SYS 6003: Optimization	**Fall 2016**

Lecture 17

Instructor: Quanquan Gu *Date: Oct* 26^{th}

Today we are going to study the projected gradient descent algorithm.

Consider the following contrained optimization problem :

$$\min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}). \tag{1}$$

If we apply the gradient descent algorithm directly, we cannot guarantee that in each iteration, $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$ will be in $\mathcal{D}$. In other words, we may end up with infeasible solutions. To ensure that the new point, $\mathbf{x}_{t+1}$, that obtained in each iteration will be always in $\mathcal{D}$, one way is to project the new point back onto the feasible set.

Let us first define the projection of a point onto a set.

Definition 1 (Projection) *The projection point of* $\mathbf{x}$ *onto a set* $\mathcal{C}$ *is defined as* $\Pi_{\mathcal{C}}(\mathbf{x}) := \arg\min_{\mathbf{y}\in\mathcal{C}} \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$.

Theorem 1 (Projection Theorem) *Let* $\mathcal{C} \subseteq \mathbb{R}^d$ *be a convex set. For any* $\mathbf{x} \in \mathbb{R}^d$ *and* $\mathbf{y} \in \mathcal{C}$*, it holds that*

(1) $(\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{y})^\top (\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{x}) \leq 0$.*;*

(2) $\|\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{y}\|_2^2 + \|\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{x}\|_2^2 \leq \|\mathbf{x}-\mathbf{y}\|_2^2$.

Proof: (1) Let $f(\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$. By the first order necessary condition of local minimum $\mathbf{y}^* = \Pi_{\mathcal{C}}(\mathbf{x})$, we have $\nabla f(\mathbf{y}^*)^\top \mathbf{d} \geq 0$ where $\mathbf{d}$ is any feasible directions at $\mathbf{y}^*$. Let $\mathbf{d} = \mathbf{y} - \Pi_{\mathcal{C}}(\mathbf{x})$. For any $\mathbf{y} \in \mathcal{C}$, it then follows that

$$\nabla f(\mathbf{y}^*)^T (\mathbf{y} - \Pi_{\mathcal{C}}(\mathbf{x})) \geq 0. \tag{2}$$

Note that $\nabla f(\mathbf{y}^*) = -(\mathbf{x} - \mathbf{y}^*) = \mathbf{y}^* - \mathbf{x}$ and $\mathbf{y}^* = \Pi_C(\mathbf{x})$. From (2), it then follows

$$\begin{aligned} &(\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{x})^\top (\mathbf{y} - \Pi_{\mathcal{C}}(\mathbf{x})) \geq 0, \quad \text{i.e.,} \\ &(\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{x})^\top (\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{y}) \leq 0. \end{aligned}$$

(2) We have

$$\begin{aligned} \|\mathbf{x}-\mathbf{y}\|_2^2 &= \|\mathbf{x} - \Pi_{\mathcal{C}}(\mathbf{x}) + \Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{y}\|_2^2 \\ &= \|\mathbf{x} - \Pi_{\mathcal{C}}(\mathbf{x})\|_2^2 + \|\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{y}\|_2^2 - 2(\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{y})^\top (\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{x}) \\ &\geq \|x - \Pi_{\mathcal{C}}(\mathbf{x})\|_2^2 + \|\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{y}\|_2^2, \end{aligned}$$

where the inequality follows from part (1). This completes the proof. ■

Remark 1 *Geometrically, the projection theorem says that the the angle between vectors* $\mathbf{y} - \Pi_{\mathcal{C}}(\mathbf{x})$ *and* $\Pi_{\mathcal{C}}(\mathbf{x}) - \mathbf{x}$ *is either acute or right.*

Algorithm 1 Projected Gradient Descent

1: **Input:** η_t
2: **Initialize:** $\mathbf{x}_1 \in \mathcal{D}$
3: **for** $t = 1$ to $T - 1$ **do**
4: $\quad \mathbf{x}_{t+1} = \Pi_{\mathcal{D}}[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)]$
5: **end for**

So we modify the updating rule of gradient descent to be $\mathbf{x}_{t+1} = \Pi_{\mathcal{D}}[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)]$, where $\Pi_{\mathcal{D}}(\mathbf{x})$ is the projection of $\mathbf{x}$ onto $\mathcal{D}$. Then we have the *projected gradient descent algorithm* shown in Algorithm 1. It is worth noting that if the gradient of f does not exist at $\mathbf{x}_t$, then in the fourth line of Algorithm 1, we can use any subgradient of f at $\mathbf{x}_t$ instead of its gradient.

The following theorem provides the convergence rate for the projected gradient descent algorithm.

Theorem 2 *Suppose that f is a convex function, and its subgradient $\mathbf{g}(\mathbf{x})$ is bounded by G, i.e., $\|\mathbf{g}(\mathbf{x})\|_2 \leq G$, for any $\mathbf{x} \in \mathcal{D}$. Then for the projected gradient descent with $\eta_t = 1/\sqrt{t}$, it holds that*

$$f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\right) - f(\mathbf{x}^*) \leq \left(\frac{R^2}{2} + G^2\right)\frac{1}{\sqrt{T}}$$

where $\mathbf{x}^*$ is the optimal solution to problem (1) and $R = \max_{\mathbf{x},\mathbf{y}\in\mathcal{D}} \|\mathbf{x} - \mathbf{y}\|_2$ is the diameter of the set convex $\mathcal{D}$.