Convergence of proximal gradient method based on proximal Polyak-Lojasiewicz Inequality

Andersen Ang

Department of Combinatorics and Optimization,
University of Waterloo, Waterloo, Canada

msxang@uwaterloo.ca, where $\mathbf{x} = \lfloor \pi \rfloor$
Homepage: angms.science

First draft: November 23, 2021 Last update: November 27, 2021

Problem setup

- Consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}) \qquad (\mathcal{P})$$

  - $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth
  - $g : \mathbb{R}^n \to \bar{\mathbb{R}}$ is convex, possibly nonsmooth and proximable[1]

- We consider solving $(\mathcal{P})$ using proximal gradient: starting with $\mathbf{x}_0$, iterate

$$\mathbf{x}_{k+1} = \operatorname{prox}_{\frac{1}{L}g}\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big),$$

  where the proximal operator $\operatorname{prox} : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$\operatorname{prox}_{\gamma g}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{z}} \frac{1}{2}\|\mathbf{z} - \mathbf{w}\|_2^2 + \gamma g(\mathbf{z})$$

- Note that in general we have $\in$ instead of $=$ in the proximal step, here it is the convexity of g leading to $=$.

[1] The proximal operator prox_g is easy and/or cheap to compute.

Equivalent form of the proximal gradient step

- Using $\text{prox}_{\gamma g}(\mathbf{w}) = \underset{\mathbf{z}}{\text{argmin}} \frac{1}{2}\|\mathbf{z}-\mathbf{w}\|_2^2 + \gamma g(\mathbf{z})$, the proximal gradient step becomes

$$\begin{aligned}
\mathbf{x}_{k+1} &= \text{prox}_{\frac{1}{L}g}\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big) \\
&= \underset{\mathbf{z}}{\text{argmin}} \frac{1}{2}\left\|\mathbf{z} - \Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big)\right\|_2^2 + \tfrac{1}{L}g(\mathbf{z}) \\
&= \underset{\mathbf{z}}{\text{argmin}} \frac{1}{2}\left\|(\mathbf{z}-\mathbf{x}_k) + \frac{1}{L}\nabla f(\mathbf{x}_k)\right\|_2^2 + \tfrac{1}{L}g(\mathbf{z})
\end{aligned}$$

- Expand the quadratic term gives

$$\mathbf{x}_{k+1} = \underset{\mathbf{z}}{\text{argmin}} \frac{1}{2}\|\mathbf{z}-\mathbf{x}_k\|_2^2 + \frac{1}{2}\left\|\frac{1}{L}\nabla f(\mathbf{x}_k)\right\|_2^2 + \langle \mathbf{z}-\mathbf{x}_k, \frac{1}{L}\nabla f(\mathbf{x}_k)\rangle + \frac{1}{L}g(\mathbf{z})$$

- For argmin we can multiply the right hand side with L, and then add or subtract constant terms

$$\mathbf{x}_{k+1} = \underset{\mathbf{z}}{\text{argmin}} \frac{L}{2}\|\mathbf{z}-\mathbf{x}_k\|_2^2 + \langle \mathbf{z}-\mathbf{x}_k, \nabla f(\mathbf{x}_k)\rangle + g(\mathbf{z}) - g(\mathbf{x}_k)$$

PL and generalized PL inequality

- Now the proximal gradient step is

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{z}} \frac{L}{2}\|\mathbf{z}-\mathbf{x}_k\|_2^2 + \langle \mathbf{z}-\mathbf{x}_k, \nabla f(\mathbf{x}_k)\rangle + g(\mathbf{z}) - g(\mathbf{x}_k) \qquad (*)$$

- Recall that the Polyak-Lojasiewicz inequality is

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|_2^2 \geq \mu\Big(f(\mathbf{x}) - f^*\Big), \quad \forall \mathbf{x} \qquad \text{(PL)}$$

see here for the details of PL.

- Based on $(*)$, we can try to generalize PL as

$$\frac{1}{2}\mathcal{D}_g(\mathbf{x}, L) \geq \mu\Big(f(\mathbf{x}) - f^*\Big), \quad \forall \mathbf{x} \qquad \text{(prox-PL)}$$

with

$$\mathcal{D}_g(\mathbf{x}, \alpha) = -2\alpha \min_{\mathbf{z}} \frac{\alpha}{2}\|\mathbf{z}-\mathbf{x}\|_2^2 + \langle \mathbf{z}-\mathbf{x}, \nabla f(\mathbf{x})\rangle + g(\mathbf{z}) - g(\mathbf{x}).$$

Why prox-PL is useful

- We can use prox-PL to prove the convergence of proximal gradient method, with a (sort-of) shorter proof then the one using the traditional approach. (e.g. here and here)

- We are going to prove the following theorem
  Theorem On solving $(\mathcal{P})$, assuming F satisfies prox-PL with parameter μ, then for the sequence $\{\mathbf{x}_k\}$ generated by the proximal gradient step $(*)$, we have

$$F(\mathbf{x}_k) - F^* \leq \left(1 - \frac{\mu}{L}\right)^k \big(F(\mathbf{x}_0) - F^*\big).$$

- Important note: this theorem only shows that $\{F(\mathbf{x}_k)\}$ converges to F^*, which does not say anything that $\{\mathbf{x}_k\}$ will converge to $\mathbf{x}^*$. In other words, if we want to show that $\{\mathbf{x}_k\}$ converge to $\mathbf{x}^*$, we have to prove it separately.

The proof

- Consider $F(\mathbf{x}_{k+1}) = f(\mathbf{x}_{k+1}) + g(\mathbf{x}_{k+1})$, then a tricky step

$$F(\mathbf{x}_{k+1}) = f(\mathbf{x}_{k+1}) + g(\mathbf{x}_{k+1}) + g(\mathbf{x}_k) - g(\mathbf{x}_k) = f(\mathbf{x}_{k+1}) + g(\mathbf{x}_k) + g(\mathbf{x}_{k+1}) - g(\mathbf{x}_k)$$

- By the fact that f is L-smooth:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2,$$

so we have

$$\begin{aligned} F(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 + g(\mathbf{x}_k) + g(\mathbf{x}_{k+1}) - g(\mathbf{x}_k) \\ &\leq F(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 + g(\mathbf{x}_{k+1}) - g(\mathbf{x}_k) \end{aligned}$$

- The next step is the most tricky step in the proof.

- We have

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 + g(\mathbf{x}_{k+1}) - g(\mathbf{x}_k)$$

- Using $\mathcal{D}_g(\mathbf{x}_k, L)$ and $\mathbf{x}_{k+1}$, we have

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \frac{1}{2L}\mathcal{D}_g(\mathbf{x}_k, L).$$

- Explanation: recall $\mathcal{D}_g(\mathbf{x}, \alpha) = -2\alpha \min_{\mathbf{z}} \frac{\alpha}{2}\|\mathbf{z} - \mathbf{x}_k\|_2^2 + \langle \mathbf{z} - \mathbf{x}_k, \nabla f(\mathbf{x}_k) \rangle + g(\mathbf{z}) - g(\mathbf{x}_k)$, so $-\frac{1}{2L}\mathcal{D}_g(\mathbf{x}_k, L)$ cancelled out the $-\frac{1}{2L}$ and left with

$$-\frac{1}{2L}\mathcal{D}_g(\mathbf{x}_k, L) = +\min_{\mathbf{z}} \frac{L}{2}\|\mathbf{z} - \mathbf{x}_k\|_2^2 + \langle \mathbf{z} - \mathbf{x}_k, \nabla f(\mathbf{x}_k) \rangle + g(\mathbf{z}) - g(\mathbf{x}_k),$$

which is larger than $\langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 + g(\mathbf{x}_{k+1}) - g(\mathbf{x}_k)$ because by definition $\mathbf{x}_{k+1}$ is the minimizer of this expression.

- Now we have $F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \frac{1}{2L}\mathcal{D}_g(\mathbf{x}_k, L)$. Using the prox-PL,

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

- Subtract F^* on both sides

$$F(\mathbf{x}_{k+1}) - F^* \leq F(\mathbf{x}_k) - F^* - \frac{\mu}{L}(F(\mathbf{x}) - F^*) = \left(1 - \frac{\mu}{L}\right)(F(\mathbf{x}) - F^*)$$

which by recursion gives

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

Last page - summary

- A function f satisfies the Proximal Polyak-Lojasiewicz inequality if there exists $\mu > 0$ that

$$\frac{1}{2}\mathcal{D}_g(\mathbf{x}, L) \geq \mu\Big(f(\mathbf{x}) - f^*\Big), \quad \forall \mathbf{x} \qquad \text{(prox-PL)}$$

  with

$$\mathcal{D}_g(\mathbf{x}, \alpha) = -2\alpha \min_{\mathbf{z}} \frac{\alpha}{2}\|\mathbf{z} - \mathbf{x}\|_2^2 + \langle \mathbf{z} - \mathbf{x}, \nabla f(\mathbf{x})\rangle + g(\mathbf{z}) - g(\mathbf{x}).$$

- Prox-PL can be used to prove convergence of proximal gradient method.

- Reference: Hamed Karimi, Julie Nutini, and Mark Schmid, "Linear Convergence of Gradient and Proximal-GradientMethods Under the Polyak Lojasiewicz Condition", 2016

End of document