COMM 616: Modern Optimization with Applications in ML and OR 2024-25 Fall

Lecture 9: Unified Convergence Analysis Framework for PGD

Instructor: Jiajin Li

1 Optimization Problem

In this class, we focus on

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) := f(\boldsymbol{x}) + g(\boldsymbol{x}). \tag{P}$$

Here, we have

(i) The function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth.

(ii) The function $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ is convex, closed, and (possibly) non-smooth.

(iii) The proximal operator $\operatorname{prox}_{tg}(\cdot)$ is easily computed. (e.g., ℓ_1, ℓ_∞ norms).

PGD is already general enough to cover widely used algorithms:

(i) Proximal Point Algorithms (PPA) when $f(\boldsymbol{x}) = 0$. Generally speaking, we focus on a pure nonsmooth convex optimization problem.

(ii) Gradient Descent (GD) when $g(\boldsymbol{x}) = 0$.

(iii) Projected Gradient Descent (PGD) when $g(\boldsymbol{x}) = \mathbb{I}_{\mathcal{X}}(\boldsymbol{x})$ where $\mathcal{X}$ is a convex and closed set.

2 Convergence Analysis

When the function is convex or strongly convex, the convergence analysis of Proximal Gradient Descent (PGD) relies on the following crucial lemma:

Lemma 1. *Suppose f is L-smooth, μ-strongly convex, and $\boldsymbol{x}^+ = \operatorname{prox}_{tg}(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x}))$. Then we have*

$$F(\boldsymbol{x}^+) \le F(\boldsymbol{z}) + L\langle \boldsymbol{x}^+ - \boldsymbol{z}, \boldsymbol{x} - \boldsymbol{x}^+\rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}^+\|^2 - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2$$

for any $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$.

Theorem 2. *Suppose that $t_k = \frac{1}{L}$ for any $k \ge 0$. Then, we have*

(i) (Convexity):

$$F(\boldsymbol{x}^K) - F^\star \le \frac{L\|x^0 - \boldsymbol{x}^\star\|^2}{2K}.$$

(ii) (μ-Strongly Convexity):

$$\|\boldsymbol{x}^K - \boldsymbol{x}^\star\|^2 \le \exp\left(-\frac{K}{\kappa}\right)\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2.$$

Remark 3. *Let κ be a constant number defined as L/μ.*

Proof. (i): Picking $\boldsymbol{z} = \boldsymbol{x}^k$ in Lemma 1 yields

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\frac{L}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2.$$

Next, we pick $\boldsymbol{z} = \boldsymbol{x}^\star$ in Lemma 1 and get

$$\begin{aligned} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^\star) &\leq -L(\boldsymbol{x}^k - \boldsymbol{x}^{k+1})^T(\boldsymbol{x}^\star - \boldsymbol{x}^{k+1}) + \frac{L}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \\ &= \frac{L}{2}(\boldsymbol{x}^k - \boldsymbol{x}^{k+1})^T(\boldsymbol{x}^k + \boldsymbol{x}^{k+1} - 2\boldsymbol{x}^\star) \\ &= \frac{L}{2}\left(\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2\right). \end{aligned}$$

Telescoping it from $k = 0$ to $k = K$, we have

$$\sum_{k=0}^{K} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^\star) \leq \frac{L}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2.$$

Moreover, as the sequence $\{F(\boldsymbol{x}^k)\}_{k\geq 0}$ is monotonically decreasing, we conclude our proof.

(ii): Similar with the proof in case (i). We have

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^\star) \leq \frac{L}{2}\left((1 - \frac{\mu}{L})\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2\right).$$

Then, we have

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2.$$

■

Without the convexity of f, the convergence analysis of PGD under the nonconvex setting differs from the case in Lemma 1. Instead, we rely on the following sufficient decrease property:

Proposition 4 (Sufficient Decrease Property)**.** *Suppose that f is L-smooth and g is convex. Let $\boldsymbol{x}^+ = \mathrm{prox}_{tg}(\boldsymbol{x} - t\nabla f(\boldsymbol{x}))$. Then, we have*

$$F(\boldsymbol{x}^+) \leq F(\boldsymbol{x}) - \left(\frac{1}{2t} - \frac{L}{2}\right)\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2.$$

Proof. By definition of the proximal operator, we have

$$\boldsymbol{x}^+ = \arg\min_{\boldsymbol{y}}\left\{\frac{1}{2t}\|\boldsymbol{y} - (\boldsymbol{x} - t\nabla f(\boldsymbol{x}))\|^2 + g(\boldsymbol{y})\right\}.$$

Since $\boldsymbol{x}^+$ is the optimal solution of the above optimization problem, we have

$$\frac{1}{2t}\|\boldsymbol{x}^+ - (\boldsymbol{x} - t\nabla f(\boldsymbol{x}))\|^2 + g(\boldsymbol{x}^+) \leq \frac{t}{2}\|\nabla f(\boldsymbol{x})\|^2 + g(\boldsymbol{x}),$$

which implies

$$\frac{1}{2t}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2 + \langle \boldsymbol{x}^+ - \boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle + g(\boldsymbol{x}^+) \leq g(\boldsymbol{x}). \tag{1}$$

Moreover, since f is L-smooth, we have

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

Combining (1) and (2) yields

$$f(\boldsymbol{x}^+) + g(\boldsymbol{x}^+) \le f(\boldsymbol{x}) + g(\boldsymbol{x}) + \left(\frac{L}{2} - \frac{1}{2t}\right) \|\boldsymbol{x}^+ - \boldsymbol{x}\|^2.$$

■

Now, we are ready to conduct the convergence analysis of PGD under the nonconvex setting. A natural task is to connect the relative change $\|\boldsymbol{x}^+ - \boldsymbol{x}\|$ with a certain optimality residual. Thus, we are able to do the average error type analysis.

Proposition 5 (Safeguard Condition)**.** *Suppose that f is L-smooth and g is convex. Let $\boldsymbol{x}^+ = \text{prox}_{tg}(\boldsymbol{x} - t\nabla f(\boldsymbol{x}))$. Then, we have*

$$\text{dist}(0, \partial F(\boldsymbol{x}^+)) \le \left(\frac{1}{t} + L\right) \|\boldsymbol{x}^+ - \boldsymbol{x}\|.$$

Proof. Recall that

$$\boldsymbol{x}^+ = \arg\min_{\boldsymbol{y}} \left\{ \frac{1}{2t} \|\boldsymbol{y} - (\boldsymbol{x} - t\nabla f(\boldsymbol{x}))\|^2 + g(\boldsymbol{y}) \right\}.$$

We drive its optimality condition as

$$\begin{aligned} 0 &\in \frac{1}{t}(\boldsymbol{x}^+ - \boldsymbol{x} + t\nabla f(\boldsymbol{x})) + \partial g(\boldsymbol{x}^+) \\ &\in \frac{1}{t}(\boldsymbol{x}^+ - \boldsymbol{x}) + (\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^+)) + \nabla f(\boldsymbol{x}^+) + \partial g(\boldsymbol{x}^+). \end{aligned}$$

Then we have,

$$\text{dist}(0, \partial F(\boldsymbol{x}^+)) \le \frac{1}{t} \|\boldsymbol{x}^+ - \boldsymbol{x}\| + L\|\boldsymbol{x}^+ - \boldsymbol{x}\|,$$

where the inequality follows from the L-Lipschitz of the gradient operator $\nabla f(\cdot)$. ■

From the sufficient decrease property, it is easy to get the final convergence result under the nonconvex setting.

Theorem 6 (Nonconvex)**.** *Suppose that f is L-smooth and g is convex and closed. Let $t = \frac{1}{2L}$. Then, we have*

$$\min_{k \in [K]} \text{dist}(0, \partial F(\boldsymbol{x}^{k+1}) \le O\left(\frac{1}{\sqrt{K}}\right).$$

Proof. From Proposition 11, we have

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \le -\frac{L}{2} \|\boldsymbol{x}^+ - \boldsymbol{x}\|^2 \le -\frac{1}{18L} \text{dist}^2(0, \partial F(\boldsymbol{x}^{k+1})),$$

where the last inequality follows from Proposition 12. Then, we do the telescoping from $k = 0$ to $k = K$ and get

$$\min_{k \in [K]} \text{dist}^2(0, \partial F(\boldsymbol{x}^{k+1}) \le \frac{F(\boldsymbol{x}^0) - F^\star}{18LK}.$$

■

So far, you can see a significant distinction between the convex and nonconvex settings. A natural question raises:

Can we provide a unified convergence analysis framework for PGD at least?

3 Unification under KŁ Framework

Definition 7 (KŁ Exponent)**.** *Suppose that the problem (P) has a nonempty solution set and a finite optimal value. We say $F(\boldsymbol{x})$ is a KŁ function with the exponent $\theta \in (0,1]$ at the point $\boldsymbol{x}$ if we have*

$$\operatorname{dist}(0, \partial F(\boldsymbol{x})) \geq \sqrt{2\mu}\left(F(\boldsymbol{x}) - \max_{x \in \mathbb{R}^d} F(\boldsymbol{x})\right)^{\theta}.$$

Remark 8. *As we discussed in the previous lectures, if $F(\boldsymbol{x})$ is μ-strongly convex, then we know $F(\boldsymbol{x})$ is a KŁ function with the exponent $\theta = \frac{1}{2}$. When the function $F(\boldsymbol{x})$ is just convex, we know that the function $F(\boldsymbol{x})$ is a KŁ function with the exponent $\theta = 1$ at all points within a bounded distance from the optimal set, i.e.,*

$$F(\boldsymbol{x}) - F(\boldsymbol{x}^\star) \leq \boldsymbol{g}^T(\boldsymbol{x} - \boldsymbol{x}^\star).$$

for all $g \in \partial F(\boldsymbol{x})$. Then, we have

$$\frac{1}{\|\boldsymbol{x} - \boldsymbol{x}^\star\|}(F(\boldsymbol{x}) - F(\boldsymbol{x}^\star)) \leq \operatorname{dist}(0, \partial F(\boldsymbol{x})). \tag{3}$$

3.1 Recover the Convergence Result (Convex)

Assumption 9. *The function $F : \mathbb{R}^d \to \mathbb{R}$ is level bounded (Coerciveness).*

Proof. From Proposition 11 and Proposition 12, we have

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\frac{1}{18L}\operatorname{dist}^2(0, \partial F(\boldsymbol{x}^{k+1})).$$

Then, we know the sequence $\{F(\boldsymbol{x}^k)\}_{k \leq 0}$ is monotonically decreasing and $F(\boldsymbol{x}^k) \leq F(\boldsymbol{x}^0)$ holds for any $k \geq 0$ from Assumption 9. WLOG, we can assume that

$$\operatorname{dist}(\boldsymbol{x}^k, \mathcal{X}^\star) \leq D, \forall k \geq 0.$$

Armed with (3), we have

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\frac{1}{18LD^2}(F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^\star))^2,$$

which implies

$$\frac{1}{F(\boldsymbol{x}^{k+1}) - F^\star} \geq \frac{k+1}{36LD^2}.$$

(Deriving by the mathematical induction). We finished the proof. ∎

4 Recover the Convergence Result (Strongly Convex)

Proof. From Proposition 11 and Proposition 12, we have

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\frac{1}{18L}\operatorname{dist}^2(0, \partial F(\boldsymbol{x}^{k+1})).$$

As F is μ-strongly convex, we have

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\frac{\mu}{9L}(F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^\star)).$$

Finally, we get

$$F(\boldsymbol{x}^{k+1}) - F^\star \leq \left(1/(1 + \frac{\mu}{9L})\right) F(\boldsymbol{x}^k) - F^\star.$$

∎

A general convergence analysis template:

Theorem 10. *Suppose that the following conditions hold.*

(i) (Sufficient Decrease Property): For any $k \geq 0$, we have

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\kappa_1 \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2,$$

for some constant $\kappa_1 > 0$. It aims to quantify the algorithmic progress at each step.

(ii) (Safeguard Condition): For any $k \geq 0$, we have

$$\operatorname{dist}(0, \partial F(\boldsymbol{x}^{k+1})) \leq \kappa_2 \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|,$$

for some constant $\kappa_2 > 0$. It links the relative change produced by the algorithm to the optimality residual.

(iii) (Growth Condition):

$$\operatorname{dist}(0, \partial F(\boldsymbol{x})) \geq \sqrt{2\mu} \left(F(\boldsymbol{x}) - \max_{x \in \mathbb{R}^d} F(\boldsymbol{x}) \right)^{\theta},$$

which depends solely on the problem and is independent of the algorithm.

Then, we have

$$F(x^K) - F^\star \leq \mathcal{O}(K^{-\frac{1}{(2\theta-1)_+}}).$$

Proof.

$$\begin{aligned} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) &\leq -\kappa_1 \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \\ &\leq -\frac{\kappa_1}{\kappa_2^2} \operatorname{dist}^2(0, \partial F(\boldsymbol{x}^{k+1})) \\ &\leq -\frac{2\kappa_1 \mu}{\kappa_2^2} (F(\boldsymbol{x}^{k+1}) - F^\star)^{2\theta}. \end{aligned}$$

Define $R(\boldsymbol{x}^k) := F(\boldsymbol{x}^k) - F^\star$. Then, we have

$$R(\boldsymbol{x}^{k+1}) - R(\boldsymbol{x}^k) \leq -\alpha R(\boldsymbol{x}^{k+1})^{2\theta},$$

where $\alpha := \frac{2\kappa_1 \mu}{\kappa_2^2}$.

Case 1: When $\theta \in (0, 1/2)$, we have

$$R(\boldsymbol{x}^{k+1}) - R(\boldsymbol{x}^k) \leq -\alpha R(\boldsymbol{x}^{k+1}) R(\boldsymbol{x}^{k+1})^{2\theta-1}.$$

As $2\theta - 1 < 0$, we have $R(\boldsymbol{x}^{k+1})^{2\theta-1} \geq R(\boldsymbol{x}^0)^{2\theta-1}$. We can get the linear convergence rate.

Case 2: When $\theta = 1/2$, we trivially get the result.

Case 3: When $\theta \in (1/2, 1]$, we have:

$$R(\boldsymbol{x}^k) - R(\boldsymbol{x}^{k+1}) \geq \alpha R(\boldsymbol{x}^{k+1})^{2\theta}.$$

Consider the following two cases:

1. If $R(\boldsymbol{x}^k) \leq 2R(\boldsymbol{x}^{k+1})$, we denote $\psi(s) = \frac{1}{2\theta-1} s^{-(2\theta-1)}$ and then

$$\begin{aligned} \psi(R(\boldsymbol{x}^{k+1})) - \psi(R(\boldsymbol{x}^k)) &= \int_{R(\boldsymbol{x}^{k+1})}^{R(\boldsymbol{x}^k)} -\psi'(s)\mathrm{d}s = \int_{R(\boldsymbol{x}^{k+1})}^{R(\boldsymbol{x}^k)} s^{-2\theta}\mathrm{d}s \\ &\geq R(\boldsymbol{x}^k)^{-2\theta}(R(\boldsymbol{x}^k) - R(\boldsymbol{x}^{k+1})) \geq \alpha \left(\frac{R(\boldsymbol{x}^{k+1})}{R(\boldsymbol{x}^k)} \right)^{2\theta} \geq \frac{\alpha}{2^{2\theta}} \geq \frac{\alpha}{4}. \end{aligned}$$

2. If $R(\boldsymbol{x}^k) > 2R(\boldsymbol{x}^{k+1})$, we have

$$\begin{aligned}\psi(R(\boldsymbol{x}^{k+1})) - \psi(R(\boldsymbol{x}^k)) &= \frac{1}{2\theta-1}\left(R(\boldsymbol{x}^{k+1})^{-(2\theta-1)} - R(\boldsymbol{x}^k)^{-(2\theta-1)}\right) \\ &\geq \frac{1}{2\theta-1}\left(R(\boldsymbol{x}^{k+1})^{-(2\theta-1)} - (2R(\boldsymbol{x}^{k+1}))^{-(2\theta-1)}\right) \\ &\geq \frac{1-2^{-(2\theta-1)}}{2\theta-1}R(\boldsymbol{x}^{k+1})^{-(2\theta-1)} \geq \frac{1-2^{-(2\theta-1)}}{2\theta-1}R(\boldsymbol{x}^0)^{-(2\theta-1)}.\end{aligned}$$

Combing these two cases, we have

$$\psi(R(\boldsymbol{x}^{k+1})) - \psi(R(\boldsymbol{x}^k)) \geq \min\left\{\frac{\alpha}{4}, \frac{1-2^{-(2\theta-1)}}{2\theta-1}R(\boldsymbol{x}^0)^{-(2\theta-1)}\right\} = \frac{C}{2\theta-1} > 0.$$

Hence, we have

$$\psi(R(\boldsymbol{x}^k)) \geq \psi(R(\boldsymbol{x}^0)) + \frac{C}{2\theta-1}k \geq \frac{C}{2\theta-1}k$$

and

$$\frac{1}{2\theta-1}R(\boldsymbol{x}^k)^{-(2\theta-1)} \geq \frac{C}{2\theta-1}k \to R(\boldsymbol{x}^k)^{-(2\theta-1)} \geq Ck.$$

Finally, we have

$$R(\boldsymbol{x}^k) \leq (ck)^{-\frac{1}{2\theta-1}}.$$

■

5 Convergence Analysis of PPA (Nonconvex)

Similar with the analysis of PGD, we only have to check **sufficient decrease** and **safeguard** properties.

Proposition 11 (Sufficient Decrease Property)**.** *Suppose that F is ρ-weakly convex. Let $\boldsymbol{x}^+ = \mathrm{prox}_{tF}(\boldsymbol{x})$ where $t^{-1} > \rho$. Then, we have*

$$F(\boldsymbol{x}^+) \leq F(\boldsymbol{x}) - \frac{1}{2}\left(\frac{1}{t} - \rho\right)\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2.$$

Proof. By definition of the proximal operator, we have

$$\boldsymbol{x}^+ = \arg\min_{\boldsymbol{y}}\left\{\frac{1}{2t}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + F(\boldsymbol{y})\right\}.$$

Since $\boldsymbol{x}^+$ is the optimal solution of the $(\frac{1}{t} - \rho)$-strongly convex optimization problem, we have

$$\frac{1}{2}\left(\frac{1}{t} - \rho\right)\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2 \leq F(\boldsymbol{x}^+) - F(\boldsymbol{x}).$$

■

Proposition 12 (Safeguard Condition)**.** *Suppose that F is ρ-weakly convex. Let $\boldsymbol{x}^+ = \mathrm{prox}_{tF}(\boldsymbol{x})$ where $t^{-1} > \rho$. Then, we have*

$$\mathrm{dist}(0, \partial F(\boldsymbol{x}^+)) \leq \frac{1}{t}\|\boldsymbol{x}^+ - \boldsymbol{x}\|.$$

Proof. Recall that

$$\boldsymbol{x}^{+} = \arg\min_{\boldsymbol{y}} \left\{ \frac{1}{2t} \|\boldsymbol{y} - \boldsymbol{x}\|^2 + F(\boldsymbol{y}) \right\}.$$

We drive its optimality condition as

$$0 \in \frac{1}{t}(\boldsymbol{x}^{+} - \boldsymbol{x}) + \partial F(\boldsymbol{x}^{+}),$$

which implies

$$\operatorname{dist}(0, \partial F(\boldsymbol{x}^{+})) \leq \frac{1}{t}\|\boldsymbol{x}^{+} - \boldsymbol{x}\|.$$

■

Theorem 13 (Nonconvex)**.** *Suppose that F is ρ-weakly convex and $t^{-1} > \rho$. Then, we have*

$$\min_{k \in [K]} \operatorname{dist}(0, \partial F(\boldsymbol{x}^{k+1}) \leq O\left(\frac{1}{\sqrt{K}}\right).$$