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.

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:
  - If f is strongly-convex (SC) then GD has linear convergence.
    - Error on iteration t is $O(\rho^t)$.

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:
  - If f is strongly-convex (SC) then GD has linear convergence.
    - Error on iteration t is $O(\rho^t)$.
- But even simple models are often not strongly-convex.
  - Least squares, logistic regression, etc.

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:
  - If f is strongly-convex (SC) then GD has 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?

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.

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.

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.

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,

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

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 $1/L$,

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

Linear Convergence of GD under the PL Inequality

- Consider the basic unconstrained smooth optimization,

$$\underset{x \in \mathbb{R}^d}{\text{argmin}}\, f(x),$$

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

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

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

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

we have

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

Linear Convergence of GD under the PL Inequality

- Consider the basic unconstrained smooth optimization,

$$\underset{x \in \mathbb{R}^d}{\text{argmin}}\ f(x),$$

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).

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?

Weaker Conditions than Strong Convexity (SC)

- How does PL inequality [1963] relate to more recent conditions?
  - EB: error bounds [Luo and Tseng, 1993].

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]

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]
    - QG plus convexity is "optimal strong convexity" [Liu & Wright, 2015].

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]
    - QG plus convexity is "optimal strong convexity" [Liu & Wright, 2015].
  - ESC: essential strong convexity [Liu et al., 2013].

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]
    - QG plus convexity is “optimal strong convexity” [Liu & Wright, 2015].
  - ESC: essential strong convexity [Liu et al., 2013].
  - RSI: restricted secant inequality [Zhang & Yin, 2013].
    - RSI plus convexity is “restricted strong convexity”.

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]
    - QG plus convexity is “optimal strong convexity” [Liu & Wright, 2015].
  - ESC: essential strong convexity [Liu et al., 2013].
  - RSI: restricted secant inequality [Zhang & Yin, 2013].
    - RSI plus convexity is “restricted strong convexity”.
  - WSC: weak strong convexity [Necoara et al., 2015].
    - Name is also sometimes used for QG plus convexity.

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]
    - QG plus convexity is "optimal strong convexity" [Liu & Wright, 2015].
  - ESC: essential strong convexity [Liu et al., 2013].
  - RSI: restricted secant inequality [Zhang & Yin, 2013].
    - RSI plus convexity is "restricted strong convexity".
  - WSC: weak strong convexity [Necoara et al., 2015].
    - Name is also sometimes used for QG plus convexity.
- Proofs are more complicated under all these conditions.
- Are they more general?

Relationships Between Conditions

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

$$(SC) \rightarrow (ESC) \rightarrow (WSC) \rightarrow (RSI) \rightarrow (EB) \rightarrow (PL),$$

and $(RSI) \rightarrow (QG)$*.*

Relationships Between Conditions

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

$$(SC) \rightarrow (ESC) \rightarrow (WSC) \rightarrow (RSI) \rightarrow (EB) \rightarrow (PL),$$

and $(RSI) \rightarrow (QG)$. If we further have that f is convex then

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

Relationships Between Conditions

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

$$(SC) \to (ESC) \to (WSC) \to (RSI) \to (EB) \to (PL),$$

and $(RSI) \to (QG)$. If we further have that f is convex then

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

- For convex functions PL covers all cases.
  - Don't need the other conditions.

Relationships Between Conditions

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

$$(SC) \rightarrow (ESC) \rightarrow (WSC) \rightarrow (RSI) \rightarrow (EB) \rightarrow (PL),$$

and $(RSI) \rightarrow (QG)$. If we further have that f is convex then

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

- For convex functions PL covers all cases.
  - Don't need the other conditions.
- For non-convex functions PL and QG are weakest.
  - But QG allows sub-optimal local minima.

Relationships Between Conditions

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

$$(SC) \rightarrow (ESC) \rightarrow (WSC) \rightarrow (RSI) \rightarrow (EB) \rightarrow (PL),$$

and $(RSI) \rightarrow (QG)$. If we further have that f is convex then

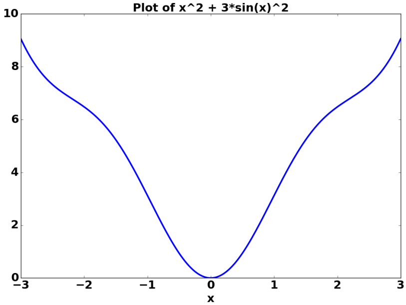
$$(RSI) \equiv (QG) \rightarrow (PL).$$

- For convex functions PL covers all cases.
  - Don't need the other conditions.
- For non-convex functions PL and QG are weakest.
  - But QG allows sub-optimal local minima.
- PL is most general that allows linear rate to global optimum.
  - Though may be other relations like $PL \rightarrow EB$ and $PL \rightarrow QG$.

PL Inequality and Invexity

- While PL 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:

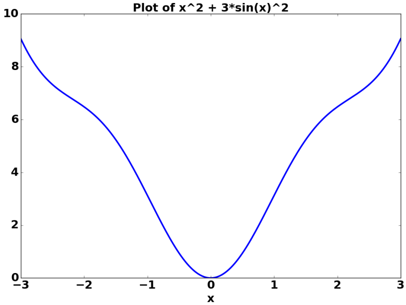
$$f(x) = x^2 + 3\sin^2(x).$$



PL Inequality and Invexity

- While PL 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

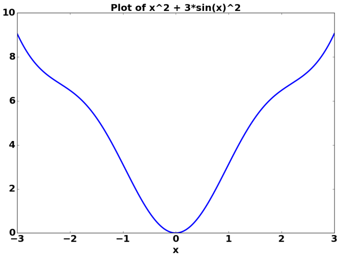
- Many important models don’t satisfy invexity.

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.

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.



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.

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.

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

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

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

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

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

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

Stochastic Gradient Methods

- Stochastic gradient (SG) methods apply to general problem

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

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*

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

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*

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

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.

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.

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.

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.

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

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

Proximal-PL Inequality

- Proximal-gradient methods apply to the problem

$$\underset{x\in\mathbb{R}^d}{\text{argmin}}\, F(x) = f(x) + g(x),$$

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

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

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

  where

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

Proximal-PL Inequality

- Proximal-gradient methods apply to the problem

$$\underset{x \in \mathbb{R}^d}{\text{argmin}}\, F(x) = f(x) + g(x),$$

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

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

Relevant Problems for Proximal-PL

- We also analyze proximal coordinate descent under PL.
  - Reddi et al. [2016] analyze proximal-SVRG and proximal-SAGA.

Relevant Problems for Proximal-PL

- We also analyze proximal coordinate descent under PL.
  - Reddi et al. [2016] analyze proximal-SVRG and proximal-SAGA.
- Proximal PL is satisfied when:
  - f is SC.
  - f satisfies PL and g is constant.
  - $f = h(Ax)$ for SC g and is indicator of convex set.
  - F is convex and satisfies QG.

Relevant Problems for Proximal-PL

- We also analyze proximal coordinate descent under PL.
  - Reddi et al. [2016] analyze proximal-SVRG and proximal-SAGA.
- Proximal PL is satisfied when:
  - f is SC.
  - f satisfies PL and g is constant.
  - $f = h(Ax)$ for SC g and is indicator of convex set.
  - F is convex and satisfies QG.
- Includes dual support vector machine (SVM) problem:
  - Implies linear rate of SDCA for SVMs.

Relevant Problems for Proximal-PL

- We also analyze proximal coordinate descent under PL.
  - Reddi et al. [2016] analyze proximal-SVRG and proximal-SAGA.
- Proximal PL is satisfied when:
  - f is SC.
  - f satisfies PL and g is constant.
  - $f = h(Ax)$ for SC g and is indicator of convex set.
  - F is convex and satisfies QG.
- Includes dual support vector machine (SVM) problem:
  - Implies linear rate of SDCA for SVMs.
- Includes L1-regularized least squares (LASSO) problem:
  - No need for RIP, homotopy, modified restricted strong convexity,...

Summary

- In 1963, Polyak proposed a condition for linear rate of GD.
  - Gives trivial proof and is weaker than more recent conditions.

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.

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.