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

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

Lecture 8

Caramanis & Mokhtari Monday, February 14, 2022

Goal: In this lecture, we talk about projected gradient method and projected Nesterov's method.

1 Convex Constrained Optimization

Consider a general constrained convex optimization problem where the constraint set is a simple convex set with no functional constraints:

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable on the convex set $\mathcal{Q}$.

2 Gradient Mapping

When we deal with a constrained optimization problem, following the update of gradient descent methods could lead to a infeasible point. Hence, we need projection steps to keep our iterates feasible. A useful tool for defining projection-based methods is the gradient mapping.

Definition 1. *For a given $\gamma > 0$ and point $\hat{\mathbf{x}}$, consider the point*

$$\mathbf{x}_{\mathcal{Q}}(\hat{\mathbf{x}}; \gamma) := \underset{\mathbf{x} \in \mathcal{Q}}{\operatorname{argmin}} \left\{ f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}}) + \frac{\gamma}{2} \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \right\}.$$

Then, the gradient mapping is defined as

$$\mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}}; \gamma) := \gamma(\hat{\mathbf{x}} - \mathbf{x}_{\mathcal{Q}}(\hat{\mathbf{x}}; \gamma)).$$

Now let's try to understand the geometry behind the new point $\mathbf{x}_{\mathcal{Q}}(\hat{\mathbf{x}}; \gamma)$ and its corresponding gradient mapping $\mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}}; \gamma)$.

Note that $\mathbf{x}_{\mathcal{Q}}(\hat{\mathbf{x}}; \gamma)$ is the unique minimizer of the strongly convex objective function $f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}}) + \frac{\gamma}{2} \|\mathbf{x} - \hat{\mathbf{x}}\|^2$ over the set $\mathcal{Q}$. By regrouping the terms we can shows that

$$\begin{aligned} \mathbf{x}_{\mathcal{Q}}(\hat{\mathbf{x}}; \gamma) &= \underset{\mathbf{x} \in \mathcal{Q}}{\operatorname{argmin}} \left\{ f(\hat{\mathbf{x}}) + \frac{\gamma}{2} \|\mathbf{x} - \hat{\mathbf{x}} + \frac{1}{\gamma} \nabla f(\hat{\mathbf{x}})\|^2 - \frac{1}{2\gamma} \|\nabla f(\hat{\mathbf{x}})\|^2 \right\} \\ &= \underset{\mathbf{x} \in \mathcal{Q}}{\operatorname{argmin}} \left\{ \|\mathbf{x} - \hat{\mathbf{x}} + \frac{1}{\gamma} \nabla f(\hat{\mathbf{x}})\|^2 \right\} \\ &= \pi_{\mathcal{Q}} \left(\hat{\mathbf{x}} - \frac{1}{\gamma} \nabla f(\hat{\mathbf{x}}) \right) \end{aligned}$$

Therefore, $\mathbf{x}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)$ is the projection of the point obtained after running one step of gradient descent on $\hat{\mathbf{x}}$ with stepsize $1/\gamma$.

The gradient mapping can be also written as

$$\mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma) = \gamma\left(\hat{\mathbf{x}} - \pi_{\mathcal{Q}}\left(\hat{\mathbf{x}} - \frac{1}{\gamma}\nabla f(\hat{\mathbf{x}})\right)\right).$$

It can be easily verified that $\mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)$ is equal to the gradient $\nabla f(\hat{\mathbf{x}})$ if $\hat{\mathbf{x}} - \frac{1}{\gamma}\nabla f(\hat{\mathbf{x}})$ belongs to the feasible set and projection is inactive.

2.1 Properties of gradient mapping

Perhaps two major inequalities that we previously used for the analysis of gradient descent were:

$$f\left(\mathbf{x} - \frac{1}{\gamma}\nabla f(\mathbf{x})\right) \leq f(\mathbf{x}) - \frac{1}{2\gamma}\|\nabla f(\mathbf{x})\|^2$$

fore $\gamma \geq L$ and

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

Next, we establish similar results for the case of projected methods.

Theorem 1. *Suppose $f \in \mathcal{F}^{1,1}_{\mu,L}(\mathcal{Q})$ and $\gamma \geq L$. Then, for any $\mathbf{x} \in \mathcal{Q}$ we have*

$$f(\mathbf{x}) \geq f\left(\mathbf{x}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)\right) + \mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)^\top(\mathbf{x} - \hat{\mathbf{x}}) + \frac{1}{2\gamma}\|\mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)\|^2 + \frac{\mu}{2}\|\mathbf{x} - \hat{\mathbf{x}}\|^2$$

Proof: Exercise. (You can also check the proof of Theorem 2.2.13 in the textbook)

The above result leads to some important inequalities for gradient mapping:

1. $f\left(\mathbf{x}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)\right) \leq f(\hat{\mathbf{x}}) - \frac{1}{2\gamma}\|\mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)\|^2$
2. $\mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)^\top(\hat{\mathbf{x}} - \mathbf{x}^*) \geq \frac{1}{2\gamma}\|\mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)\|^2 + \frac{\mu}{2}\|\mathbf{x}^* - \hat{\mathbf{x}}\|^2 + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)\|^2$

3 Projected Gradient Method

The update of projected gradient method is formally given as

$$\mathbf{x}_{k+1} = \pi_{\mathcal{Q}}\left(\mathbf{x}_k - \frac{1}{\gamma}\nabla f(\mathbf{x}_k)\right) = \mathbf{x}_k - \frac{1}{\gamma}\mathbf{g}_{\mathcal{Q}}(\hat{\mathbf{x}};\gamma)$$

Note that the lower bounds for the unconstrained settings indeed hold for the constrained setting, as $\mathbb{R}^n$ is a closed convex set and a special case of our constrained problem.

3.1 Analysis for the Strongly-Convex and Smooth Setting

Next, we analyze the convergence rate of projected gradient descent (PGD) method for the case of strongly convex and smooth setting.

Theorem 2. *Let $f \in \mathcal{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$. If $\gamma \geq (\mu + L)/2$, then the iterates of PGD satisfy*

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \left(1 - \frac{\mu}{\gamma}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|$$

Proof. The proof is very similar to the proof for the unconstrained case:

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2 &= \|\pi_{\mathcal{Q}}\left(\mathbf{x}_k-\frac{1}{\gamma}\nabla f(\mathbf{x}_k)\right)-\mathbf{x}^*\|^2\\ &= \|\pi_{\mathcal{Q}}\left(\mathbf{x}_k-\frac{1}{\gamma}\nabla f(\mathbf{x}_k)\right)-\pi_{\mathcal{Q}}(\mathbf{x}^*-\frac{1}{\gamma}\nabla f(\mathbf{x}^*))\|^2\\ &\leq \|\mathbf{x}_k-\frac{1}{\gamma}\nabla f(\mathbf{x}_k)-(\mathbf{x}^*-\frac{1}{\gamma}\nabla f(\mathbf{x}^*))\|^2\end{aligned}$$

where the second equality follows from the result in the previous lecture that $\mathbf{x}^*=\pi_{\mathcal{Q}}(\mathbf{x}^*-\frac{1}{\gamma}\nabla f(\mathbf{x}^*))$ and the inequality follows by the contraction property of the projection. The rest of the proof is identical to the proof for the unconstrained setting. In particular, we have

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2 &\leq \|\mathbf{x}_k-\mathbf{x}^*\|^2-2(\mathbf{x}_k-\mathbf{x}^*)^\top(\nabla f(\mathbf{x}_k)-\nabla f(\mathbf{x}^*))+\frac{1}{\gamma^2}\|\nabla f(\mathbf{x}_k)-\nabla f(\mathbf{x}^*)\|^2\\ &\leq \left(1-\frac{2}{\gamma}\frac{\mu L}{\mu+L}\right)\|\mathbf{x}_k-\mathbf{x}^*\|^2+\left(\frac{1}{\gamma^2}-\frac{2}{\gamma}\frac{1}{\mu+L}\right)\|\nabla f(\mathbf{x}_k)-\nabla f(\mathbf{x}^*)\|^2\\ &\leq \left(1-\frac{2}{\gamma}\frac{\mu L}{\mu+L}\right)\|\mathbf{x}_k-\mathbf{x}^*\|^2+\mu^2\left(\frac{1}{\gamma^2}-\frac{2}{\gamma}\frac{1}{\mu+L}\right)\|\mathbf{x}_k-\mathbf{x}^*\|^2\\ &= \left(1-\frac{\mu}{\gamma}\right)^2\|\mathbf{x}_k-\mathbf{x}^*\|^2\end{aligned}$$

and the claim follows. □

Indeed, by setting $\gamma=\frac{L+\mu}{2}$ we obtain

$$\|\mathbf{x}_k-\mathbf{x}^*\|\leq\left(\frac{L-\mu}{L+\mu}\right)^k\|\mathbf{x}_0-\mathbf{x}^*\|$$

3.2 Analysis for the Smooth Setting

Theorem 3. *Let $f\in\mathcal{F}_L^{1,1}(\mathbb{R}^n)$. If $\gamma=L$, then the iterates of PGD satisfy*

$$f(\mathbf{x}_k)-f(\mathbf{x}^*)\leq\frac{3L\|\mathbf{x}_0-\mathbf{x}^*\|^2+f(\mathbf{x}_0)-f(\mathbf{x}^*)}{k}$$

Proof. Based on the first side result of Theorem 1 we have

$$f(\mathbf{x}_{k+1})\leq f(\mathbf{x}_k)-\frac{1}{2L}\|\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k;L)\|^2$$

which implies that

$$f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)\leq f(\mathbf{x}_k)-f(\mathbf{x}^*)-\frac{1}{2L}\|\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k;L)\|^2 \tag{2}$$

Again based on Theorem 1 we have (set $\mathbf{x}=\mathbf{x}^*$ and $\mu=0$)

$$f(\mathbf{x}^*)\geq f\left(\mathbf{x}_{k+1}\right)+\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k;L)^\top(\mathbf{x}^*-\mathbf{x}_k)+\frac{1}{2L}\|\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k;L)\|^2$$

which implies that

$$f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)\leq\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k;L)^\top(\mathbf{x}_k-\mathbf{x}^*)-\frac{1}{2L}\|\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k;L)\|^2\leq\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k;L)^\top(\mathbf{x}_k-\mathbf{x}^*)$$

Hence, we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \|\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k; L)\| \|\mathbf{x}_k - \mathbf{x}^*\| \tag{3}$$

Now we show that $\|\mathbf{x}_k - \mathbf{x}^*\|$ is decreasing:

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_k - \frac{1}{L}\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k; L) - \mathbf{x}^*\|^2 = \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \frac{2}{L}\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k; L)^\top (\mathbf{x}_k - \mathbf{x}^*) + \frac{1}{L^2}\|\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k; L)\|^2 \leq \|\mathbf{x}_k - \mathbf{x}^*\|^2$$

where the last inequality follows from the second side-result of Theorem 1. Hence, we can write

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \|\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k; L)\| \|\mathbf{x}_0 - \mathbf{x}^*\| \tag{4}$$

which implies that

$$\frac{f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)}{\|\mathbf{x}_0 - \mathbf{x}^*\|} \leq \|\mathbf{g}_{\mathcal{Q}}(\mathbf{x}_k; L)\| \tag{5}$$

and therefore, by replacing the lower bound in (5) into (2) we obtain that

$$f(\mathbf{x}_{k+1}) - f^* \leq f(\mathbf{x}_k) - f^* - \frac{1}{2L}\frac{(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*))^2}{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}$$

Now using a similar trick that we did for the unconstrained case we can show that

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{3L\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + f(\mathbf{x}_0) - f(\mathbf{x}^*)}{k}$$

and the proof is complete. □

4 Projected Accelerated Gradient Descent Method

Similarly one can extend the proof of the accelerated gradient descent method to the constrained setting by using the projection scheme. In fact, the update in this case, is given by:

$$\mathbf{y}_k = \mathbf{x}_k + \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}(\mathbf{x}_k - \mathbf{x}_{k-1})$$

$$\mathbf{x}_{k+1} = \pi_{\mathcal{Q}}\left(\mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)\right)$$

Note that in this case, only the iterates $\mathbf{x}_k$ are feasible and the intermediate iterates $\mathbf{y}_k$ are only used for gradient computation and they may not be feasible. The details of extending the analysis of the above algorithm can be found in Section 2.2.5 of the textbook. It is worth noting that the projected accelerated gradient descent method obtains the optimal complexity bound for both convex smooth and strongly-convex smooth settings.