Polyak-Łojasiewicz inequality

Andersen Ang

Department of Combinatorics and Optimization, University of Waterloo, Canada

ms$\boldsymbol{x}$ang@uwaterloo.ca, where $\boldsymbol{x} = \lfloor \pi \rfloor$ Homepage: angms.science

First draft: November 3, 2020 Last update: September 7, 2022

Table of Contents

Usage of PŁ: (shorter) proof of convergence of gradient descent

- Set up
  - Problem: unconstrained minimization $\quad (\mathcal{P}) \; : \; \underset{\boldsymbol{x} \in \mathbb{R}^d}{\operatorname{argmin}} f(\boldsymbol{x}).$
  - Assumptions
    - f is L-smooth: f has L-Lipschitz gradient
    - $\varnothing \neq \mathcal{X}^* := \operatorname{argmin} f$
    - f is PŁ
  - We solve $(\mathcal{P})$ using gradient descent with constant stepsize $\frac{1}{L}$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{L} \nabla f(\boldsymbol{x}_k). \tag{GD}$$

- We can use PŁ to show GD has a linear convergence rate as

$$f(\boldsymbol{x}_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(\boldsymbol{x}_0) - f^*\right)$$

  where $f^* := f(\boldsymbol{x}^*)$.

- Important: we didn't assume f is convex.

Remarks on the setup

- f has L-Lipschitz gradient means
  - ∇f exists everywhere and it is (globally) L-Lipschitz,
  - equivalently, for all $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$,

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

  - if f is twice-differentiable, then $\lambda_{\nabla^2 f(\boldsymbol{x})} \leq L$, i.e., the eigenvalues of Hessian matrix at $\boldsymbol{x}$ are all upper bounded by L.

  See here for more information.

- $\varnothing \neq \mathcal{X}^* := \operatorname{argmin} f$ means the set $\mathcal{X}^*$, defined as the solution set of $(\mathcal{P})$, is non-empty. It means that there exists (at least one) minimizer $\boldsymbol{x}^*$

- Generally $f^* < +\infty$.

- GD with general stepsize $\alpha_k > 0$ is $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$.

Polyak-Łojasiewicz inequality

- A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies Polyak-Łojasiewicz (PŁ) inequality if there exists a scalar $\mu > 0$ such that

$$\frac{1}{2}\|\nabla f(\boldsymbol{x})\|^2 \geq \mu\Big(f(\boldsymbol{x}) - f^*\Big) \qquad \forall \boldsymbol{x} \in \operatorname{dom} f, \tag{PŁ}$$

  where $f^* := f(\boldsymbol{x}^*)$ and $\boldsymbol{x}^* \in \mathcal{X}^*$ is a minimizer of f.

- It links the norm of the gradient $\|\nabla f\|_2$, the measure of how close is $\boldsymbol{x}$ to a stationary point, to $f(\boldsymbol{x}) - f^*$, the measure of how close f at $\boldsymbol{x}$ to the optimal value f^*.

- The scaling factor μ is called PŁ constant.

- If f is σ-strongly convex, the f is σ-PŁ. We will prove this later.

PŁ implies all stationary points are global minimizers

- Since $f^* := \inf f$ is the smallest (global) achievable function value, thus
$$\frac{1}{2}\|\nabla f(\boldsymbol{x})\|^2 \geq \mu\Big(f(\boldsymbol{x}) - f^*\Big) \geq 0.$$

- At a point $\boldsymbol{x}$ that $\nabla f(\boldsymbol{x}) = \mathbf{0}$, we have
$$0 = \frac{1}{2}\|\nabla f(\boldsymbol{x})\|^2 \geq \mu\Big(f(\boldsymbol{x}) - f^*\Big) \geq 0.$$
By squeezing theorem we have $f(\boldsymbol{x}) = f^*$, meaning that such $\boldsymbol{x}$ is a global minimizer.

- The statement "$\|\nabla f(\boldsymbol{x})\|_2 = 0 \implies \boldsymbol{x}$ is a global minimizer" is the classical 1st-order optimality condition in convex smooth optimization. Note that here in PŁ we didn't assume f is convex.

- In fact, PŁ is related to *invex function*: a function is invex if and only if every stationary point is a global minimum.

What functions are PŁ?

- Given a function f, how do we know is f satisfies PŁ?
- Determine "is f PŁ" for *a very general* class of function f is an open problem.
- We are doing optimization so we only focus on functions we deal with most of the time.
- In optimization, we have a nice sufficient condition.

  If f is σ-strongly convex, then f is σ-PŁ.

  We prove this now.

σ-strongly convex functions are σ-PŁ

- Let $\sigma > 0$. If f is σ-strongly convex , then for all $\boldsymbol{x}, \boldsymbol{y}$,
$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\sigma}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{SC}$$
Geometrically, it means f is not "too flat": it is bounded below by a quadratic function. See here for more discussion.

- Now we show SC implies KŁ. Recall that in KŁ we have f^*, so we need to create f^* in SC. This can be done by just taking $\min_{\boldsymbol{y}}$ on both sides of SC:
$$\min_{\boldsymbol{y}} \Big\{ f(\boldsymbol{y}) \Big\} \overset{\text{SC}}{\geq} \min_{\boldsymbol{y}} \Big\{ f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\sigma}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \Big\},$$
which gives
$$f^* \geq f(\boldsymbol{x}) - \frac{1}{2\sigma} \|\nabla f(\boldsymbol{x})\|_2^2,$$
i.e.,
$$\frac{1}{2} \|\nabla f(\boldsymbol{x})\|_2^2 \geq \sigma \Big(f(\boldsymbol{x}) - f^* \Big). \tag{PŁ}$$

Table of Contents

Polyak 1963's short proof of linear convergence of GD

$$
\begin{aligned}
f(\boldsymbol{y}) &\le f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 && f \text{ is } L\text{-smooth} \\
f(\boldsymbol{x}_{k+1}) &\le f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 && \text{put } \boldsymbol{y} = \boldsymbol{x}_{k+1},\ \boldsymbol{x} = \boldsymbol{x}_k \\
f(\boldsymbol{x}_{k+1}) &\le f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 && \text{GD } \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k) \\
f(\boldsymbol{x}_{k+1}) &\le f(\boldsymbol{x}_k) - \frac{\mu}{L}\big(f(\boldsymbol{x}_k) - f^*\big) && \text{PŁ} \\
f(\boldsymbol{x}_{k+1}) - f^* &\le f(\boldsymbol{x}_k) - \frac{\mu}{L}\big(f(\boldsymbol{x}_k) - f^*\big) - f^* && \text{subtract both side by } f^* \\
&= \big(1 - \frac{\mu}{L}\big)\big(f(\boldsymbol{x}_k) - f^*\big). \\
f(\boldsymbol{x}_{k+1}) - f^* &\le \big(1 - \frac{\mu}{L}\big)^k\big(f(\boldsymbol{x}_0) - f^*\big) && \text{recursion}
\end{aligned}
$$

Comments

- The proof also applies to optimal stepsize, since

$$f(\boldsymbol{x}_{k+1}) = \min_{\alpha} f\Big(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)\Big) \leq f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big),$$

  where the $\leq$ is by definition of the optimal stepsize.

- PŁ does not
  - assume f is convex.
  - assume the minimizer $\boldsymbol{x}^*$ is unique.

  In contrast, strong convexity (SC) assumes f
  - is convex
  - is strongly convex, which implies strictly convex and thus implies the minimizer is unique

- SC $\implies$ PŁ (we just proved it), so the same convergence rate holds if f is μ-SC. However, proving such convergence rate using SC is tedious. See the long proof here.

Prove convergence of randomized coordinate descent (rCD) by PŁ

- Set up
  - Same problem: $(\mathcal{P}) \ : \ \underset{\boldsymbol{x}\in\mathbb{R}^d}{\operatorname{argmin}} f(\boldsymbol{x})$.
  - Assumptions
    - f is coordinate-wise L-smooth

$$f(\underbrace{\boldsymbol{x}+\alpha\boldsymbol{e}_i}_{\boldsymbol{y}}) \le f(\boldsymbol{x}) + \langle \nabla_i f(\boldsymbol{x})\boldsymbol{e}_i, \underbrace{\alpha\boldsymbol{e}_i}_{\boldsymbol{y}-\boldsymbol{x}}\rangle + \frac{L}{2}\|\underbrace{\alpha\boldsymbol{e}_i}_{\boldsymbol{y}-\boldsymbol{x}}\|^2$$

    - $\varnothing \neq \mathcal{X}^* := \operatorname{argmin} f$ and f is PŁ
  - rCD with constant stepsize $\frac{1}{L}$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{L}\nabla_{i_k} f(\boldsymbol{x}_k)\boldsymbol{e}_{i_k}$$

    picking coordinate index i_k is based on uniform random probability.

- We can use PŁ to show rCD has linear convergence rate in expectation as

$$\mathbb{E}\Big(f(\boldsymbol{x}_{k+1}) - f^*\Big) \le \Big(1 - \frac{\mu}{dL}\Big)^k \big(f(\boldsymbol{x}_0) - f^*\big).$$

Short proof

$$
\begin{aligned}
f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \alpha \nabla_i f(\boldsymbol{x}_k) + \frac{L}{2}\alpha^2 && f \text{ is coordinate-wise } L\text{-smooth} \\
f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - \frac{1}{2L}|\nabla_i f(\boldsymbol{x}_k)|^2 && \text{rCD update } \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \underbrace{\frac{1}{L}\nabla_{i_k} f(\boldsymbol{x}_k)}_{\alpha}\boldsymbol{e}_{i_k} \\
\mathbb{E} f(\boldsymbol{x}_{k+1}) &\leq \mathbb{E} f(\boldsymbol{x}_k) - \mathbb{E}\tfrac{1}{2L}|\nabla_i f(\boldsymbol{x}_k)|^2 && \text{take expectation} \\
&= f(\boldsymbol{x}_k) - \frac{1}{2L}\mathbb{E}|\nabla_i f(\boldsymbol{x}_k)|^2 && \text{expectation is a linear operator} \\
&= f(\boldsymbol{x}_k) - \frac{1}{2L}\sum_i \frac{1}{d}|\nabla_i f(\boldsymbol{x}_k)|^2 && \text{uniform probability} \\
&= f(\boldsymbol{x}_k) - \frac{1}{2dL}\|\nabla f(\boldsymbol{x}_k)\|^2 \\
\mathbb{E} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - \frac{\mu}{dL}\big(f(\boldsymbol{x}_k) - f^*\big) && -\frac{1}{2}\|\nabla f(\boldsymbol{x}_k)\|^2 \overset{\text{PŁ}}{\leq} -\mu\big(f(\boldsymbol{x}_k) - f^*\big)
\end{aligned}
$$

Then similar to GD: subtract both side by f^*, rearrange and perform recursion will finish the proof.

Table of Contents

Last page - summary

Discussed

- Polyak-Łojasiewicz inequality and its applications.

Not discussed

- Proximal version of PŁ: see here
- Relationship between PŁ and the more general Kurdyka-Łojasiewicz inequality.

Reference

- Hamed Karimi, Julie Nutini, Mark Schmidt, "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Lojasiewicz Condition".

End of document