Linear Convergence under the Polyak-Łojasiewicz Inequality

Hamed Karimi, Julie Nutini and Mark Schmidt

The University of British Columbia

LCI Forum

February 28^{th}, 2017

Linear Convergence of Gradient-Based Methods

- Fitting most machine learning models involves optimization.
- Most common algorithm is gradient descent (GD) and variants:
  - e.g., stochastic gradient, quasi-Newton, coordinate descent, etc.

Linear Convergence of Gradient-Based Methods

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

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

  - Error on iteration k is $O(\rho^k)$.

Linear Convergence of Gradient-Based Methods

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

$$\text{Smoothness} \quad + \quad \text{Strong-Convexity} \quad \Rightarrow \quad \text{Linear Convergence}$$

  - Error on iteration k is $O(\rho^k)$.
- But even simple models are often not strongly-convex.
  - e.g., least-squares, logistic regression, etc.

Linear Convergence of Gradient-Based Methods

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

$$\text{Smoothness} \quad + \quad \text{Strong-Convexity} \quad \Rightarrow \quad \text{Linear Convergence}$$

  - Error on iteration k is $O(\rho^k)$.
- But even simple models are often not strongly-convex.
  - e.g., least-squares, logistic regression, etc.

⋆ **This talk**: How much can we relax strong-convexity?

$$\text{Smoothness} \quad + \quad \overset{\textbf{???}}{\cancel{\text{Strong-Convexity}}} \quad \Rightarrow \quad \text{Linear Convergence}$$

Polyak-Łojasiewicz Inequality

- Polyak [1963] showed linear convergence of GD assuming

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

i.e., the gradient grows as a quadratic function of sub-optimality.

Polyak-Łojasiewicz Inequality

- Polyak [1963] showed linear convergence of GD assuming

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

i.e., the gradient grows as a quadratic function of sub-optimality.

- Holds for strongly-convex problem, 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 Inequality

- Polyak [1963] showed linear convergence of GD assuming
$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\left(f(x) - f^*\right),$$
i.e., the gradient grows as a quadratic function of sub-optimality.
- Holds for strongly-convex problem, 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 Łojasiewicz' inequality [1963].
    - We call this the Polyak-Łojasiewicz (PL) inequality.

Polyak-Łojasiewicz Inequality

- Polyak [1963] showed linear convergence of GD assuming

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

  i.e., the gradient grows as a quadratic function of sub-optimality.

- Holds for strongly-convex problem, 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 Łojasiewicz' inequality [1963].
  - We call this the Polyak-Łojasiewicz (PL) inequality.
- Using the PL inequality, we show

Smoothness + **PL Inequality** ~~Strong-Convexity~~ ⇒ Linear Convergence

- Consider the basic unconstrained smooth optimization problem,

$$\min_{x \in \mathbb{R}^d} \; f(x),$$

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

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

Linear Convergence of GD under the PL Inequality

- Consider the basic unconstrained smooth optimization problem,

$$\min_{x \in \mathbb{R}^d} f(x),$$

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 $1/L$,

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

Linear Convergence of GD under the PL Inequality

- Consider the basic unconstrained smooth optimization problem,

$$\min_{x \in \mathbb{R}^d} f(x),$$

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

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

Linear Convergence of GD under the PL Inequality

- Consider the basic unconstrained smooth optimization problem,

$$\min_{x \in \mathbb{R}^d} \; f(x),$$

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

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

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

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

Linear Convergence under the PL Inequality

- Proof is simple (simpler than with strong-convexity).

Linear Convergence under the PL Inequality

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

Linear Convergence under the PL Inequality

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

Weaker Conditions than Strong Convexity

- How does the PL inequality [1963] relate to more recent conditions?

Weaker Conditions than Strong Convexity

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

Weaker Conditions than Strong Convexity

- How does the PL inequality [1963] relate to more recent conditions?
  - EB: error bounds [Luo & Tseng, 1993].
  - QG: quadratic growth [Anitescu, 2000].
    - QG $+$ convexity is “optimal strong convexity” [Liu & Wright, 2015].

- How does the PL inequality [1963] relate to more recent conditions?
  - EB: error bounds [Luo & Tseng, 1993].
  - QG: quadratic growth [Anitescu, 2000].
    - QG $+$ convexity is “optimal strong convexity” [Liu & Wright, 2015].
  - ESC: essential strong convexity [Liu et al., 2013].

Weaker Conditions than Strong Convexity

- How does the PL inequality [1963] relate to more recent conditions?
  - EB: error bounds [Luo & Tseng, 1993].
  - QG: quadratic growth [Anitescu, 2000].
    - QG + convexity is “optimal strong convexity” [Liu & Wright, 2015].
  - ESC: essential strong convexity [Liu et al., 2013].
  - RSI: restricted secant inequality [Zhang & Yin, 2013].
    - RSI + convexity is “restricted strong convexity”.

Weaker Conditions than Strong Convexity

- How does the PL inequality [1963] relate to more recent conditions?
  - EB: error bounds [Luo & Tseng, 1993].
  - QG: quadratic growth [Anitescu, 2000].
    - QG $+$ convexity is "optimal strong convexity" [Liu & Wright, 2015].
  - ESC: essential strong convexity [Liu et al., 2013].
  - RSI: restricted secant inequality [Zhang & Yin, 2013].
    - RSI $+$ convexity is "restricted strong convexity".
  - WSC: weak strong convexity [Necoara et al., 2015].
    - Also sometimes used for QG $+$ convexity.

Weaker Conditions than Strong Convexity

- How does the PL inequality [1963] relate to more recent conditions?
  - EB: error bounds [Luo & Tseng, 1993].
  - QG: quadratic growth [Anitescu, 2000].
    - QG $+$ convexity is “optimal strong convexity” [Liu & Wright, 2015].
  - ESC: essential strong convexity [Liu et al., 2013].
  - RSI: restricted secant inequality [Zhang & Yin, 2013].
    - RSI $+$ convexity is “restricted strong convexity”.
  - WSC: weak strong convexity [Necoara et al., 2015].
    - Also sometimes used for QG $+$ convexity.
- Proofs are more complicated under these conditions.
- Are they more general?

Relationships Between Conditions

Theorem

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

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

Relationships Between Conditions

Theorem

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

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

If we further assume that f is convex, then

$$\text{(RSI)} \equiv \text{(EB)} \equiv \text{(PL)} \equiv \text{(QG)}.$$

Relationships Between Conditions

Theorem

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

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

If we further assume that f is convex, then

$$\text{(RSI)} \equiv \text{(EB)} \equiv \text{(PL)} \equiv \text{(QG)}.$$

- QG is the weakest condition but allows non-global local minima.

Relationships Between Conditions

Theorem

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

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

If we further assume that f is convex, then

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

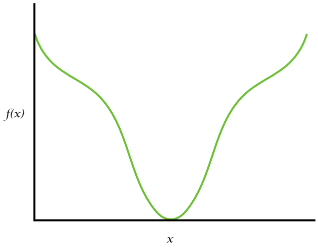
- QG is the weakest condition but allows non-global local minima.
- PL $\equiv$ EB are most general conditions.
  - Allow linear convergence to global minimizer.

PL Inequality for Invex and Non-Convex Functions

- While PL inequality does not imply convexity, it implies invexity.

- While PL inequality does not imply convexity, it implies invexity.
  - For smooth f, invexity $\iff$ all stationary points are global optimum.
  - Example of invex but non-convex function satisfying PL:

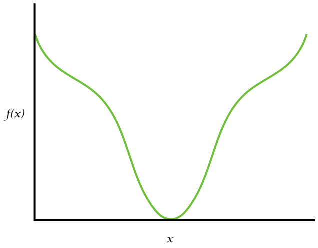
$$f(x) = x^2 + 3\sin^2(x)$$



PL Inequality for Invex and Non-Convex Functions

- While PL inequality does not imply convexity, it implies invexity.
  - For smooth f, invexity $\iff$ all stationary points are global optimum.
  - Example of invex but non-convex function satisfying PL:

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

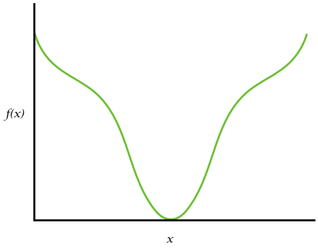


- Many important models don't satisfy invexity.

PL Inequality for Invex and Non-Convex Functions

- While PL inequality does not imply convexity, it implies invexity.
  - For smooth f, invexity $\iff$ all stationary points are global optimum.
  - Example of invex but non-convex function satisfying PL:

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

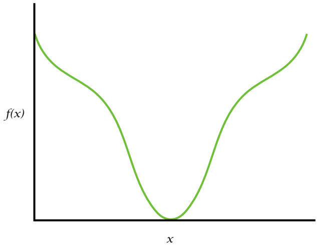


- Many important models don't satisfy invexity.
- For these problems we often divide analysis into two phases:
  - Global convergence: iterations needed to get "close" to minimizer.
  - Local convergence: how fast does it converge near the minimizer?

PL Inequality for Invex and Non-Convex Functions

- While PL inequality does not imply convexity, it implies invexity.
  - For smooth f, invexity $\iff$ all stationary points are global optimum.
  - Example of invex but non-convex function satisfying PL:

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



- 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 strong-convexity near minimizer.
  - If we assume PL, then local convergence phase may be much earlier.

Convergence for Huge-Scale Methods

- For large datasets, we typically don't use GD.
  - But the PL inequality can be used to analyze other algorithms.

- For large datasets, we typically don't use GD.
  - But the PL inequality can be used to analyze other algorithms.
- We will use PL for coordinate descent and stochastic gradient.
  - Garber & Hazan [2015] consider Franke-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

$$\mathbb{E}\left[f(x^k) - f^*\right] \leq \left(1 - \frac{\mu}{dL_c}\right)^k \left[f(x^0) - f^*\right],$$

  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

$$\mathbb{E}\left[f(x^k) - f^*\right] \leq \left(1 - \frac{\mu}{dL_c}\right)^k \left[f(x^0) - f^*\right],$$

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 a faster rate

$$f(x^k) - f^* \leq \left(1 - \frac{\mu_1}{L_c}\right)^k \left[f(x^0) - f^*\right],$$

- For randomized coordinate descent under PL we have

$$\mathbb{E}\left[f(x^k) - f^*\right] \leq \left(1 - \frac{\mu}{dL_c}\right)^k \left[f(x^0) - f^*\right],$$

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 a faster rate

$$f(x^k) - f^* \leq \left(1 - \frac{\mu_1}{L_c}\right)^k \left[f(x^0) - f^*\right],$$

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

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

Random and Greedy Coordinate Descent

- For randomized coordinate descent under PL we have

$$\mathbb{E}\left[f(x^k) - f^*\right] \leq \left(1 - \frac{\mu}{dL_c}\right)^k \left[f(x^0) - f^*\right],$$

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 a faster rate

$$f(x^k) - f^* \leq \left(1 - \frac{\mu_1}{L_c}\right)^k \left[f(x^0) - f^*\right],$$

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

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

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

- Stochastic gradient (SG) methods apply to general problems

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

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

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

- Stochastic gradient (SG) methods apply to general problems

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

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

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

- 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.

Theorem

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

$$\mathbb{E}\left[f(x^k) - f^*\right] \le \frac{L\sigma^2}{2k\mu^2},$$

Theorem

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

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

while with α_k set to constant α we have

$$\mathbb{E}\left[f(x^k) - f^*\right] \leq (1 - 2\mu\alpha)^k \left[f(x^0) - f^*\right] + \frac{L\sigma^2\alpha}{4\mu}.$$

Stochastic Gradient Methods

Theorem

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

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

while with α_k set to constant α we have

$$\mathbb{E}\left[f(x^k) - f^*\right] \leq (1 - 2\mu\alpha)^k \left[f(x^0) - f^*\right] + \frac{L\sigma^2\alpha}{4\mu}.$$

- $O(1/k)$ rate without strong-convexity (or even convexity).
- Fast reduction of sub-optimality under small constant step-size.

Stochastic Gradient Methods

Theorem

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

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

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

$$\mathbb{E}\left[f(x^k) - f^*\right] \leq (1 - 2\mu\alpha)^k \left[f(x^0) - f^*\right] + \frac{L\sigma^2\alpha}{4\mu}.$$

- $O(1/k)$ 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.
- Attouch & 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.
- Attouch & 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 a different PL generalization giving a simple global rate.

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 and g is a potentially non-smooth convex function.

  - E.g., ℓ_1-regularization, bound constraints, etc.

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 and g is a potentially non-smooth convex function.
  - E.g., ℓ_1-regularization, bound constraints, etc.
- We say that F satisfies the proximal-PL inequality if

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

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 and g is a potentially non-smooth convex function.
  - E.g., ℓ_1-regularization, bound constraints, etc.
- We say that F satisfies the proximal-PL inequality if

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

  where

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

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 and g is a potentially non-smooth convex function.

  - E.g., ℓ_1-regularization, bound constraints, etc.

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

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

where

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

- Condition yields extremely-simple proof:

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

- 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 strongly-convex.
  - f satisfies PL and g is constant.
  - $f = h(Ax)$ for strongly-convex h and g is indicator of polyhedral set.
  - F is convex and satisfies QG.
  - F satisfies the proximal-EB condition or KL inequality

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 strongly-convex.
  - f satisfies PL and g is constant.
  - $f = h(Ax)$ for strongly-convex h and g is indicator of polyhedral set.
  - F is convex and satisfies QG.
  - F satisfies the proximal-EB condition or KL inequality
- Includes dual support vector machines (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 strongly-convex.
  - f satisfies PL and g is constant.
  - $f = h(Ax)$ for strongly-convex h and g is indicator of polyhedral set.
  - F is convex and satisfies QG.
  - F satisfies the proximal-EB condition or KL inequality
- Includes dual support vector machines (SVM) problem:
  - Implies linear rate of SDCA for SVMs.
- Includes ℓ_1-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 gradient descent.
  - Gives trivial proof and is weaker than more recent conditions.
  - Weakest condition that guarantees global minima.

Summary

- In 1963, Polyak proposed a condition for linear rate of gradient descent.
  - Gives trivial proof and is weaker than more recent conditions.
  - Weakest condition that guarantees global minima.
- 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 gradient descent.
  - Gives trivial proof and is weaker than more recent conditions.
  - Weakest condition that guarantees global minima.
- 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.

Thank you!