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

Hamed Karimi (UBC, 1QBit), Julie Nutini (UBC) and Mark Schmidt (UBC)

OVERVIEW: Linear Convergence Without Strong-Convexity

- Fitting most machine learning models involves some sort of optimization problem.
  - Most common methods in ML are gradient descent and variants: coordinate descent, stochastic gradient.
- Well-known for these methods,

Smoothness + Strong-Convexity ⇒ Linear Convergence

  - However, many objectives of ML problems are not strongly-convex.
- Motivated alternative conditions for linear convergence:
  - Error bounds (EB) [Luo & Tseng, 1993]:

$$\|\nabla f(x)\| \geq \mu \|x_p - x\| \quad \forall\, x$$

  - Essential strong-convexity (ESC) [Liu et al., 2014]:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{\mu}{2}\|y - x\|^2 \quad \forall\, x, y \text{ such that } x_p = y_p$$

  - Weak strong-convexity (WSC) [Necoara et al., 2015]:

$$f^* \geq f(x) + \langle \nabla f(x), x_p - x\rangle + \tfrac{\mu}{2}\|x_p - x\|^2 \quad \forall\, x$$

  - Restricted secant inequality (RSI) [Zhang & Yin, 2013]:

$$\langle \nabla f(x), x - x_p\rangle \geq \mu \|x_p - x\|^2 \quad \forall\, x$$

  - Quadratic growth (QG) [Anitescu, 2000]:

$$f(x) - f^* \geq \tfrac{\mu}{2}\|x_p - x\|^2 \quad \forall\, x$$

⋆ In this work, we consider the Polyak-Łojasiewicz (PL) condition:

Smoothness + PL Condition ~~Strong-Convexity~~ ⇒ Linear Convergence

  - Simple proof of linear convergence.
  - For convex functions, equivalent to several of the above conditions.
  - For non-convex functions, weakest assumption while still guaranteeing global minimizer.

⋆ We generalize the PL condition to analyze proximal-gradient methods.

⋆ We give simple new analyses in a variety of settings:
  - Least-squares and logistic regression.
  - Randomized coordinate descent.
  - Greedy coordinate descent and variants of boosting.
  - Stochastic gradient (diminishing or constant step-size).
  - Stochastic variance-reduced gradient (SVRG).
  - Proximal-gradient and LASSO.
  - Coordinate minimization with separable non-smooth term (bound constraints or L1-regularization).
  - Linear convergence rate of training SVMs with SDCA.

PL-Inequality and Linear Convergence

- We first consider the optimization problem

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

where ∇f is Lipschitz-continuous

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n.$$

- A function f satisfies the Polyak-Łojasiewicz (PL) condition if

$$\tfrac{1}{2}\|\nabla f(x)\|^2 \geq \mu\,(f(x) - f^*), \quad \forall\, x \in \mathbb{R}^n.$$

- For gradient descent with constant step size,

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

these assumptions give a simple proof of linear convergence,

$$\begin{aligned}
f(x^{k+1}) - f^* &\leq f(x^k) - f^* + \nabla f(x^k)(x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 && \text{(Lipschitz } \nabla f\text{)}\\
&= f(x^k) - f^* + \frac{1}{2L}\|\nabla f(x^k)\|^2 && \text{(Definition of } x^{k+1}\text{)}\\
&\leq f(x^k) - f^* - \frac{\mu}{L}\left(f(x^k) - f^*\right) && \text{(PL condition of } f\text{)}\\
&= \left(1 - \frac{\mu}{L}\right)\left(f(x^k) - f^*\right).
\end{aligned}$$

Convergence of Huge-Scale Methods

- Randomized coordinate descent under PL conditions satisfies

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

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

- Greedy coordinate descent under the ∞-norm PL condition satisfies

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

giving new rates for variants of boosting.

- Stochastic gradient with decreasing step-size $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ satisfies

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

and stochastic gradient with constant step-size α has a linear rate plus error,

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

- We give a rate for stochastic variance-reduced gradient (SVRG) for finite sums,

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

Comparison to Other Conditions for Obtaining Linear Convergence

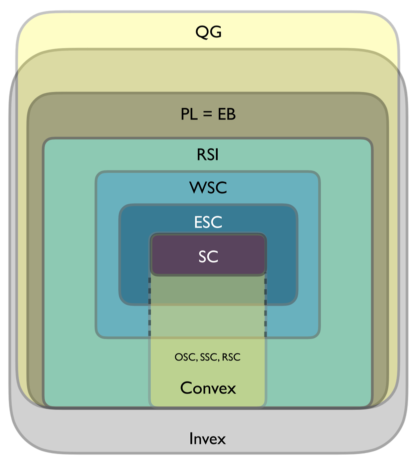
Theorem

For a function f with a Lipschitz-continuous gradient, the following implications hold:

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

If we further assume that f is convex then we have

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



- QG is weakest but does not imply invexity (allows non-global local minima).
- PL ≡ EB are most general conditions that allow linear convergence to global minimizer.

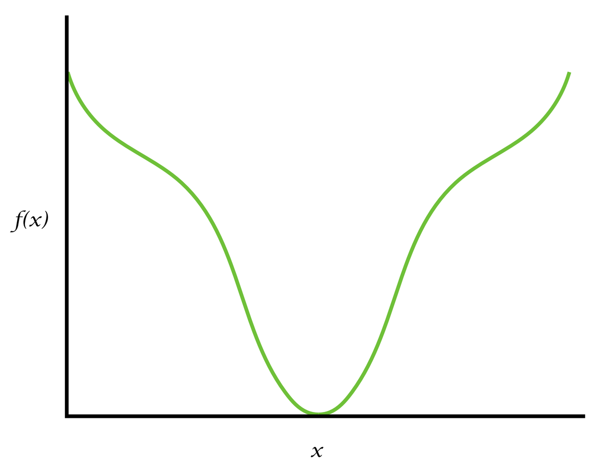
Functions Satisfying the PL Condition

- **Strongly-convex functions**:
  - By minimizing both sides of the strong-convexity condition,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2,$$

we obtain the PL-inequality with the same constant μ.

- $f(x) = g(Ax)$ **for strongly convex** g:
  - Satisfies the PL condition by the Hoffman [1952] bound.
  - Includes least-squares with singular matrix.
- $f(x) = g(Ax)$ **for strictly-convex** g:
  - Satisfies the PL condition on bounded sets.
  - Includes logistic regression when iterations/solutions are finite.



- Some non-convex functions also satisfy the inequality:
  - Figure: $f(x) = x^2 + 3\sin^2(x)$ has $L = 8$ and $\mu = 1/32$ even though $f''(x)$ can be negative.

→ For general non-convex problems, implies radius of linear convergence is larger than with SC.

Proximal-Gradient Generalization

- Consider the more general problem,

$$\min_x F(x) \equiv f(x) + g(x),$$

where f has an L-Lipschitz gradient and g is a simple non-smooth convex function.
  - E.g., bound constraints, probability simplex constraints, L1-regularization.
- For this setting, we introduce the proximal-PL condition,

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

where

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

- For proximal-gradient with constant step-size $1/L$,

$$x^{k+1} = \operatorname*{argmin}_y \left[\langle \nabla f(x^k), y - x^k\rangle + \frac{L}{2}\|y - x^k\|^2 + g(y) - g(x^k)\right],$$

assuming L-Lipschitz continuity and proximal-PL, we can *easily* prove linear convergence,

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

- This proximal-gradient proof is much simpler than previous analyses.

Functions Satisfying the Proximal-PL Inequality

The proximal-PL inequality is satisfied if:

1. f satisfies the PL inequality and g is constant.
2. f is strongly convex.
3. f has the form $f(x) = h(Ax)$ for a strongly convex function h and a matrix A, while g is an indicator function for a polyhedral set.
4. F is convex and satisfies the QG property.
   - Implies linear convergence for $L1$-regularized least squares (LASSO) and the SVM dual.

Proximal Coordinate Minimization and Support Vector Machines

- If we do proximal coordinate descent/minimization, then we have

$$\mathbb{E}\left[F(x^k) - F^*\right] \leq \left(1 - \frac{\mu}{Ln}\right)\left[F(x^0) - F^*\right].$$

- Implies linear convergence of shooting algorithm for general LASSO problems.
- Another important example is support vector machines,

$$\operatorname*{argmin}_{x \in \mathbb{R}^d} \frac{\lambda}{2}x^T x + \sum_i^l \max(0, 1 - y_i w^T x_i),$$

whose associated dual problem is

$$\min_{y_i \in [0, U]} \frac{1}{2} y^T M y - \sum_i y_i.$$

- Dual problem satisfies QG property.
  - Obtain linear convergence rate on primal by showing SDCA has global linear convergence rate on dual.