CS 7545: Machine Learning Theory **Fall 2019**

Lecture 12: More Online Convex Optimization

Lecturer: Jacob Abernethy *Scribes: Jihui Jin & Sihan Zeng*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

12.1 Reminder of Online Convex Optimization Framework

The online convex optimization framework involves the following protocol. Given a convex set $\mathcal{K} \subseteq \mathbb{R}^d$
For $t = 1, ..., T$:

1. Alg selects $x_t \in \mathcal{K}$
2. Nature reveals convex function $f_t : \mathcal{K} \to \mathbb{R}$

Goal: Minimize $\text{Regret}_T = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x)$

Definition 12.1 (projection) *For any $y \in \mathbb{R}^d$, the* ***projection*** *onto a set $\mathcal{K}$ is defined as*

$$\Pi_{\mathcal{K}}(y) = \underset{x \in \mathcal{K}}{\operatorname{argmin}} \|x - y\|_2$$

Lemma 12.2 (Pythagorean Theorem for Bregman Divergences) *For any $x \in \mathcal{K}$,*

$$\|x - \Pi_{\mathcal{K}}(y)\| \leq \|x - y\|$$

, and the equation holds iff $y \in K$.

Proof: (Exercise) ■

12.2 Online Gradient Descent

Assumptions:

- f_t is $G-$Lipschitz: $\|\nabla f_t\| \leq G$
- $\mathcal{K}$ has diameter D: $\|x - y\| \leq D \quad \forall x, y \in \mathcal{K}$

Algorithm 1: Online Gradient Descent

Init $x_1 \in \mathcal{K}$;
for $i \leftarrow 1$ **to** T **do**
  $\eta_t = \frac{D}{G\sqrt{t}}$;
  $\tilde{x}_{t+1} = x_t + \eta_t \nabla f_t(x_t)$;
  $x_{t+1} = \Pi_{\mathcal{K}}(\tilde{x}_{t+1})$;
end

Theorem 12.3 *Assume that $\|\nabla f_t\| \leq G$, $\|x - y\| \leq D$ for all x, y in $\mathcal{K}$. Then*

$$\text{Regret}_T(OGD) \leq \frac{3}{2} GD\sqrt{T}$$

Proof: Let $x^* = \text{argmin}_{x\in\mathcal{K}} \sum_{t=1}^T f_t(x)$. Define $\nabla_t := \nabla f_t(x_t)$
By convexity of f_t at x_t:

$$f(x_t) - f_t(x^*) \le \nabla_t^\top (x_t - x^*)$$

By Lemma 12.2, we have that

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|\Pi_{\mathcal{K}}(x_t - \eta_t\nabla_t) - x^*\|^2 \\
&\le \|x_t - \eta_t\nabla_t - x^*\|^2 \\
&= \|x_t - x^*\|^2 + \eta_t^2\|\nabla_t\|^2 - 2\eta_t\nabla_t^\top(x_t - x^*)
\end{aligned}$$

Combining these two gives that

$$f_t(x_t) - f_t(x^*) \le \frac{\|x_t - x^*\|^2}{2\eta_t} + \frac{\eta_t\|\nabla_t\|^2}{2} - \frac{\|x_{t+1} - x^*\|^2}{2\eta_t}$$

$$\begin{aligned}
\text{Regret}_T &= \sum_{t=1}^T f_t(x_t) - f_t(x^*) \\
&\le \frac{1}{2}\sum_{t=1}^T \left(\frac{\|t_x - x^*\|^2 - \|x_{t+1} - x^*\|^2}{\eta_t}\right) + \frac{G^2}{2}\sum_{t=1}^T \eta_t \\
&\le \frac{1}{2}\sum_{t=1}^T \|x_t - x^*\|^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2}\sum_{t=1}^T \eta_t \\
&\le \frac{D^2}{2}\sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2}\sum_{t=1}^T \eta_t \\
&= \frac{D^2}{2}\left(\frac{1}{\eta_T} - \frac{1}{\eta_0}\right) + \frac{G^2}{2}\sum_{t=1}^T \frac{D}{G\sqrt{t}} \\
&\le \frac{1}{2}DG\sqrt{T} + \frac{G^2}{2}\left(2\frac{D}{G}\sqrt{T}\right) \\
&= \frac{3}{2}DG\sqrt{T}
\end{aligned}$$

Where in the second to last step we make use of the inequality that

$$\sum_{t=1}^T \frac{1}{\sqrt{t}} \le 2\sqrt{T}$$

■

Corollary 12.4 *Gradient descent algorithms for minimizing one convex function f with "averaging" converges at a rate of $O(\frac{DG}{\sqrt{T}})$*

12.3 Stochastic Gradient Descent (SGD)

Let $h(\cdot;\cdot)$ be some convex loss function. Consider the SGD algorithm to minimize $f(x) = \mathop{\mathbb{E}}_{\zeta}[h(x;\zeta)]$, where ζ denotes samples from the distribution.

The steps of the SGD algorithm at each iteration t is as follows:

$$\begin{aligned}&(1)\ \text{sample } \zeta_t\\&(2)\ x_{t+1} \leftarrow x_t - \eta_t \nabla h(x_t;\zeta_t)\\&(3)\ \text{output:} \bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t\end{aligned}$$

Claim 12.5 *SGD converges at* $O(\frac{DG}{\sqrt{T}})$.

Proof:

$$\begin{aligned}\mathop{\mathbb{E}}_{\zeta_{1:T}}[f(\bar{x}_T)] &\le \mathop{\mathbb{E}}_{\zeta}[\frac{1}{T}\sum_{t=1}^{T} f(x_t)] \ \text{ by the convexity of the loss function}\\&= \mathop{\mathbb{E}}_{\zeta}[\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\zeta_{1:t}}[f(x_t)|\zeta_{1:t-1}]] \ \text{ since } x_t \text{ only depends on samples before } t\\&= \mathop{\mathbb{E}}_{\zeta}[\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\zeta_{1:t}}[h(x_t;\zeta_t)|\zeta_{1:t-1}]] \ \text{ using the definition of } f\\&= \mathop{\mathbb{E}}_{\zeta}[\frac{1}{T}\sum_{t=1}^{T} f_t(x_t)] \ \text{ using the fact that } \mathbb{E}[\mathbb{E}[A|B]] = \mathbb{E}[A]\\&\le \mathop{\mathbb{E}}_{\zeta}[\frac{1}{T}\sum_{t=1}^{T} f_t(x^*)] + \frac{Regret_T(OGD)}{T} \ \text{ since we use the OGD protocol to update x at each iteration}\\&= f(x^*) + \frac{3DG}{2\sqrt{T}}\end{aligned}$$

Therefore, the convergence rate $\mathbb{E}[f(\bar{x}_T)] - f(x^*) \le \frac{3DG}{2\sqrt{T}}$.

■