AI-3001: Advanced Topics in Machine Learning **Aug-Nov 2022**

Lecture 8: Introduction to Online Convex Optimization

Lecturer: Ganesh Ghalme *Scribes: Ganesh Ghalme*

8.1 OCO Setting

Algorithm 1: OCO Setting

Input: Action set: $\mathcal{K}$, Function class: $\mathcal{F}$

for $t = 1, 2, \cdots$ **do**

- **Algorithm plays** $x_t \in \mathcal{K}$;
- **Environment reveals** $f_t \in \mathcal{F}$;
- **Algorithm incurs a loss** $f_t(x_t) \in \mathbb{R}$;

end

- We are trying to optimize over a function class hence the regret is given as

$$\mathcal{R}_T(\textsc{Alg}) = \sup_{f_1, f_2, \cdots f_T \in \mathcal{F}} \Big[\sum_{t=1}^{T} f_t(x_t) - \inf_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x) \Big] \tag{8.1}$$

- It is worth noting that if we fix function class with a single function we recover the actions setting.
- In OCO framework we will focus on setting where $\mathcal{K}$ is convex and $\mathcal{F}$ is a collection of convex functions.

8.1.1 In this module (6 lectures)

- Online gradient Descent (OGD) and its regret guarantee
- Regret minimization inspired OCO techniques; FTL, FTRL, FTPL (time permitting)
- Online Mirrored Descent (OMD)
- Connection between FTRL and OMD

8.1.2 Today: OGD and its regret guarantee

Algorithm 2: OGD Algorithm

Input: action set: $\mathcal{K}$, $x_1 \in \mathcal{K}$, step size sequence $\{\eta_t\}$

for $t = 1, 2, \cdots$ **do**

- **Algorithm plays** $x_t \in \mathcal{K}$;
- **Environment reveals** $f_t \in \mathcal{F}$;
- **Algorithm incurs a loss** $f_t(x_t) \in \mathbb{R}$;
- **Update**

$$y_{t+1} = x_t - \eta_t \nabla f_t(x_t) \tag{8.2}$$
$$x_{t+1} = \Pi_{\mathcal{K}}(y_{t+1}) \tag{8.3}$$

end

Theorem 8.1 (Pythagorean Theorem)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a closed, convex set and $\Pi_{\mathcal{K}}(x) \in \mathcal{K}$ denote the projection of point x on $\mathcal{K}$. Then*

$$||\Pi_{\mathcal{K}}(x) - y||^2 \leq ||y - x||^2 \text{ for every } y \in \mathcal{K} \text{ for every } x \in \mathbb{R}^d.$$

Proof. We begin with the following supporting result.

Claim 8.2. *We have*

1. *$\Pi_{\mathcal{K}}(x)$ is a projection of x on set $\mathcal{K}$ iff $\langle x - \Pi_{\mathcal{K}}(x), y - \Pi_{\mathcal{K}}(x)\rangle \leq 0$ for all $y \in \mathcal{K}$, and*
2. *$\Pi_{\mathcal{K}}(x)$ is unique.*

Proof of the claim. Note that $\Pi_{\mathcal{K}}(x) \in \arg\min_{x' \in \mathcal{K}} \frac{1}{2}||x - x'||^2$.
(1) $\Rightarrow$: Let $y \in \mathcal{K}$ be such that $y \neq \Pi_{\mathcal{K}}(x)$. By convexity of the set $\mathcal{K}$ we have that $(1-\alpha)\Pi_{\mathcal{K}}(x) + \alpha y = \Pi_{\mathcal{K}}(x) + \alpha(y - \Pi_{\mathcal{K}}(x)) \in \mathcal{K}$.

$$\begin{aligned}
||x - \Pi_{\mathcal{K}}(x)||^2 &\leq ||x - (\Pi_{\mathcal{K}}(x) + \alpha(y - \Pi_{\mathcal{K}}(x)))||^2 && \text{(By definition of } \Pi_{\mathcal{K}}(x)) \\
&= ||x - \Pi_{\mathcal{K}}(x)||^2 + \alpha^2 ||y - \Pi_{\mathcal{K}}(x)||^2 - 2\alpha\langle x - \Pi_{\mathcal{K}}(x), y - \Pi_{\mathcal{K}}(x)\rangle && (8.4) \\
\implies \langle x - \Pi_{\mathcal{K}}(x), y - \Pi_{\mathcal{K}}(x)\rangle &\leq \frac{\alpha}{2}||y - \Pi_{\mathcal{K}}(x)||^2 && (8.5)
\end{aligned}$$

The above result holds for any value of α. Letting $\alpha \downarrow 0$ we have $\langle x - \Pi_{\mathcal{K}}(x), y - \Pi_{\mathcal{K}}(x)\rangle \leq 0$
(1) $\Leftarrow$: Let $y \in \mathcal{K}$ such that $y \neq \Pi_{\mathcal{K}}(x)$.

$$\begin{aligned}
||x - y||^2 - ||x - \Pi_{\mathcal{K}}(x)||^2 &= ||(x - \Pi_{\mathcal{K}}(x)) + (\Pi_{\mathcal{K}}(x) - y)||^2 - ||x - \Pi_{\mathcal{K}}(x)||^2 \\
&= ||\Pi_{\mathcal{K}}(x) - y||^2 - 2\langle x - \Pi_{\mathcal{K}}(x), y - \Pi_{\mathcal{K}}(x)\rangle \\
&\geq 0 && \text{(since } \langle x - \Pi_{\mathcal{K}}(x), y - \Pi_{\mathcal{K}}(x)\rangle \leq 0.)
\end{aligned}$$

(2) For contradiction, let x_1 and $x_2 \neq x_1$ are two projections of point x on convex set $\mathcal{K}$. From (1) we have

$$\langle x_2 - x_1, x - x_1\rangle \leq 0 \tag{8.6}$$
$$\langle x_1 - x_2, x - x_2\rangle \leq 0 \tag{8.7}$$

From above equation we have $\langle x_2 - x_1, x_2 - x_1\rangle = ||x_2 - x_1||^2 \leq 0$. Which is true only when $x_2 = x_1$, a contradiction. □

We now comlpete the proof of Pythagorean theorem. We have,

$$\begin{aligned}
||y-x||^2 &= ||\Pi_{\mathcal{K}}(x) - x + y - \Pi_{\mathcal{K}}(x)||^2 \\
&= ||\Pi_{\mathcal{K}}(x) - x||^2 + ||\Pi_{\mathcal{K}}(x) - y||^2 - 2\langle x - \Pi_{\mathcal{K}}(x)\rangle\langle y - \Pi_{\mathcal{K}}(x)\rangle \\
&\geq ||\Pi_{\mathcal{K}}(x) - x||^2 + ||\Pi_{\mathcal{K}}(x) - y||^2 && \text{(From above claim)} \\
&\geq ||\Pi_{\mathcal{K}}(x) - y||^2 && \text{(Non-negativity of norm function)}
\end{aligned}$$

□

Let $G := \sup_{f\in\mathcal{F}} \sup_{x\in\mathcal{K}} ||\nabla f(x)||$ and $D := \sup_{x,x'\in\mathcal{K}} ||x - x'||$.

Theorem 8.3. *The OGD algorithm with step size $\eta_t = \frac{D}{G\sqrt{t}}$ gives the following upper bound on regret*

$$\mathcal{R}_T(OGD) \leq \frac{3}{2} GD\sqrt{T} \tag{8.8}$$

Proof. Let $x^* = \arg\min_{x\in\mathcal{K}} \sum_{t=1}^T f_t(x)$ and $\nabla_t = \nabla f_t(x_t)$. By convexity of f_t we have

$$f_t(x_t) - f_t(x^*) \leq \nabla_t^T (x_t - x^*)$$

We now upper bound the RHS.

$$||x_{t+1} - x^*||^2 = ||\Pi_{\mathcal{K}}(x_t - \eta_t\nabla_t) - x^*||^2 \leq ||x_t - \eta_t\nabla_t - x^*||^2 = ||x_t - x^*||^2 + \eta_t^2||\nabla_t||^2 - 2\eta_t\langle\nabla_t, x_t - x^*\rangle$$

$$\implies 2\nabla_t^T(x_t - x^*)\rangle \leq \frac{||x_t - x^*||^2 - ||x_{t+1} - x^*||^2}{\eta_t} + \eta_t||\nabla_t||^2 \tag{8.9}$$

$$\begin{aligned}
\implies 2\mathcal{R}_T(OGD) &= \sum_{t=1}^T 2(f_t(x_t) - f_t(x^*)) \leq \sum_{t=1}^T \frac{||x_t - x^*||^2 - ||x_{t+1} - x^*||^2}{\eta_t} + G^2 \sum_{t=1}^T \eta_t \\
&\leq D^2 \sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^T \eta_t \\
&\leq D^2 \sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^T \eta_t - \underbrace{\frac{||x_{T+1} - x^*||^2}{\eta_T}}_{\geq 0} \\
&\leq \frac{D^2}{\eta_T} + G^2 \sum_{t=1}^T \eta_t \\
&= DG\sqrt{T} + DG \sum_{t=1}^T \frac{1}{\sqrt{t}} && \text{(substituting } \eta_t = \tfrac{D}{G\sqrt{t}}) \\
&\leq 3DG\sqrt{T}
\end{aligned}$$

□

Theorem 8.4. *If the loss function is α-strongly convex then OGD algorithm with $\eta_t = \frac{1}{\alpha t}$ gives the following regret guarantee,*

$$\mathcal{R}_T(OGD) \leq \frac{G^2}{2\alpha}(1 + \log(T)). \tag{8.10}$$

Proof. From α-strong convexity property we have

$$2(f_t(x_t) - f_t(x^*)) \leq 2\nabla_t^T(x_t - x^*) - \alpha||x_t - x^*||^2$$

We will use the upper bound on the first term from previous theorem (obtained in Eq. 8.9). We have

$$\begin{aligned}
2\mathcal{R}_T(OGD) = \sum_{t=1}^{T} 2(f_t(x_t) - f_t(x^*)) &\leq \sum_{t=1}^{T}\Big(\frac{||x_t - x^*||^2 - ||x_{t+1} - x^*||^2}{\eta_t} - \alpha||x_t - x^*||^2\Big) + G^2\sum_{t=1}^{T}\eta_t - \frac{||x_{T+1} - x^*||^2}{\eta_T} \\
&\leq D^2 \underbrace{\sum_{t=1}^{T}\Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}} - \alpha\Big)}_{=0} + G^2\sum_{t=1}^{T}\eta_t \\
&= \frac{G^2}{\alpha}\sum_{t=1}^{T}\frac{1}{t} \\
&\leq \frac{G^2}{\alpha}(1 + \log(T))
\end{aligned}$$

$\square$

References

[Bub11] Sébastien Bubeck. Introduction to online optimization. *Lecture Notes*, 2, 2011.

[Haz19] Elad Hazan. Introduction to online convex optimization, 2019.

[SS+12] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.