Convergence Analysis of PGM

2022年2月25日 9:08

The basic PGM for solving

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ F(\theta) \triangleq \ell(\theta) + R(\theta) \right\} \quad -(*)$$

(ℓ: convex, $\nabla \ell$ L-Lipschitz; R: convex)

is given by

$$\theta^{k+1} = \operatorname{prox}_{\alpha_k R}\left(\theta^k - \alpha_k \nabla \ell(\theta^k)\right)$$

$$= \underset{\gamma \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2} \| \theta^k - \alpha_k \nabla \ell(\theta^k) - \gamma \|_2^2 + \alpha_k R(\gamma) \right\}.$$

Proposition 1: Suppose that $\{\alpha_k\}$ satisfy $0 < \underline{\alpha} < \alpha_k < \bar{\alpha} < \frac{1}{L}$ for some $\underline{\alpha}, \bar{\alpha}$. Assuming $(*)$ is solvable,

(a) (Sufficient decrease) There exists a constant $K_1 > 0$ s.t.

$$F(\theta^k) - F(\theta^{k+1}) \geq K_1 \| \theta^k - \theta^{k+1} \|_2^2$$

(b) (Safeguard; relative error) There exists a constant $K_2 > 0$ s.t.

$$\| E(\theta^k) \|_2 \leq K_2 \| \theta^k - \theta^{k+1} \|_2, \qquad E(\theta) = \operatorname{prox}_R(\theta - \nabla \ell(\theta)) - \theta$$

Consequences

(1) $F(\theta^k) \to \hat{v} \triangleq F(\hat{\theta})$,

(2) Every accumulation point of $\{\theta^k\}$ is optimal for $(*)$.

Goal: Determine the convergence rate of PGM.

Need a measure of the progress of the method.

Proposition 2: Let $\mathcal{H}$ be the optimal solution set of $(*)$, assumed to be non-empty (thus, convex and closed). Under the setting of Proposition 1,

(c) (Cost-to-go estimate) There exists a $K_3 > 0$ s.t.

$$F(\theta^{k+1}) - \hat{v} \le \kappa_3 \left\{ \operatorname{dist}(\theta^k, \Theta)^2 + \|\theta^{k+1} - \theta^k\|_2^2 \right\},$$

where $\operatorname{dist}(\theta, \Theta) = \inf_{\gamma \in \Theta} \|\theta - \gamma\|_2$.

Proof: Let $\bar{\theta}^k = \Pi_\Theta(\theta^k)$. By definition, $F(\theta^{k+1}) = \mathcal{L}(\theta^{k+1}) + R(\theta^{k+1})$

and since $\theta^{k+1} = \operatorname{prox}_{\alpha_k R}(\theta^k - \alpha_k \nabla \mathcal{L}(\theta^k))$,

$$\frac{1}{2}\|\theta^{k+1} - \theta^k + \alpha_k \nabla \mathcal{L}(\theta^k)\|_2^2 + \alpha_k R(\theta^{k+1})$$

$$\le \frac{1}{2}\|\bar{\theta}^k - \theta^k + \alpha_k \nabla \mathcal{L}(\theta^k)\|_2^2 + \alpha_k R(\bar{\theta}^k)$$

$$\Rightarrow R(\theta^{k+1}) - R(\bar{\theta}^k) + \nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \bar{\theta}^k) \leftarrow$$

$$\le \frac{1}{2\alpha_k}\|\bar{\theta}^k - \theta^k\|_2^2 \le \frac{1}{2\underline{\alpha}} \operatorname{dist}(\theta^k, \Theta)^2.$$

On the other hand, by the mean-value theorem, there exists $\hat{\theta}^k \in [\bar{\theta}^k, \theta^{k+1}]$ s.t.

$$\mathcal{L}(\theta^{k+1}) - \mathcal{L}(\bar{\theta}^k) = \nabla \mathcal{L}(\hat{\theta}^k)^T(\theta^{k+1} - \bar{\theta}^k)$$

θ^{k+1}, $\hat{\theta}^k$, $\bar{\theta}^k = \Pi_\Theta(\theta^k)$

It follows that

$$F(\theta^{k+1}) - \hat{v} = F(\theta^{k+1}) - F(\bar{\theta}^k)$$

$$= \mathcal{L}(\theta^{k+1}) - \mathcal{L}(\bar{\theta}^k) + R(\theta^{k+1}) - R(\bar{\theta}^k)$$

$$= \nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \bar{\theta}^k) + R(\theta^{k+1}) - R(\bar{\theta}^k) \leftarrow$$

$$+ (\nabla \mathcal{L}(\hat{\theta}^k) - \nabla \mathcal{L}(\theta^k))^T(\theta^{k+1} - \bar{\theta}^k)$$

C-S: $\le \|\nabla \mathcal{L}(\hat{\theta}^k) - \nabla \mathcal{L}(\theta^k)\|_2 \cdot \|\theta^{k+1} - \bar{\theta}^k\|_2$

$$\le \frac{1}{2\underline{\alpha}} \operatorname{dist}(\theta^k, \Theta)^2 + L \cdot \|\hat{\theta}^k - \theta^k\|_2 \cdot \|\theta^{k+1} - \bar{\theta}^k\|_2$$

Now, note that

$$\|\hat{\theta}^k - \theta^k\|_2 \le \|\theta^{k+1} - \theta^k\|_2 + \|\theta^{k+1} - \hat{\theta}^k\|_2 \quad \text{(triangle inequality)}$$

$$\leq \|\theta^{k+1} - \theta^k\|_2 + \|\theta^{k+1} - \bar{\theta}^k\|_2 ,$$

$$\|\theta^{k+1} - \bar{\theta}^k\|_2 \leq \|\theta^{k+1} - \theta^k\|_2 + \|\theta^k - \bar{\theta}^k\|_2$$

$$= \|\theta^{k+1} - \theta^k\|_2 + \operatorname{dist}(\theta^k, \Theta)$$

Hence,

$$\|\hat{\theta}^k - \theta^k\|_2 \, \|\theta^{k+1} - \bar{\theta}^k\|_2 \leq 2\left[\operatorname{dist}(\theta^k, \Theta) + \|\theta^{k+1} - \theta^k\|_2\right]^2$$

$$\leq 4\left[\operatorname{dist}(\theta^k, \Theta)^2 + \|\theta^{k+1} - \theta^k\|_2^2\right]$$

Since $\forall a, b \in \mathbb{R}$, $(a+b)^2 \leq 2(a^2+b^2)$

Observe that $\operatorname{dist}(\theta^k, \Theta)$ is not easy to estimate. To circumvent this difficulty, let us assume the following:

(A4) (Local Error Bound Condition)

For any $v \geq \hat{v}$, there exist $\mu, \varepsilon > 0$ s.t.

$$\operatorname{dist}(\theta, \Theta) \leq \mu \cdot \|E(\theta)\|_2 \qquad (E(\theta) = \operatorname{prox}_R(\theta - \nabla f(\theta)) - \theta)$$

for any $\theta \in \mathbb{R}^d$ satisfying $F(\theta) \leq v$, $\|E(\theta)\|_2 \leq \varepsilon$.

Remarks

(1) (A4) says when θ is "close" to Θ (as determined by the conditions $F(\theta) \leq v$, $\|E(\theta)\|_2 \leq \varepsilon$), the measure $\|E(\theta)\|_2$ is a good surrogate of $\operatorname{dist}(\theta, \Theta)$. However, if μ is independent of v and $\varepsilon = +\infty$, then we say that the error bound is global.

(2) (A4) is a property of the problem and is algorithm independent.

Using Propositions 1 and 2 and (A4), we can get the convergence rate of PGM. Indeed, for θ^k close to Θ in the sense of (A4), we have

$$F(\theta^{k+1}) - \hat{v} \overset{\substack{(A4)\\+\\(c)}}{\leq} K_4 \left[\|E(\theta^k)\|_2^2 + \|\theta^k - \theta^{k+1}\|_2^2 \right]$$

$$\overset{(b)}{\leq} K_5 \|\theta^k - \theta^{k+1}\|_2^2$$

$$\overset{(a)}{\leq} K_6 \left[(F(\theta^k) - \hat{v}) - (F(\theta^{k+1}) - \hat{v}) \right]$$

Hence,

$$F(\theta^{k+1}) - \hat{v} \leq \underbrace{\frac{K_6}{1+K_6}}_{<1} (F(\theta^k) - \hat{v})$$

$\Rightarrow$ $\{F(\theta^k) - \hat{v}\}$ converges to 0 geometrically

Similarly, by (a),

$$\|\theta^k - \theta^{k+1}\|_2^2 \leq \frac{1}{K_1} \left[(F(\theta^k) - \hat{v}) - \underbrace{(F(\theta^{k+1}) - \hat{v})}_{\geq 0} \right]$$

$$\leq \frac{1}{K_1} (F(\theta^k) - \hat{v})$$

Thus,

$$\text{dist}(\theta^k, \Theta) \overset{(A4)}{\leq} \mu \cdot \|E(\theta^k)\|_2 \overset{(b)}{\leq} \mu' \cdot \|\theta^k - \theta^{k+1}\|_2$$

$$\leq \mu'' \cdot \sqrt{F(\theta^k) - \hat{v}}$$

This shows $\{\theta^k\}$ approaches Θ geometrically.