence of GD under the PL Inequality

- Consider the basic unconstrained smooth optimization,

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

where f satisfies the PL inequality and ∇f is Lipschitz continuous,

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

- Applying GD with a constant step-size of $1/L$,

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

we have

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

- Subtracting f^* and applying recursively gives global linear rate,

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k [f(x^0) - f^*].$$

Linear Convergence under the PL Inequality

- Proof is simple (simpler than than with SC).
- Does not require uniqueness of solution (unlike SC).
- Does not imply convexity (unlike SC).

Weaker Conditions than Strong Convexity (SC)

- How does PL inequality [1963] relate to more recent conditions?
  - EB: error bounds [Luo and Tseng, 1993].
  - QG: quadratic growth [Anitescu, 2000]
  - ESC: essential strong convexity [Liu et al., 2013].
  - RSI: restricted secant inequality [Zhang & Yin, 2013].
    - RSI plus convexity is "restricted strong-convexity".
  - Semi-strong convexity [Gong & Ye, 2014].
    - Equivalent to QG plus convexity.
  - Optimal strong convexity [Liu & Wright, 2015].
    - Equivalent to QG plus convexity.
  - WSC: weak strong convexity [Necoara et al., 2015].
- Proofs are often more complicated under these conditions.
- Are they more general?

Relationships Between Conditions

For a function f with a Lipschitz-continuous gradient, we have:

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

If we further assume that f is convex, then

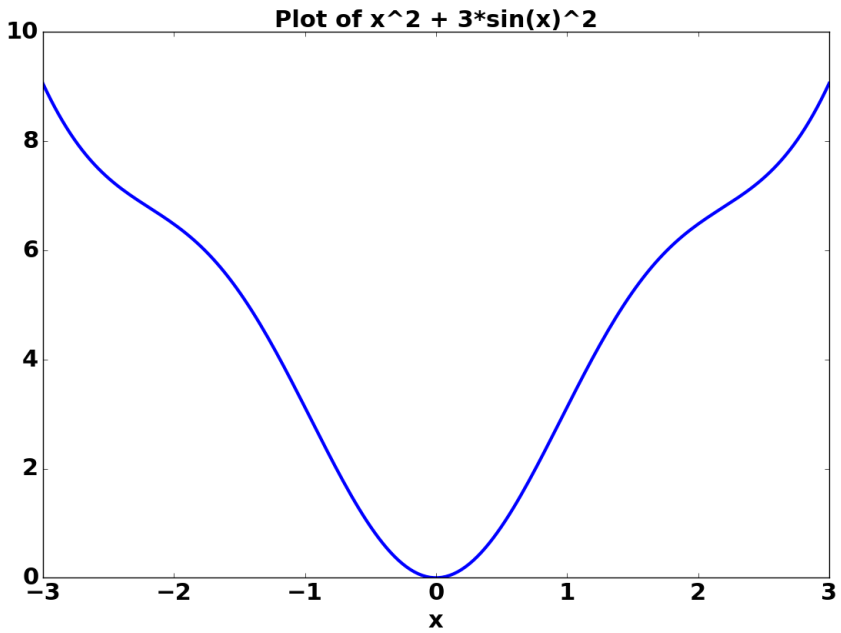
$$(RSI) \equiv (EB) \equiv (PL) \equiv (QG).$$

- QG is the weakest condition but allows non-global local minima.
- PL $\equiv$ EB are most general conditions giving global min.
  - More recent condition are not needed.

PL Inequality and Invexity

- While PL imply doesn't convexity, it implies invexity.
  - For smooth f, invexity $\leftrightarrow$ all stationary points are global optimum.
- Example of invex but non-convex function satisfying PL:

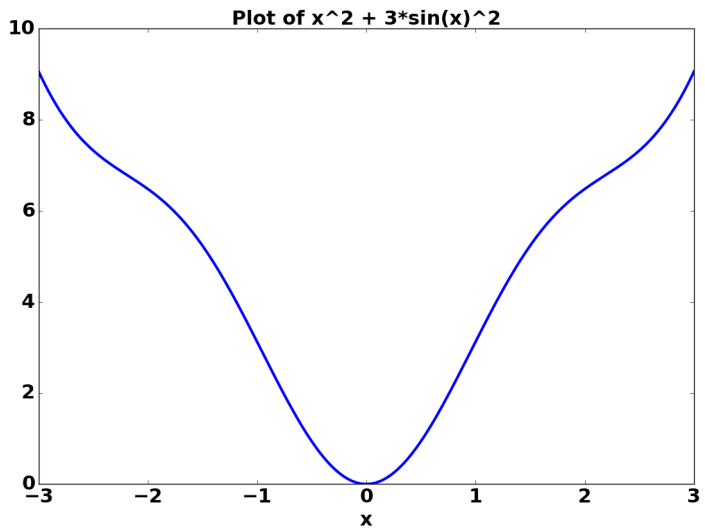
$$f(x) = x^2 + 3\sin^2(x).$$



- Maybe "strong invexity" is a better name?

PL Inequality and Non-Convex Functions

- Many important models don't satisfy invexity.
- For these problems we often divide analysis into two phases:
  - Global convergence: iterations needed to get "close" to minimizer.
  - Local convergence: how fast does it converge near the minimizer.
- Usually, local convergence assumes SC near minimizer.
  - If we assume PL, local convergence phase may be much earlier.



  - Holds for $\mathbb{R}$ on this function even though non-convex on $[-3, 3]$.

Convergence of Huge-Scale Methods

- For large datasets, we typically don't use GD.
  - But the PL inequality can be used to analyze other algorithms.
- We'll use PL for coordinate descent and stochastic gradient.
  - Garber & Hazan [2015] consider Frank-Wolfe.
  - Reddi et al. [2016] consider other stochastic algorithms.
  - In Karimi et al. [2016] we consider sign-based gradient methods.

Random and Greedy Coordinate Descent

- For randomized coordinate descent under PL we have

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

  where L_c is coordinate-wise Lipschitz constant of ∇f.

- Faster than GD rate if iterations are d times cheaper.

- For greedy coordinate descent under PL we have faster rate

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

  where μ_1 is the PL constant in the L_∞-norm,

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

- Gives rate for some boosting variants [Meir and Rätsch, 2003].

Stochastic Gradient Methods

- Stochastic gradient (SG) methods apply to general problem

$$\operatorname*{argmin}_{x \in \mathbb{R}^d} f(x) = \mathbb{E}[f_i(x)],$$

and we usually focus on the special case of a finite sum

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

- SG methods use the iteration

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

where ∇f_{i_k} is an unbiased gradient approximation.

Stochastic Gradient Methods

With $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ *the SG method satisfies*

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

while with α_k *set to constant* α *we have*

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

- $O(1/k)$ rate without strong-convexity (or even convexity).
- Fast reduction of sub-optimality under small constant step size.
- Our work and Reddi et al. [2016] consider finite sum case:
  - Analyze stochastic variance-reduced gradient (SVRG) method.
  - Obtain linear convergence rates.

PL Generalizations for Non-Smooth Problems

- What can we say about non-smooth problems?
  - Well-known generalization of PL is the KL inequality.
- Attach and Bolte [2009] show linear rate for proximal-point.
- But proximal-gradient methods are more relevant for ML.
  - KL inequality has been used to show local rate for this method.
- We propose different PL generalization giving simple global rate.

Proximal-PL Inequality

- Proximal-gradient methods apply to the problem

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

where ∇f is L-Lipschitz but g may be non-smooth.

  - For example, L1-regularization or simple convex constraints.

- We say that F satisfies the proximal-PL inequality if

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

where

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

- Condition is ugly but it yields extremely-simple proof:

$$\begin{aligned}
F(x_{k+1}) &= f(x_{k+1}) + g(x_k) + g(x_{k+1}) - g(x_k) \\
&\leq F(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} ||x_{k+1} - x_k||^2 + g(x_{k+1}) - g(x_k) \\
&\leq F(x_k) - \frac{1}{2L} \mathcal{D}_g(x_k, L) \\
&\leq F(x_k) - \frac{\mu}{L}[F(x_k) - F^*] \Rightarrow F(x^k) - F^* \leq \left(1 - \frac{\mu}{L}\right)^k [F(x^0) - F^*].
\end{aligned}$$

Relevant Problems for Proximal-PL

- Proximal PL is satisfied when:
  - f is strongly-convex.
  - f satisfies PL and g is constant.
  - $f = h(Ax)$ for SC g and is indicator of polyhedral set.
  - F is convex and satisfies QG.
  - Any function satisfying KL inequality.
    - We've shown that proxlmal-PL and KL inequality are equivalent.
- Includes dual support vector machine (SVM) problem.
- Includes L1-regularized least squares (LASSO) problem:
  - No need for RIP, homotopy, modified restricted strong convexity,. . .

- We also analyze proximal coordinate descent under PL.
  - Implies linear rate of SDCA for SVMs.
- Reddi et al. [2016] analyze proximal-SVRG and proximal-SAGA.

Summary

- In 1963, Polyak proposed a condition for linear rate of GD.
  - Gives trivial proof and is weaker than more recent conditions.
- We can use the inequality to analyze huge-scale methods:
  - Coordinate descent, stochastic gradient, SVRG, etc.
- We give proximal-gradient generalization:
  - Standard algorithms have linear rate for SVM and LASSO.

Boris Polyak boris@ipu.ru via cs.ubc.ca 9/2/16

to hamedkarim, jnutini, schmidtm

Dear colleagues,

it was a pleasure to read your paper **Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Lojasiewicz Condition.**

Actually it was a surprise for me why the simple condition from my old publication attracted so little attention, and now

your work exhibits its importance.

You cite my paper in Russian, find attached its English translation. I also attach our paper with Nesterov where some extensions of the condition (and some its applications) can be found.

Best regards,

Boris Polyak

Thank you for the invitation!