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

Lecture 14: Mirror Descent Continued

Lecturer: Jacob Abernethy *Scribes: Sebastian Bischoff, Jie Zhang*

The algorithm for online convex optimization is

for $t = 1, ..., T$ **do**
  Choose $X_t \in K \subseteq \mathbb{R}^d$;
  Observe convex loss function $f_t(\cdot)$
end

We want to minimize the regret:

$$\sum_{t=1}^{T} f_t(X_t) - \min_{x \in K} \sum_{t=1}^{T} f_t(x)$$

14.1 Online Mirror Descent Algorithm

Given some convex regularizer $R(\cdot)$
Need $K \subseteq \text{int}(\text{dom}(R))$
Assume $R(\cdot)$ is λ-strongly convex w.r.t $\|\cdot\|$
Pick $X_1 \in K$ arbitrarily

At time t: $X_{t+1} = \arg\min_{X \in K} \eta_t \langle x, \nabla_t \rangle + D_R(X, X_t)$ where $\eta_1, \ldots, \eta_T > 0$ is a sequence of learning rates and $\nabla_t := \nabla f_t(X_t)$

Definition 14.1 (Bregman Divergence) *For $R(\cdot)$ being λ-strongly convex w.r.t $\|\cdot\|$*

$$D_R(x, y) = R(x) - R(y) - \langle \nabla R(y), x - y \rangle$$

Remark 1 (Previous lecture)

$$R(\cdot) \textit{ is } \lambda\textit{-strongly convex w.r.t } \|\cdot\| \iff D_R(x, y) \geq \frac{\lambda}{2} \|x - y\|^2$$

Lemma 14.2 (From Homework 1) *For all $a, b, c \in \text{dom}(R)$ and for any convex differentiable function R*

$$D_R(c, a) + D_R(a, b) - D_R(c, b) = \langle \nabla R(b) - \nabla R(a), c - a \rangle$$

Lemma 14.3 (FOOC) *For x_t, x_{t+1} be chosen via Online Mirror Descent algorithm, then*

$$\langle \nabla R(x_{t+1}) - \nabla R(x_t) + \eta_t \nabla_t, u - x_{t+1} \rangle \geq 0$$

Lemma 14.4 (Covered in last class) *Let v, w be any vectors, $\| \cdot \|, \| \cdot \|_*$ a dual norm pair, then for any $\lambda > 0$, we have*

$$\langle v, w \rangle \leq \frac{\lambda \|v\|^2}{2} + \frac{\|w\|_*^2}{2\lambda}$$

Lemma 14.5 (Big Lemma) *Let x_t, x_{t+1} chosen by Online Mirror Descent. Let $u \in K$, then*

$$\eta_t (f_t(x_t) - f_t(u)) \leq \langle \eta_t \nabla_t, x_t - u \rangle \leq D_R(u, x_t) - D_R(u, x_{t+1}) + \frac{\eta_t^2}{2\lambda} \|\nabla_t\|_*^2$$

Proof: The first inequality is by convexity.
We now show the second inequality.

$$\begin{aligned}\langle \eta_t \nabla_t, x_t - u\rangle &= \langle \nabla R(x_t) - \nabla R(x_{t+1}) - \eta_t \nabla_t, u - x_{t+1}\rangle \\ &\quad - \langle \nabla R(x_t) - \nabla R(x_{t+1}), u - x_{t+1}\rangle + \langle \eta_t \nabla_t, x_t - x_{t+1}\rangle \\ &\leq 0 - (D_R(u, x_{t+1}) + D_R(x_t, x_{t+1}) - D_R(u, x_t)) + \langle \eta_t \nabla_t, x_t - x_{t+1}\rangle \\ &\leq (D_R(u, x_{t+1}) + D_R(x_t, x_{t+1}) - D_R(u, x_t)) + \frac{\lambda \|x_t - x_{t+1}\|^2}{2} + \frac{\eta_t^2 \|\nabla_t\|_*^2}{2\lambda} \\ &\leq D_R(u, x_t) - D_R(x, x_{t+1}) + \frac{\eta_t^2 \|\nabla_t\|_*^2}{2\lambda}\end{aligned}$$

The first inequality is from 14.3:

$$\langle \nabla R(x_t) - \nabla R(x_{t+1}) - \eta_t \nabla_t, u - x_{t+1}\rangle \leq 0$$

The second inequality is from 14.4, with $w = \eta_t \nabla_t$ and $v = x_t - x_{t+1}$. The third inequality is because

$$-D_R(x_t, x_{t+1}) + \frac{\lambda \|x_t - x_{t+1}\|^2}{2} \leq 0$$

by definition of λ strong convexity. ■

Theorem 14.6 *Let $u \in K$ be arbitrary.*
Let $\eta_1, \ldots, \eta_T \geq 0$ be a decreasing sequence.
Let $x_1, \ldots, x_T$ be chosen by online mirror descent.
Let $d = \sqrt{max_{t=1,\ldots,T} D_R(u, x_t)}$, then

$$\sum_{t=1}^{T} (f_t(x_t) - f_t(u)) \leq \frac{d^2}{\eta_T} + \frac{1}{2\lambda} \sum_{t=1}^{T} \eta_T \|\nabla_t\|_*^2$$

Proof: Here's a proof of this big statement. It basically follows from Lemma 14.5.

$$\begin{aligned}\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(u) &\leq \sum_{t=1}^{T} (\frac{1}{\eta_t} D_R(u, x_t) - \frac{1}{\eta_t} D_R(u, x_{t+1}) + \frac{\eta_t}{2\lambda} \|\nabla_t\|_*^2) \\ &= \frac{1}{\eta_1} D_R(u, x_1) - \frac{1}{\eta_T} D_R(u, x_{t+1}) + \sum_{t=1}^{T-1} (\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}) D_R(u, x_{t+1}) + \sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\nabla_t\|_*^2 \\ &\leq \frac{d^2}{\eta_1} + d^2 \sum_{t=1}^{T-1} (\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}) + \sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\nabla_t\|_*^2 \\ &= \frac{d^2}{\eta_1} + \frac{d^2}{\eta_T} - \frac{d^2}{\eta_1} + \sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\nabla_t\|_*^2 \\ &= \frac{d^2}{\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2\lambda} \|\nabla_t\|_*^2\end{aligned}$$

Inequality in first line is from 14.5. Inequality in the third line follows from $D_R(u, x_1) \leq d^2$, $-\frac{1}{\eta_T} D_R(u, x_{t+1}) \leq 0$, $D_R(u, x_{t+1}) \leq d^2$. ■

Corollary 14.7 *Let $\eta_t = \frac{d\sqrt{\lambda}}{\sqrt{\sum_{s=1}^{T} \|\nabla_s\|_*^2}}$, then $Regret_T \leq 2\frac{d}{\sqrt{\lambda}} \sqrt{\sum_{s=1}^{T} \|\nabla_s\|_*^2}$*

Remark 2 *Online Mirror Descent is quite general. For example, we can let K be any bounded convex set, $R(x) = \frac{1}{2}\|x\|^2$, then*

$$argmin_{x \in K} \eta_t \langle x, \nabla_t \rangle + D_R(x, x_t) = \Pi_k(x_t - \eta_t \nabla_t)$$

where the left hand side is online mirror descent and the right hand side is online gradient descent.

Remark 3 *Let $K = \Delta_n$, $R(x) = \sum_{i=1}^n x_i \log x_i$, then*

$$D_R(x, x_t) = \sum_{i=1}^n x(i) \log \frac{x(i)}{x_t(i)}$$

and therefore

$$\arg\min_{x \in \Delta_n} \eta_t \langle x, \nabla_t \rangle + D_R(x, x_t) = \arg\min_{x \in \Delta_n} \eta_t \langle x, \nabla_t \rangle + \sum_{i=1}^n x_i * \log(\frac{x(i)}{x_t(i)}).$$

It follows that

$$(\eta_t \nabla_t)_i = \frac{\log x(i)}{\log x_t(i)}$$

because

$$\nabla_p(\sum p_i * \log \frac{p_i}{q_i}) = \log(\frac{p_i}{q_i}) + p_i \frac{q_i}{p_i} \frac{1}{q_i} = \log \frac{p_i}{q_i} + 1.$$

Why do we choose this regularizer? Entropy is 1-strongly convex w.r.t $\|\cdot\|$ which yields

$$Regret_T \leq \frac{d}{\sqrt{\lambda}} \sqrt{\sum \|\nabla_t\|_*^2}$$