The University of Texas at Austin
Department of Electrical and Computer Engineering

ECE 381V: Large-Scale Optimization II — Spring 2022

Lecture 13

Caramanis & Mokhtari Wednesday, March 2, 2022

Goal: In this lecture, we talk about the proximal gradient method and its convergence properties.

1 Beyond black-box optimization and Composite Optimization

Recall the following general convex (nonsmooth) constrained problem:

$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t } \mathbf{x} \in \mathcal{Q}, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex on $\mathbb{R}^n$ and the set $\mathcal{Q}$ is convex. We showed that the lower bound for finding an ϵ-accurate solution for the above problem is $1/\epsilon^2$ and projected subgradient method achieves this complexity upto a log factor.

One thing that we have to have in mind is that the above bounds are for the black-box setting, where we don't exactly know the objective function f and its structure and we simply query the objective function value and its subgradient at a point. In this lecture, we focus on the case that the explicit expression for the objective function is given and it has some nice properties.

In particular, we consider the following composite optimization problem where

$$\begin{aligned} &\min f(\mathbf{x}) := \phi(\mathbf{x}) + h(\mathbf{x}) \\ &\text{s.t } \mathbf{x} \in \mathcal{Q}, \end{aligned} \tag{2}$$

where ϕ is a convex and differentiable function that is smooth with L, and h is a convex and nonsmooth function that its proximal operator/mapping is easy to compute (will be defined later). In this case, we show that the above nonsmooth problem, that has a specific structure, can be solved as fast as a smooth convex problem. Note that we further assume that we have access to the gradient of $f(\mathbf{x})$ at a specific query point and we know the function $h(\mathbf{x})$. Before, formally defining how we can efficiently solve the above problem, we first introduce the proximal mapping.

2 Proximal Mapping (Prox-Operator)

The proximal mapping of a closed convex function $h : \mathbf{E} \to \mathbb{R}$, where $\mathbf{E}$ is a Euclidean space, is formally defined as

$$\text{prox}_h(\mathbf{x}) = \underset{\mathbf{u}}{\text{argmin}} \left\{ h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\}$$

In other words, the prox of function of h at point $\mathbf{x}$ is the minimizer of the following strictly convex function

$$g(\mathbf{u}; \mathbf{x}) := h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2$$

Indeed, as g is strictly convex, $\text{prox}_h(\mathbf{x})$ exists and its unique.

2.1 Examples

Example 1. Consider the convex quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$, it can be easily verified that its prox is given by

$$\text{prox}_f(\mathbf{x}) = \underset{\mathbf{u}}{\text{argmin}} \left\{ \frac{1}{2}\mathbf{u}^\top \mathbf{A}\mathbf{u} + \mathbf{b}^\top \mathbf{u} + c + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\} = (\mathbf{A} + \mathbf{I})^{-1}(\mathbf{x} - \mathbf{b})$$

Example 2. Consider the scalar function $f(x) = \lambda|x|$. It can be easily verified that

$$\text{prox}_f(x) = \underset{u}{\text{argmin}} \left\{ \lambda|u| + \frac{1}{2}|u - x|^2 \right\} = [|x| - \lambda]_+ \text{sign}(x)$$

this operator is also known as the shrinkage operator.

Example 3. Considering the above example, we can find the prox of ℓ_1 norm. Consider the function $f(\mathbf{x}) = \lambda\|\mathbf{x}\|$. It can be easily verified that

$$\text{prox}_f(\mathbf{x}) = \underset{\mathbf{u}}{\text{argmin}} \left\{ \lambda\|\mathbf{u}\|_1 + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\} = \underset{\mathbf{u}}{\text{argmin}} \left\{ \lambda \sum_{i=1}^{n} (|u_i| + \frac{1}{2}|u_i - x_i|^2) \right\}$$

Hence, if we denote $\mathbf{u}^* = \text{prox}_f(\mathbf{x})$ we can easily show that its i-th coordinate is given by

$$u_i^* = [|x_i| - \lambda]_+ \text{sign}(x_i)$$

Example 4. Consider the indicator function of a convex set $\mathcal{C}$ that we denote by $\delta_{\mathcal{C}}(\mathbf{x})$, which is zero for $\mathbf{x} \in \mathcal{C}$ and ∞ for $\mathbf{x} \notin \mathcal{C}$. In this case we have

$$\text{prox}_f(\mathbf{x}) = \underset{\mathbf{u}}{\text{argmin}} \left\{ \delta_{\mathcal{C}}(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\} = \underset{\mathbf{u} \in \mathcal{C}}{\text{argmin}} \left\{ \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\} = \pi_{\mathcal{C}}(\mathbf{x})$$

2.2 Properties of proximal mapping

Theorem 1. *If we use the notation* $\hat{\mathbf{x}} := \mathit{prox}_h(\mathbf{x})$ *then we have*

$$\mathbf{0} \in \partial h(\hat{\mathbf{x}}) + (\hat{\mathbf{x}} - \mathbf{x}) \quad \iff \quad \mathbf{x} - \hat{\mathbf{x}} \in \partial h(\hat{\mathbf{x}}) \quad \iff \quad h(\mathbf{z}) - h(\hat{\mathbf{x}}) \geq (\mathbf{x} - \hat{\mathbf{x}})^\top (\mathbf{z} - \hat{\mathbf{x}}) \;\; \text{for all } \mathbf{z}$$

Theorem 2. *If f is a proper convex function, then we have*

$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathbf{E}}{\text{argmin}}\, f(\mathbf{x}) \iff \mathbf{x}^* = \mathit{prox}_f(\mathbf{x}^*)$$

Proof.

$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathbf{E}}{\text{argmin}}\, f(\mathbf{x}) \iff \mathbf{0} \in \partial f(\mathbf{x}^*) \iff \mathbf{x}^* - \mathbf{x}^* \in \partial f(\mathbf{x}^*) \iff \mathbf{x}^* = \text{prox}_f(\mathbf{x}^*)$$

□

Theorem 3. *Let f be a convex function. Then for any $\mathbf{x}$ and $\mathbf{y}$ we have*

1. $(\mathbf{x} - \mathbf{y})^\top (\mathit{prox}_f(\mathbf{x}) - \mathit{prox}_f(\mathbf{y})) \geq \|\mathit{prox}_f(\mathbf{x}) - \mathit{prox}_f(\mathbf{y})\|^2$
2. $\|\mathit{prox}_f(\mathbf{x}) - \mathit{prox}_f(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$

Proof. To prove the first result, simply use the fact that

$$\mathbf{x} - \text{prox}_f(\mathbf{x}) \in \partial f(\text{prox}_f(\mathbf{x})), \qquad \mathbf{y} - \text{prox}_f(\mathbf{y}) \in \partial f(\text{prox}_f(\mathbf{y}))$$

Using the monotonicity of (subgradient) subdifferential we have

$$((\mathbf{x} - \text{prox}_f(\mathbf{x})) - (\mathbf{y} - \text{prox}_f(\mathbf{y})))^\top (\text{prox}_f(\mathbf{x}) - \text{prox}_f(\mathbf{y})) \geq 0$$

which leads to the first claim. The second claim simply follows from the first result and the Cauchy–Schwarz inequality. □

3 Proximal Gradient Method for Composite Optimization

Composite Optimization. In this section, we formally study the composite optimization problem

$$\min f(\mathbf{x}) := \phi(\mathbf{x}) + h(\mathbf{x}) \tag{3}$$

- ϕ is convex and and L-smooth.
- h is convex (possibly nondifferentiable) and its prox operator is easy to compute.
- The optimal solution set is nonempty.

Perhaps the most special case of the above problem and algorithm is the ℓ_1 regularization problem:

$$\min f(\mathbf{x}) := \phi(\mathbf{x}) + \lambda \|\mathbf{x}\|_1$$

Proximal gradient method (PGM). To solve the above problem, we follow the update:

$$\mathbf{x}_{k+1} = \underset{\mathbf{u}}{\text{argmin}} \left\{ \phi(\mathbf{x}_k) + \nabla\phi(\mathbf{x}_k)^\top (\mathbf{u} - \mathbf{x}_k) + \frac{1}{2\eta_k} \|\mathbf{u} - \mathbf{x}_k\|^2 + h(\mathbf{u}) \right\}$$

which is also equivalent to the following update

$$\mathbf{x}_{k+1} = \underset{\mathbf{u}}{\text{argmin}} \left\{ \eta_k h(\mathbf{u}) + \|\mathbf{u} - (\mathbf{x}_k - \eta_k \nabla\phi(\mathbf{x}_k))\|^2 \right\}$$

and by definition is equivalent to

$$\mathbf{x}_{k+1} = \text{prox}_{\eta_k h}(\mathbf{x}_k - \eta_k \nabla\phi(\mathbf{x}_k))$$

You can also think of it as

$$\begin{aligned} \mathbf{z}_{k+1} &= \mathbf{x}_k - \eta_k \nabla\phi(\mathbf{x}_k) \\ \mathbf{x}_{k+1} &= \text{prox}_{\eta_k h}(\mathbf{z}_k) \end{aligned}$$

3.1 Special cases of PGM

Case 1. When $h \equiv 0$, then the algorithm becomes the gradient descent method for minimizing the smooth function f.

Case 2. When $h(\mathbf{x}) = \delta_{\mathcal{C}}(\mathbf{x})$, then the algorithm becomes the projected gradient method for minimizing the function f over the set $\mathcal{C}$. Why?

$$\mathbf{x}_{k+1} = \underset{\mathbf{u}}{\text{argmin}} \left\{ \eta_k \delta_{\mathcal{C}}(\mathbf{u}) + \|\mathbf{u} - (\mathbf{x}_k - \eta_k \nabla\phi(\mathbf{x}_k))\|^2 \right\} = \underset{\mathbf{u}\in\mathcal{C}}{\text{argmin}} \left\{ \|\mathbf{u} - (\mathbf{x}_k - \eta_k \nabla\phi(\mathbf{x}_k))\|^2 \right\} = \pi_{\mathcal{C}}(\mathbf{x}_k - \eta_k \nabla\phi(\mathbf{x}_k))$$

Case 3. When we deal with ℓ_1 norm minimization as out nonsmooth part, then the PGM method is also known as Iterative Shrinkage-Thresholding Algorithms (ISTA), which is very useful for solving linear inverse problems arising in signal/image processing.

4 Convergence Analysis of PGM

First, we state a fundamental lemma that we use in the convergence analysis of PGM.

Lemma 1. *Consider the function $f(\mathbf{x}) = \phi(\mathbf{x}) + h(\mathbf{x})$ under the assumptions that ϕ is L-smooth and h is convex. Then, for any $\eta \leq 1/L$ we have*

$$f(\mathbf{x}) - f(prox_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))) \geq \frac{1}{2\eta}\|\mathbf{x} - prox_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))\|^2 - \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2 + \phi(\mathbf{x}) - \phi(\mathbf{y}) - \nabla\phi(\mathbf{y})^\top(\mathbf{x} - \mathbf{y})$$

Proof. Note that using L-smoothness of ϕ we have

$$\phi(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))) \leq \phi(\mathbf{y}) + \nabla\phi(\mathbf{y})^\top(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y})) - \mathbf{y}) + \frac{1}{2\eta}\|\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y})) - \mathbf{y}\|^2 \tag{4}$$

Now consider the function $\zeta(\mathbf{u})$ defined as

$$\zeta(\mathbf{u}) = \phi(\mathbf{y}) + \nabla\phi(\mathbf{y})^\top(\mathbf{u} - \mathbf{y}) + h(\mathbf{u}) + \frac{1}{2\eta}\|\mathbf{u} - \mathbf{y}\|^2$$

Indeed this function is strongly convex and its minimizer is $\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))$. Now using the fact that $\zeta(\mathbf{u})$ is strongly convex with constant $1/\eta$, and its minimizer is $\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))$, we can write

$$\zeta(\mathbf{x}) - \zeta(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))) \geq \frac{1}{2\eta}\|\mathbf{x} - \text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))\|^2$$

Now note that

$$\begin{aligned}
&\zeta(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))) \\
&= \phi(\mathbf{y}) + \nabla\phi(\mathbf{y})^\top(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y})) - \mathbf{y}) + h(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))) + \frac{1}{2\eta}\|\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y})) - \mathbf{y}\|^2 \\
&\geq \phi(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))) + h(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))) \\
&= f(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y})))
\end{aligned}$$

Using this result we have

$$\zeta(\mathbf{x}) - f(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))) \geq \frac{1}{2\eta}\|\mathbf{x} - \text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))\|^2$$

Now if we replace $\zeta(\mathbf{x})$ by its definition we obtain that

$$\phi(\mathbf{y}) + \nabla\phi(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}) + h(\mathbf{x}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2 - f(\text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))) \geq \frac{1}{2\eta}\|\mathbf{x} - \text{prox}_{\eta h}(\mathbf{y} - \eta\nabla\phi(\mathbf{y}))\|^2$$

By adding $\phi(\mathbf{x})$ to both sides and regrouping the terms the claim follows. □

Note that the above result also holds for the case that ϕ is not convex.

4.1 Convex setting

Next, we use this result to show the convergence rate of PGM for the convex setting.

Theorem 4. *Consider PGM applied to the function $f(\mathbf{x}) = \phi(\mathbf{x}) + h(\mathbf{x})$ with stepsize $\eta_k = \eta \leq 1/L$, for the case that ϕ is L-smooth and convex and h is convex. Then, we have*

$$f(\mathbf{x}_N) - f(\mathbf{x}^*) \leq \frac{1}{2\eta}\frac{\|\mathbf{x}^* - \mathbf{x}_0\|^2}{N}$$

Proof. In the result of Lemma 1, replace $\mathbf{x}$ by $\mathbf{x}^*$, $\mathbf{y}$ by $\mathbf{x}_k$ and $\text{prox}_{\eta h}(\mathbf{y} - \eta \nabla \phi(\mathbf{y}))$ by $\mathbf{x}_{k+1}$. Then, we obtain

$$f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|^2 - \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_k\|^2 + \phi(\mathbf{x}^*) - \phi(\mathbf{x}_k) - \nabla\phi(\mathbf{x}_k)^\top(\mathbf{x}^* - \mathbf{x}_k)$$

Now by convexity of ϕ the term $\phi(\mathbf{x}^*) - \phi(\mathbf{x}_k) - \nabla\phi(\mathbf{x}_k)^\top(\mathbf{x}^* - \mathbf{x}_k)$ is indeed non-negative hence we have

$$f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|^2 - \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_k\|^2$$

A simple telescope argument implies

$$\sum_{k=0}^{N-1}(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \leq \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_0\|^2 - \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_N\|^2 \leq \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_0\|^2$$

Now in Lemma 1, set $\mathbf{x} = \mathbf{y} = \mathbf{x}_k$ to obtain

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2\eta}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2$$

Hence, $f(\mathbf{x}_k)$ is monotonically deceasing and we have

$$N(f(\mathbf{x}_N) - f(\mathbf{x}^*)) \leq \sum_{k=0}^{N-1}(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \leq \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_0\|^2$$

and the claim follows. □

4.2 Strongly-convex setting

Theorem 5. *Consider PGM applied to the function $f(\mathbf{x}) = \phi(\mathbf{x}) + h(\mathbf{x})$ with stepsize $\eta_k = \eta \leq 1/L$, for the case that ϕ is L-smooth and μ-strongly convex and h is convex. Then, we have*

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq (1 - \eta\mu)\|\mathbf{x}_k - \mathbf{x}^*\|^2$$

Proof. In the result of Lemma 1, replace $\mathbf{x}$ by $\mathbf{x}^*$, $\mathbf{y}$ by $\mathbf{x}_k$ and $\text{prox}_{\eta h}(\mathbf{y} - \eta \nabla \phi(\mathbf{y}))$ by $\mathbf{x}_{k+1}$. Then, we obtain

$$f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|^2 - \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_k\|^2 + \phi(\mathbf{x}^*) - \phi(\mathbf{x}_k) - \nabla\phi(\mathbf{x}_k)^\top(\mathbf{x}^* - \mathbf{x}_k)$$

Now using the strong convexity of ϕ with parameter μ we have

$$f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|^2 - \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_k\|^2 + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_k\|^2$$

Now as $f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) \leq 0$ we have

$$0 \geq \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|^2 - \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_k\|^2 + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_k\|^2$$

and the claim follows by regrouping the terms. □