Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Chao Qin
Email: cq2199@columbia.edu

SCRIBE
8

Projected Gradient Descent

In the previous lectures, we have studied the gradient descent algorithm and its analysis under three conditions. In this lecture, we consider the general convex feasible set K, and propose the projected gradient descent. In addition, we analyze this algorithm under the same conditions.

1. Projected Gradient Descent

We first introduce the projection operator from $\mathbb{R}^d$ to the feasible set K.

Definition 8.1 The projection operator $\Pi_K : \mathbb{R}^d \to K$ is defined as

$$\Pi_K(y) = \operatorname*{argmin}_{x \in K} \|y - x\|$$

Now we are ready to propose the projected gradient descent for general convex feasible set K.

Param: $\eta > 0$, which is the stepsize;
Init: $x_0 \in K$ arbitrary;
for $t = 0, 1, 2, \ldots$ **do**
    $y_{t+1} = x_t - \eta \nabla f(x_t)$;
    $x_{t+1} = \Pi_K(y_{t+1})$
end
return x_T (or some combination of $x_0, \ldots, x_T$)

Algorithm 1: Projected Gradient Descent

2. Analysis of Projected GD

To analyze projected GD, we need the following property of the projection operator.

Lemma 8.0.1 (Version of Pythagoras) For any $y \in \mathbb{R}^d$ and $x \in K$,

$$\|y - x\|^2 \geq \|y - \Pi_K(y)\|^2 + \|\Pi_K(y) - x\|^2.$$

Proof. Since $\Pi_K(y)$ is a minimizer of $f(x) = \|x-y\|^2$ on K, by the first-order condition, we have

$$\nabla f(\Pi_K(y))(x - \Pi_K(y)) = (\Pi_K(y) - y)(x - \Pi_K(y)) \geq 0.$$

Hence,

$$\begin{aligned}\|y - x\|^2 &= \|y - \Pi_K(y)\|^2 + \|\Pi_K(y) - x\|^2 + (y - \Pi_K(y))(\Pi_K(y) - x) \\ &\geq \|y - \Pi_K(y)\|^2 + \|\Pi_K(y) - x\|^2.\end{aligned}$$

□

By applying this result with a choice of $x = x^*$ and $y = y_{t+1}$, we have

$$\|y_{t+1} - x^*\|^2 \geq \|y_{t+1} - x_{t+1}\|^2 + \|x_{t+1} - x^*\|^2 \tag{1}$$

Now we are ready to analyze projected GD under three conditions. We have

$$\|y_{t+1} - x^*\|^2 = \|x_t - \eta\nabla f(x_t) - x^*\|^2 = \|x_t - x^*\|^2 + \eta^2\|\nabla f(x_t)\|^2 - 2\eta\nabla f(x_t)^\top(x_t - x^*),$$

which leads to

$$\nabla f(x_t)^\top(x_t - x^*) = \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(x_t)\|^2.$$

2.1 f is Lipschitz with constant L

By convexity,

$$\begin{aligned}f(x_t) - f(x^*) \leq& \nabla f(x_t)^\top(x_t - x^*) \\ =& \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(x_t)\|^2.\end{aligned}$$

By Equation (1),

$$\|y_{t+1} - x^*\|^2 \geq \|y_{t+1} - x_{t+1}\|^2 + \|x_{t+1} - x^*\|^2 \geq \|x_{t+1} - x^*\|^2.$$

where $x_{t+1} = \Pi_K(y_{t+1})$. This leads to

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(x_t)\|^2.$$

Then the analysis is exactly the same as before.

2.2 f is β-smooth

Suppose we run projected gradient descent with $\eta = \frac{1}{\beta}$. We need the following helpful lemma:

Lemma 8.0.2 For $x \in K$, suppose $y = x - \frac{1}{\beta}\nabla f(x)$ and $x' = \Pi_K(y)$. Then,

$$f(x') \leq f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2 + \frac{\beta}{2}\|y - x'\|^2.$$

Proof. By β-smoothness,

$$
\begin{aligned}
f(x') &\le f(x) + \nabla f(x)^\top (x' - x) + \frac{\beta}{2}\|x' - x\|^2 \\
&= f(x) + \nabla f(x)^\top \left(x' - y - \frac{1}{\beta}\nabla f(x)\right) + \frac{\beta}{2}\left\|x' - y - \frac{1}{\beta}\nabla f(x)\right\|^2 \\
&= f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2 + \frac{\beta}{2}\|y - x'\|^2
\end{aligned}
$$

□

By applying this result with a choice of $x = x_t$, we have

$$\|\nabla f(x_t)\|^2 \le 2\beta(f(x_t) - f(x_{t+1})) + \beta^2\|y_{t+1} - x_{t+1}\|^2. \tag{2}$$

Since $\eta = \frac{1}{\beta}$,

$$
\begin{aligned}
f(x_t) - f(x^*) &\le \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right) + f(x_t) - f(x_{t+1}) + \frac{\beta}{2}\|y_{t+1} - x_{t+1}\|^2 \\
&\le \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) + f(x_t) - f(x_{t+1})
\end{aligned}
$$

where the last inequality uses Equation (1). Then we have

$$f(x_{t+1}) - f(x^*) \le \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right),$$

and the analysis is exactly the same as before.

2.3 f is α-strongly convex and β-smooth

By α-strongly convexity,

$$
\begin{aligned}
f(x_t) - f(x^*) \le& \nabla f(x_t)^\top (x_t - x^*) - \frac{\alpha}{2}\|x_t - x^*\|^2 \\
=& \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(x_t)\|^2 - \frac{\alpha}{2}\|x_t - x^*\|^2 \\
\le& \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) + f(x_t) - f(x_{t+1}) - \frac{\alpha}{2}\|x_t - x^*\|^2
\end{aligned}
$$

where the last inequality uses Equations (2) and (1), and the fact that $\eta = \frac{1}{\beta}$. This is equivalent to

$$f(x_{t+1}) - f(x^*) \le \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) - \frac{\alpha}{2}\|x_t - x^*\|^2, \tag{3}$$

Since $f(x_{t+1}) \ge f(x^*)$, the above inequality implies

$$\|x_{t+1} - x^*\|^2 \le \left(1 - \frac{\alpha}{\beta}\right)\|x_t - x^*\|^2,$$

and thus

$$\|x_t - x^*\|^2 \le \left(1 - \frac{\alpha}{\beta}\right)^t \|x_0 - x^*\|^2, \tag{4}$$

for all $t = 0, 1, \ldots, T$. Next, applying (3) to $t = T-1$, we get

$$\begin{aligned} f(x_T) - f(x^*) &\leq \frac{1}{2\eta}\left(\|x_{T-1} - x^*\|^2 - \|x_T - x^*\|^2\right) - \frac{\alpha}{2}\|x_{T-1} - x^*\|^2 \\ &\leq \frac{\beta - \alpha}{2}\|x_{T-1} - x^*\|^2 \\ &\leq \frac{\beta - \alpha}{2}\left(1 - \frac{\alpha}{\beta}\right)^{T-1}\|x_0 - x^*\|^2 \\ &= \frac{\beta}{2}\left(1 - \frac{\alpha}{\beta}\right)^{T}\|x_0 - x^*\|^2. \end{aligned}$$

The second inequality follows by using $\eta = \frac{1}{\beta}$ and dropping the non-positive term $-\frac{\beta}{2}\|x_T - x^*\|^2$. The third inequality follows from (4). Setting $D := \|x_0 - x^*\|$, exactly as in the unconstrained case, after

$$\begin{aligned} T &= \frac{\log\left(\frac{2\epsilon}{D^2\beta}\right)}{-\log\left(1 - \frac{\alpha}{\beta}\right)} \\ &\approx \frac{\beta}{\alpha}\log\left(\frac{\beta D^2}{2\epsilon}\right), \quad \text{since } \log(1-x) \approx -x \end{aligned} \tag{5}$$

iterations we can achieve ϵ-sub-optimality.