Polyak-Lojasiewicz inequality

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft: November 3, 2020
Last update: November 4, 2020

Using PL to prove convergence of gradient descent

- Set up
  - Problem: unconstrained minimization

$$(\mathcal{P}) \ : \ \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}).$$

  - Assumptions on the problem:
    - f has L-Lipschitz gradient
    - $\mathcal{X}^* \neq \varnothing$
    - f is PL
  - Gradient descent with constant stepsize $\frac{1}{L}$

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

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

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

Remarks on the setup

- f has L-Lipschitz gradient means

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

  and if f is twice-differentiable, then $\lambda(\nabla^2 f(\mathbf{x})) \leq L$, i.e., the eigenvalues of Hessian matrix at $\mathbf{x}$ are all upper bounded by L. See here for more information.

- $\mathcal{X}^* \neq \varnothing$ means the solution set of $(\mathcal{P})$ is non-empty. It means that there exists at least one minimizer $\mathbf{x}^*$, and $f^* = f(\mathbf{x}^*) < +\infty$.

- Gradient descent with general stepsize $\alpha_k > 0$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k).$$

Polyak-Lojasiewicz inequality

- A function $f : \mathbb{R}^n \to \mathbb{R}$ that is differentiable (i.e. $\nabla f(\mathbf{x})$ exists for $\mathbf{x} \in \operatorname{dom} f$) satisfies Polyak-Lojasiewicz (PL) inequality if there exists a positive scalar $\mu > 0$ such that

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

  for all $\mathbf{x} \in \operatorname{dom} f$, where $f^* = f(\mathbf{x}^*)$ and $\mathbf{x}^*$ is a minimizer of f.

- It means the gradient grows faster than a quadratic function (scaled by a scalar $\mu > 0$) as we move $\mathbf{x}$ away from $\mathbf{x}^*$.

- Side note: the L should be the Polish L (L with Stroke).

Polyak 1963's short proof of linear convergence of GD

$$\begin{array}{rrcl} f \text{ has } L\text{-Lips. grad} & f(\mathbf{y}) & \leq & f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \\ \text{GD update} & \mathbf{x}_{k+1} & = & \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k) \end{array}$$

- Put $\mathbf{y} = \mathbf{x}_{k+1}$, $\mathbf{x} = \mathbf{x}_k$ in the first inequality, then plug in the second equation gives

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2.$$

- PL inequality: $-\frac{1}{2}\|\nabla f(\mathbf{x}_k)\|^2 \leq -\mu\big(f(\mathbf{x}_k) - f^*\big)$, so

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\mu}{L}\big(f(\mathbf{x}_k) - f^*\big).$$

- Subtract both side by f^*

$$f(\mathbf{x}_{k+1}) - f^* \leq f(\mathbf{x}_k) - \frac{\mu}{L}\big(f(\mathbf{x}_k) - f^*\big) - f^* = \big(1 - \frac{\mu}{L}\big)\big(f(\mathbf{x}_k) - f^*\big).$$

- Recursion:

$$f(\mathbf{x}_{k+1}) - f^* \leq \big(1 - \frac{\mu}{L}\big)^k \big(f(\mathbf{x}_0) - f^*\big).$$

Comments

- The proof also applies to optimal stepsize, since

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

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

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

  In contrast, strong convexity (SC) assumes f is convex and yields the minimizer is unique.

- If f is μ-SC, then it is μ-PL with the same μ. That is, SC implies PL.

Using PL to prove convergence of randomized coordinate descent (rCD)

- Set up
  - Same problem: $(\mathcal{P}) \ : \ \underset{\mathbf{x}\in\mathbb{R}^d}{\operatorname{argmin}} \ f(\mathbf{x})$.
  - Assumptions on the problem:
    - f has coordinate-wise L-Lipschitz gradient

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

    - $\mathcal{X}^* \neq \varnothing$
    - f is PL
  - rCD: Coordinate-wise gradient descent with constant stepsize $\frac{1}{L}$

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

    with uniform random rule on picking i_k,

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

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

Short proof of linear convergence of rCD

$$
\begin{array}{rrcl}
f \text{ has coor.-wise } L\text{-Lips. grad} & f(\mathbf{x}_{k+1}) & \leq & f(\mathbf{x}_k) + \alpha \nabla_i f(\mathbf{x}_k) + \frac{L}{2}\alpha^2 \\
\text{rCD update} & \mathbf{x}_{k+1} & = & \mathbf{x}_k - \underbrace{\frac{1}{L}\nabla_{i_k} f(\mathbf{x}_k)}_{\alpha} \mathbf{e}_{i_k}
\end{array}
$$

- Put α in the second equation to the first inequality gives

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2L}|\nabla_i f(\mathbf{x}_k)|^2.$$

- Take expectation

$$
\begin{array}{rcl}
\mathbb{E} f(\mathbf{x}_{k+1}) & \leq & \mathbb{E} f(\mathbf{x}_k) - \mathbb{E}\frac{1}{2L}|\nabla_i f(\mathbf{x}_k)|^2 \\
& = & f(\mathbf{x}_k) - \frac{1}{2L}\mathbb{E}|\nabla_i f(\mathbf{x}_k)|^2 \\
& = & f(\mathbf{x}_k) - \frac{1}{2L}\sum_i \frac{1}{d}|\nabla_i f(\mathbf{x}_k)|^2 \\
& = & f(\mathbf{x}_k) - \frac{1}{2dL}\|\nabla f(\mathbf{x}_k)\|^2
\end{array}
$$

- Apply PL inequality

$$\mathbb{E} f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\mu}{dL}\big(f(\mathbf{x}_k) - f^*\big).$$

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

Last page - summary

- Polyak-Lojasiewicz inequality and its applications.
- Not discussed: PL inequality also works for proximal-gradient method.
- Not discussed: the relationship between PL inequality and the (more general) Kurdyka-Lojasiewicz inequality.
- Reference: Hamed Karimi, Julie Nutini, Mark Schmidt, "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Lojasiewicz Condition", arXiv:1608.04636

End of document