Mathematical Foundations of Machine Learning(CS 4783/5783)

Lecture 12: Stochastic Gradient Descent

1 Stochastic Optimization

One of the practical advantages of online learning methods is that they are simple and computationally efficient, and can be used for statistical learning. The stochastic gradient descent algorithm is closely related to online gradient descent method. In this lecture, we will see how online gradient descent can be used for statistical learning. The single pass SGD algorithm can be described as follows. Given sample $S = (x_1, y_1), \ldots, (x_n, y_n)$ drawn iid from fixed distribution $\mathbf{D}$, the idea is to simply go over the sample one at a time from $t = 1$ to n as if it were produced by an adversary and use the algorithm:

$$\mathbf{f}_{t+1} = \mathbf{f}_t - \eta \nabla \ell(\mathbf{f}_t; (x_t, y_t))$$

We will simply use $\mathcal{F}$ to be all of $\mathbb{R}^d$ for this lecture. Throughout we also make the assumption that for any $\mathbf{f}$:

$$\mathbb{E}_{(x,y)\sim\mathbf{D}} \|\nabla \ell(\mathbf{f}; (x, y)) - \nabla L_{\mathbf{D}}(\mathbf{f})\|_2^2 \leq \sigma^2 \tag{1}$$

where $L_{\mathbf{D}}(\mathbf{f}) = \mathbb{E}_{(x,y)\sim\mathbf{D}} [\ell(\mathbf{f}; (x, y))]$

2 Convex Lipschitz Problems

The first problem we consider is one where $L_{\mathbf{D}}$ is a convex function and is L-Lipschitz, that is, for any $\mathbf{f}, \mathbf{f}'$,

$$|L_{\mathbf{D}}(\mathbf{f}) - L_{\mathbf{D}}(\mathbf{f}')| \leq L\|\mathbf{f} - \mathbf{f}'\|_2$$

First, using Taylor's theorem, we can conclude that the above Lipschitz property implies that

$$\|\nabla L_{\mathbf{D}}(f)\|_2 \leq L \tag{2}$$

For this problem setting we will now show that the online gradient descent or one pass stochastic gradient algorithm is successful.

Lemma 1. *If the gradient variance condition in. Eq. 1 is satisfied and the risk $L_{\mathbf{D}}$ is convex and L-Lipschitz then, using online gradient descent/one pass SGD we get that for any $\mathbf{f}^*$:*

$$\mathbb{E}_S \left[L_{\mathbf{D}} \left(\frac{1}{n} \sum_{t=1}^{n} \mathbf{f}_t \right) \right] - L_{\mathbf{D}}(f^*) \leq \frac{\|\mathbf{f}_1 - \mathbf{f}^*\|_2 \sqrt{L^2 + \sigma^2}}{\sqrt{n}}$$

Proof. From the proof of online gradient descent we have:

$$\nabla \ell(\mathbf{f}_t; (x_t, y_t))^\top (\mathbf{f}_t - \mathbf{f}^*) \leq \frac{1}{2\eta} \left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2 \right) + \frac{\eta}{2} \|\nabla \ell(\mathbf{f}_t, (x_t, y_t))\|_2^2$$

Now note that $\mathbf{f}_t$ only depends on samples $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ and so

$$\mathbb{E}_{(x_t, y_t)} \left[\nabla \ell(\mathbf{f}_t; (x_t, y_t))^\top (\mathbf{f}_t - \mathbf{f}^*) \right] = \mathbb{E}_{(x_t, y_t)} \left[\nabla \ell(\mathbf{f}_t; (x_t, y_t)) \right]^\top (\mathbf{f}_t - \mathbf{f}^*) = \nabla L_{\mathbf{D}}(\mathbf{f}_t)^\top (\mathbf{f}_t - \mathbf{f}^*)$$

Using this in the above by taking expectation over x_t, y_t on both sides we get that:

$$\nabla L_{\mathbf{D}}(\mathbf{f}_t)^\top (\mathbf{f}_t - \mathbf{f}^*) \leq \frac{1}{2\eta} \left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t, y_t)} \left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2 \right] \right) + \frac{\eta}{2} \mathbb{E}_{(x_t, y_t)} \left[\|\nabla \ell(\mathbf{f}_t, (x_t, y_t))\|_2^2 \right]$$

Now convexity of $L_{\mathbf{D}}$ implies that $\nabla L_{\mathbf{D}}(\mathbf{f}_t)^\top (\mathbf{f}_t - \mathbf{f}^*) \geq L_{\mathbf{D}}(\mathbf{f}_t) - L_{\mathbf{D}}(\mathbf{f}^*)$ and so using this,

$$\begin{aligned}
&L_{\mathbf{D}}(\mathbf{f}_t) - L_{\mathbf{D}}(\mathbf{f}^*) \\
&\quad \leq \frac{1}{2\eta} \left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t, y_t)} \left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2 \right] \right) + \frac{\eta}{2} \mathbb{E}_{(x_t, y_t)} \left[\|\nabla \ell(\mathbf{f}_t, (x_t, y_t))\|_2^2 \right] \\
&\quad = \frac{1}{2\eta} \left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t, y_t)} \left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2 \right] \right) + \frac{\eta}{2} \mathbb{E}_{(x_t, y_t)} \left[\|\nabla \ell(\mathbf{f}_t, (x_t, y_t)) - \nabla L_{\mathbf{D}}(\mathbf{f}_t) + \nabla L_{\mathbf{D}}(\mathbf{f}_t)\|_2^2 \right] \\
&\quad = \frac{1}{2\eta} \left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t, y_t)} \left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2 \right] \right) \\
&\qquad + \frac{\eta}{2} \mathbb{E}_{(x_t, y_t)} \left[\|\nabla \ell(\mathbf{f}_t, (x_t, y_t)) - \nabla L_{\mathbf{D}}(\mathbf{f}_t)\|_2^2 + \|\nabla L_{\mathbf{D}}(\mathbf{f}_t)\|_2^2 + 2(\nabla \ell(\mathbf{f}_t, (x_t, y_t)) - \nabla L_{\mathbf{D}}(\mathbf{f}_t))^\top \nabla L_{\mathbf{D}}(\mathbf{f}_t) \right] \\
&\quad = \frac{1}{2\eta} \left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t, y_t)} \left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2 \right] \right) + \frac{\eta}{2} \left(\mathbb{E}_{(x_t, y_t)} \left[\|\nabla \ell(\mathbf{f}_t, (x_t, y_t)) - \nabla L_{\mathbf{D}}(\mathbf{f}_t)\|_2^2 \right] + \|\nabla L_{\mathbf{D}}(\mathbf{f}_t)\|_2^2 \right)
\end{aligned}$$

where the last equality holds because the cross term is 0 in expectation. That is,
$\mathbb{E}_{(x_t, y_t)} \left[(\nabla \ell(\mathbf{f}_t, (x_t, y_t)) - L_{\mathbf{D}}(\mathbf{f}_t))^\top \nabla L_{\mathbf{D}}(\mathbf{f}_t) \right] = 0$ Hence using the variance bound we can conclude that:

$$L_{\mathbf{D}}(\mathbf{f}_t) - L_{\mathbf{D}}(\mathbf{f}^*) \leq \frac{1}{2\eta} \left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t, y_t)} \left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2 \right] \right) + \frac{\eta}{2} \left(\sigma^2 + \|\nabla L_{\mathbf{D}}(\mathbf{f}_t)\|_2^2 \right) \tag{3}$$

At this point, use the fact that since $L_{\mathbf{D}}$ is L-Lipschitz, $\|\nabla L_{\mathbf{D}}(f)\|_2 \leq L$ and so,

$$L_{\mathbf{D}}(\mathbf{f}_t) - L_{\mathbf{D}}(\mathbf{f}^*) \leq \frac{1}{2\eta} \left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t, y_t)} \left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2 \right] \right) + \frac{\eta}{2} \left(\sigma^2 + L^2 \right)$$

Taking expectation over entire sample on both sides and averaging over t we conclude that:

$$\begin{aligned}
\mathbb{E}_S \left[\frac{1}{n} \sum_{t=1}^{n} L_{\mathbf{D}}(\mathbf{f}_t) \right] - L_{\mathbf{D}}(\mathbf{f}^*) &\leq \frac{1}{2\eta n} \sum_{t=1}^{n} \left(\mathbb{E}_S \left[\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 \right] - \mathbb{E}_S \left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2 \right] \right) + \frac{\eta}{2} \left(\sigma^2 + L^2 \right) \\
&= \frac{1}{2\eta n} \left(\mathbb{E}_S \left[\|\mathbf{f}_1 - \mathbf{f}^*\|_2^2 \right] - \mathbb{E}_S \left[\|\mathbf{f}_{n+1} - \mathbf{f}^*\|_2^2 \right] \right) + \frac{\eta}{2} \left(\sigma^2 + L^2 \right) \\
&\leq \frac{\|\mathbf{f}_1 - \mathbf{f}^*\|_2^2}{2\eta n} + \frac{\eta}{2} \left(\sigma^2 + L^2 \right)
\end{aligned}$$

Choosing $\eta = \frac{\|\mathbf{f}_1 - \mathbf{f}^*\|_2}{\sqrt{\sigma^2 + L^2} \sqrt{n}}$ we can conclude the result. Note that we assume we know $\|\mathbf{f}_1 - \mathbf{f}^*\|_2$, if we only know an upper bound on this distance, we can use that for setting η and replace $\|\mathbf{f}_1 - \mathbf{f}^*\|_2$ by the upper bound. Finally, since $L_{\mathbf{D}}$ is convex, we can use Jensen's inequality to push the average over iterations inside to conclude the final statement. □

3 Smooth Convex Problems

In this section we will still assume that $L_{\mathbf{D}}$ is convex. But instead of assuming it is Lipschitz, we assume that $L_{\mathbf{D}}$ is H-smooth. That is, we assume that the gradient of $L_{\mathbf{D}}$ is H-Lipschitz. That is:

$$\|\nabla L_{\mathbf{D}}(\mathbf{f}) - \nabla L_{\mathbf{D}}(\mathbf{f}')\|_2 \leq H\|\mathbf{f} - \mathbf{f}'\|_2$$

For instance, when the domain is unbounded, square loss while smooth is not Lipschitz and so such a setting is useful to study.

Before we proceed to the bound that we can get in such a case for SGD, let us first prove a useful inequality about smooth functions.

Proposition 2. *Let $G : \mathbb{R}^d \mapsto \mathbb{R}$ be a H-smooth function. In this case, for any $\mathbf{f} \in \mathbb{R}^d$*

$$\|\nabla G(\mathbf{f})\|_2^2 \leq 2H\left(G(\mathbf{f}) - \min_{\mathbf{f}} G(\mathbf{f})\right)$$

Proof. Note that by fundamental theorem of calculus, for any function $g : \mathbb{R} \mapsto \mathbb{R}$, $g(1) - g(0) = \int_0^1 g'(t)dt$. Now given $\mathbf{f}, \mathbf{f}'$ let $g(t) = G(\mathbf{f} + t(\mathbf{f}' - \mathbf{f}))$, using fundamental theorem of calculus,

$$G(\mathbf{f}') - G(\mathbf{f}) = g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 \frac{d}{dt} G(\mathbf{f} + t(\mathbf{f}' - \mathbf{f}))dt = \int_0^1 \nabla G(\mathbf{f} + t(\mathbf{f}' - \mathbf{f}))^\top (\mathbf{f}' - \mathbf{f})dt$$

Next we add and subtract $\nabla G(\mathbf{f}')$ inside the integral to get:

$$\begin{aligned}
G(\mathbf{f}') - G(\mathbf{f}) &= \int_0^1 \left(\nabla G(\mathbf{f} + t(\mathbf{f}' - \mathbf{f})) - \nabla G(\mathbf{f})\right)^\top (\mathbf{f}' - \mathbf{f})dt + \int_0^1 \nabla G(\mathbf{f})^\top (\mathbf{f}' - \mathbf{f})dt \\
&= \int_0^1 \left(\nabla G(\mathbf{f} + t(\mathbf{f}' - \mathbf{f})) - \nabla G(\mathbf{f})\right)^\top (\mathbf{f}' - \mathbf{f})dt + \nabla G(\mathbf{f})^\top (\mathbf{f}' - \mathbf{f}) \\
&\leq \int_0^1 \left\|\nabla G(\mathbf{f} + t(\mathbf{f}' - \mathbf{f})) - \nabla G(\mathbf{f})\right\|_2 \ \|\mathbf{f} - \mathbf{f}'\|_2 dt + \nabla G(\mathbf{f})^\top (\mathbf{f}' - \mathbf{f})
\end{aligned}$$

Using H-smoothness,

$$\begin{aligned}
&\leq \int_0^1 tH \left\|\mathbf{f} - \mathbf{f}'\right\|_2 \ \|\mathbf{f} - \mathbf{f}'\|_2 dt + \nabla G(\mathbf{f})^\top (\mathbf{f}' - \mathbf{f}) \\
&= \frac{H}{2} \left\|\mathbf{f} - \mathbf{f}'\right\|_2^2 + \nabla G(\mathbf{f})^\top (\mathbf{f}' - \mathbf{f})
\end{aligned}$$

Setting $\mathbf{f}' = \mathbf{f} - \frac{\nabla G(\mathbf{f})}{H}$ we get:

$$G(\mathbf{f}') - G(\mathbf{f}) \leq \frac{H}{2} \frac{\|\nabla G(\mathbf{f})\|_2^2}{H^2} - \frac{1}{H}\|\nabla G(\mathbf{f})\|_2^2$$

Rearranging we conclude that:

$$\|\nabla G(\mathbf{f})\|_2^2 \leq 2H(G(\mathbf{f}) - G(\mathbf{f}')) \leq 2H(G(\mathbf{f}) - \min_{\mathbf{f}} G(\mathbf{f}))$$

□

Now given the above proposition we are ready to prove a bound for SGD.

Lemma 3. *If the gradient variance condition in. Eq. 1 is satisfied and the risk $L_{\mathbf{D}}$ is convex and H-smooth, then, using online gradient descent/one pass SGD we get that for any $\mathbf{f}^*$:*

$$\mathbb{E}_S\left[L_{\mathbf{D}}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{f}_t\right)\right] - \min_{\mathbf{f}} L_{\mathbf{D}}(\mathbf{f}) \leq \frac{2H\|\mathbf{f}_1 - \mathbf{f}^*\|_2^2}{n} + \frac{2\sigma\ \|\mathbf{f}_1 - \mathbf{f}^*\|_2}{\sqrt{n}}$$

Proof. We start the proof from Eq. 3 to have:

$$L_{\mathbf{D}}(\mathbf{f}_t) - \min_{\mathbf{f}} L_{\mathbf{D}}(\mathbf{f}) \leq \frac{1}{2\eta}\left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t,y_t)}\left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2\right]\right) + \frac{\eta}{2}\left(\sigma^2 + \|\nabla L_{\mathbf{D}}(\mathbf{f}_t)\|_2^2\right)$$

Now using Proposition 2 on $L_{\mathbf{D}}$ we can conclude that:

$$\|\nabla L_{\mathbf{D}}(\mathbf{f}_t)\|_2^2 \leq 2H\left(L_{\mathbf{D}}(\mathbf{f}_t) - \min_{\mathbf{f}} L_{\mathbf{D}}(\mathbf{f})\right)$$

and using this above we get,

$$L_{\mathbf{D}}(\mathbf{f}_t) - \min_{\mathbf{f}} L_{\mathbf{D}}(\mathbf{f}) \leq \frac{1}{2\eta}\left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t,y_t)}\left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2\right]\right) + \frac{\eta}{2}\sigma^2 + \eta H\left(L_{\mathbf{D}}(\mathbf{f}_t) - \min_{\mathbf{f}} L_{\mathbf{D}}(\mathbf{f})\right)$$

Hence we conclude that

$$(1 - \eta H)\left(L_{\mathbf{D}}(\mathbf{f}_t) - \min_{\mathbf{f}} L_{\mathbf{D}}(\mathbf{f})\right) \leq \frac{1}{2\eta}\left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t,y_t)}\left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2\right]\right) + \frac{\eta}{2}\sigma^2$$

We will set η such that $\eta H < \frac{1}{2}$. Hence, we have that:

$$L_{\mathbf{D}}(\mathbf{f}_t) - \min_{\mathbf{f}} L_{\mathbf{D}}(\mathbf{f}) \leq \frac{1}{\eta}\left(\|\mathbf{f}_t - \mathbf{f}^*\|_2^2 - \mathbb{E}_{(x_t,y_t)}\left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2\right]\right) + \eta\sigma^2$$

Taking average over t and expectation over sample we have:

$$\begin{aligned}
\mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n} L_{\mathbf{D}}(\mathbf{f}_t)\right] - \min_{\mathbf{f}} L_{\mathbf{D}}(\mathbf{f}) &\leq \frac{1}{\eta}\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}_S\left[\|\mathbf{f}_t - \mathbf{f}^*\|_2^2\right] - \mathbb{E}_S\left[\|\mathbf{f}_{t+1} - \mathbf{f}^*\|_2^2\right]\right) + \eta\sigma^2 \\
&= \frac{1}{\eta n}\left(\mathbb{E}_S\left[\|\mathbf{f}_1 - \mathbf{f}^*\|_2^2\right] - \mathbb{E}_S\left[\|\mathbf{f}_{n+1} - \mathbf{f}^*\|_2^2\right]\right) + \eta\sigma^2 \\
&= \frac{1}{\eta n}\|\mathbf{f}_1 - \mathbf{f}^*\|_2^2 + \eta\sigma^2
\end{aligned}$$

Now we simply set $\eta = \min\{\frac{1}{2H}, \frac{\|\mathbf{f}_1 - \mathbf{f}^*\|_2}{\sigma\sqrt{n}}\}$. This gives us the bound.

$$\mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n} L_{\mathbf{D}}(\mathbf{f}_t)\right] - \min_{\mathbf{f}} L_{\mathbf{D}}(\mathbf{f}) \leq \frac{2H\|\mathbf{f}_1 - \mathbf{f}^*\|_2^2}{n} + \frac{2\sigma\ \|\mathbf{f}_1 - \mathbf{f}^*\|_2}{\sqrt{n}}$$

Just as in previous section, using Jensen we can move the average inside $L_{\mathbf{D}}$ to obtain the final statement. □

Remark 3.1. *In both the sections, notice that beyond the variance bound, we only made assumptions like convexity and smoothness or Lipschitzness on $L_{\mathbf{D}}$ and made no other assumptions on ℓ itself. Also note that convexity in both the sections can be relaxed to so called one point convexity around $\mathbf{f}^*$ or in other words, we just need that $L_{\mathbf{D}}(\mathbf{f}_t) - L_{\mathbf{D}}(\mathbf{f}^*) \leq \nabla L_{\mathbf{D}}(\mathbf{f}_t)^\top(\mathbf{f}_t - \mathbf{f}^*)$ hold when we compare loss of any $\mathbf{f}_t$ to that of $\mathbf{f}^*$ and not for all pairs which can be a significantly weaker assumption than convexity that seems to hold in many applications at least locally.*