COS-511: Learning Theory **Spring 2017**

Lecture 14: SGD Analysis

Lecturer: Roi Livni

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

14.1 Proof of Thm. 13.3

Before proving Thm. 13.3, we prove a Lemma, that will play a major role in the next lectures, when we begin our discussion on Online Learning.

Lemma 14.1. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_T$ be an arbitrary sequence of vectors (perhaps random) such that $\mathbb{E}\left[\|\mathbf{v}_t\|^2\right] \leq G^2$. Let $\mathcal{K}$ be a convex set of diameter D. Set:*

$$\mathbf{y}_{t+1} = \mathbf{x}^{(t)} - \eta_t \mathbf{v}_t$$
$$\mathbf{x}^{(t+1)} = \Pi_{\mathcal{K}}(\mathbf{y}_{t+1})$$

Then

$$\sum \mathbf{v}_t \cdot \mathbf{x}^{(t)} - \min_{\mathbf{x}^* \in \mathcal{K}} \mathbf{v}_t \cdot \mathbf{x}^* \leq DG\sqrt{T}$$

Proof. We first upper bound $\mathbf{v}_t \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*)$ using the update rule for $\mathbf{x}^{(t+1)}$

$$\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 = \|\Pi_{\mathcal{K}}\left(\mathbf{x}^{(t)} - \eta_t \mathbf{v}_t - \mathbf{x}^*\right)\|^2 \leq \|\mathbf{x}^{(t)} - \eta_t \mathbf{v}_t - \mathbf{x}^*\|^2 \tag{14.1}$$

Hence,

$$\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \leq \|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2 + \eta_t^2 \|\mathbf{v}_t\|^2 - 2\eta_t \mathbf{v}_t \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*)$$

Rearranging terms we get:

$$2\,\mathbb{E}\left[\mathbf{v}_t \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*)\right] \leq \mathbb{E}\left[\frac{\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{(t+1)} - \mathbf{x}^*\|^2}{\eta_t}\right] + \eta_t G^2 \tag{14.2}$$

Summing Eq. 14.1 and Eq. 14.2 from $t = 1$ to T, and setting $\eta_t = \frac{D}{G\sqrt{T}}$:

$$\begin{aligned}
\mathbb{E}\left[\sum \mathbf{v}_t \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*)\right] &\leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{\|\mathbf{x}^{(t)} - \mathbf{x}^*\| - \|\mathbf{x}^{(t+1)} - \mathbf{x}^*\|^2}{2\eta_t}\right] + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2\right]\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq D^2 \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^{T} \eta_t \quad \text{Telescoping Series} \\
&\leq D^2 \frac{1}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \\
&\leq 3DG\sqrt{T}
\end{aligned}$$

□

Proof of Thm. 13.3 By convexity of f we have that

$$f(\bar{\mathbf{x}^{(t)}}) - f(\mathbf{x}^*) \leq \frac{1}{T}\sum f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{1}{T}\sum \nabla f(\mathbf{x}^{(t)}) \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*)$$

Taking expectation we obtain that

$$\mathbb{E}\left[f(\bar{\mathbf{x}^{(t)}}) - f(\mathbf{x}^*)\right] \leq \frac{1}{T}\sum \mathbb{E}\left[\nabla f(\mathbf{x}^{(t)}) \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*)\right]$$

Now recall that for **fixed** $\mathbf{x}^{(t)}$ we have that given past observations $\mathbb{E}\left[\hat{\nabla}_t | \hat{\nabla}_{1:t}\right] = \nabla(f(\mathbf{x}^{(t)}))$. Since $\mathbf{x}^{(t)}$ is determined by past events we have that:

$$\mathbb{E}\left[\hat{\nabla}_t \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*) | \hat{\nabla}_{1:t-1}\right] = \mathbb{E}\left[\nabla f(\mathbf{x}^{(t)}) \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*) | \hat{\nabla}_{1:t-1}\right]$$

Over all we obtain:

$$\mathbb{E}\left[f(\bar{\mathbf{x}^{(t)}}) - f(\mathbf{x}^*)\right] \leq \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\hat{\nabla}_{1:t-1}} \mathop{\mathbb{E}}_{\hat{\nabla}_t}\left[\nabla f(\mathbf{x}^{(t)}) \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*) | \hat{\nabla}_{1:t-1}\right] \leq$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\hat{\nabla}_{1:t-1}} \mathop{\mathbb{E}}_{\hat{\nabla}_t}\left[\hat{\nabla}_t \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*) | \hat{\nabla}_{1:t-1}\right] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \hat{\nabla}_t \cdot (\mathbf{x}^{(t)} - \mathbf{x}^*)\right]$$

Applying Lem. 14.1 with $\mathbf{v}_t = \hat{\nabla}_t$ we obtain

$$\mathbb{E}\left[f(\bar{\mathbf{x}^{(t)}}) - f(\mathbf{x}^*)\right] \leq \frac{1}{T} 3DG\sqrt{T} = \frac{3DG}{\sqrt{T}}$$

14.1.1 A glimpse at Online Learning

A striking fact of our proof is that the backbone relied on a bound over the gradients in Lem. 14.1 which is *completely deterministic*!!! Therefore, it is worth asking how far we can take the proof of Thm. 13.3. Suppose now, that instead of trying to minimize a stationary function f, we wish to minimize a sequence of function $f_1, \ldots, f_T$. Namely consider the following setting:

Online Convex Optimization: Probem Setup

At each round t we get to choose a point $\mathbf{x}^{(t)}$ and suffer a loss given by an arbitrary convex function: $f_t(\mathbf{x}^{(t)})$. The objective of the learner would be to minimize the following term (which is called *regret*):

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}^{(t)}) - \min_{\mathbf{x}^* \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^*)$$

Consider a learner who chooses at each step

$$\mathbf{y}_{t+1} = \mathbf{x}^{(t)} - \eta_t \nabla f_t(\mathbf{x}^{(t)})$$
$$\mathbf{x}^{(t+1)} = \Pi_{\mathcal{K}}(\mathbf{y}_{t+1}).$$

This learner is depicted in Alg 1. Then again by convexity we have as before

$$\sum f_t(\mathbf{x}^{(t)}) - f_t(\mathbf{x}^*) \leq \sum \nabla_t(\mathbf{x}^{(t)} - \mathbf{x}^*)$$

Where we denote $\nabla_t = \nabla f_t(\mathbf{x}^{(t)})$. Applying Lem. 14.1 we obtain the following guarantee over the regret

$$\text{Regret}_T \leq 3DG\sqrt{T}$$

We summerize this in the following result:

Theorem 14.2. *Consider the Online Convex Optimization setting. Suppose that at each round t: f_t is a G-Lipschitz convex function and let $\mathcal{K}$ be a set of diameter D. Apply Alg. 1 to the sequence with step size $\eta_t = \frac{DG}{\sqrt{t}}$, then for the output $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}$ we have that:*

$$\textit{Regret}_T = \sum f_t(\mathbf{x}^{(t)}) - \min_{\mathbf{x}^*} \sum f_t(\mathbf{x}^*) = O(DG\sqrt{T})$$

Algorithm 1 Online Gradient Descent

Input: A sequence of arbitrary functions $f_1, \ldots, f_T$ step sizes $\{\eta_t\}$, $\mathbf{x}_1 \in \mathcal{K}$
for $t = 1, 2 \ldots T$ **do**
Observe f_t and suffer cost $f_t(\mathbf{x}^{(t)})$.
Set $\nabla_t = \nabla f_t(\mathbf{x}^{(t)})$.
Update and project:

$$\mathbf{y}_{t+1} = \mathbf{x}^{(t)} - \eta_t \nabla_t$$
$$\mathbf{x}^{(t)} = \Pi_{\mathcal{K}}(\mathbf{y}_{t+1})$$

end for
return $\mathbf{x}_1, \ldots, \mathbf{x}^{(t)}$.

14.1.2 Online Learning

In terms of learning, we can consider a sequential setting where a learner chooses at each iteration t a classifier $\mathbf{w}_t$: Then an adversery outputs an arbitrary point $(\mathbf{x}_t, y_t)$, and the learner suffers loss $\ell(\mathbf{w}_t \cdot \mathbf{x}_t, y_t)$. The performance of the learner is then measured in terms of regret:

$$\text{Regret}_T = \sum_{t=1}^{T} \ell(\mathbf{w}_t \cdot \mathbf{x}_t, y_t) - \min_{\mathbf{w}^* \in \mathcal{K}} \ell(\mathbf{w}^* \cdot \mathbf{x}_t, y_t)$$

By Thm. 14.2 we using the OGD algorithm, we can obtain a regret bound of $O(\sqrt{T})$[1]: This is completely comparable with the bound we have obtained so far Statistical Learning Theory. Namely if $T > \frac{1}{\epsilon^2}$ we obtain that

$$\frac{1}{T}\text{Regret}_T \leq O(\epsilon)$$

[1]Here we neglect all terms that depend on Lipschitness and boundness of the class

In future lecture we will see how to derive the Statistical Learning Result directly from these regret bounds. Note however, that unlike the Statistical setting: an ERM algorithm will tend to fail (as the points are not choosing through some distribution but are adverserially picked). On the other hand, we can still construct algorithms for the online setting that achieve similar guarantees as in the stochastic setting. Moreover, this setting is much more general as it does not assume that the points to be learnt are picked from a stationary distribution.