CS861: Theoretical Foundations of Machine Learning	Lecture 23 - 10/28/2024
University of Wisconsin–Madison, Fall 2024	

Lecture 23: Online Convex Optimization

Lecturer: Kirthevasan Kandasamy — *Scribed by: Hongyi Huang, Shuqi Bi*

In this lecture, we will introduce online convex optimization. We will first introduce the basic concepts about convexity and then move to two motivating examples, the online linear classification, and the expert problem, and give a `unified framework` for online convex optimization. Then, we will discuss two methods, `Follow the Leader`(FTL), and `Follow the Regularized Leader`(FTRL). Finally, we will use several examples to show how to `choose the regularizer`.

1 Convexity review

Definition 1 (Convex function)**.** *We will present two equivalent definitions of convex functions:*

(i.) *A function $f : \Omega \to \mathbb{R}$ is convex if Ω is a convex set for $\forall\ \alpha \in [0, 1]$, and $\forall\ u, v \in \Omega$, we have:*

$$f(\alpha u + (1 - \alpha)v) \leq \alpha f(u) + (1 - \alpha)f(v).$$

(ii.) *Equivalently f is convex if $\forall\ w \in \Omega, \exists\ g \in \mathbb{R}^n, s.t., \forall\ w' \in \Omega$, we have:*

$$f(w') \geq f(w) + g^T(w' - w)$$

Definition 2 (Sub-gradients and sub-differential)**.** *We will present the definition for sub-gradient and sub-differential.*

(i.) *Any $g \in \mathbb{R}^n$ which satisfies (ii) in the above definition is called a subgradient of f at w.*

(ii.) *The set of subgradient of f at w are called sub-differential and denote $\partial f(w)$.*

Remark Some useful facts about sub-gradients:

(i.) If f is differentiable, $\partial f(w) = \{\nabla f(w)\}$.

(ii.) $0 \in \partial f(w) \Leftrightarrow w \in \underset{w\in\Omega}{\arg\min} f(w)$.

(iii.) For finite-valued convex functions[1] (f_1, f_2) and positive scalars (α_1, α_2), if $g_1 \in \partial f_1(w)$ and $g_2 \in \partial f_2(w)$, then $\alpha_1 g_1 + \alpha_2 g_2 \in \partial h(w)$, where $h = \alpha_1 g_1 + \alpha_2 g_2$.

Definition 3 (Strong Convexity)**.** *A convex function $f : \Omega \to \mathbb{R}$ is α-strongly convex in some norm $||\cdot||$, if $f(w') \geq f(w) + g^T(w' - w) + \frac{\alpha}{2}||w' - w||^2$, $\forall g \in \partial f(w)$.*

Example 1. Some examples of strongly-convex functions:

(i.) $f(w) = \frac{1}{2}||w||_2^2$ is 1-strongly convex is $||\cdot||_2$.

[1] Refer to Theorem 8.11 here: `https://people.eecs.berkeley.edu/~brecht/opt4ml_book/O4MD_08_Subgradients.pdf`

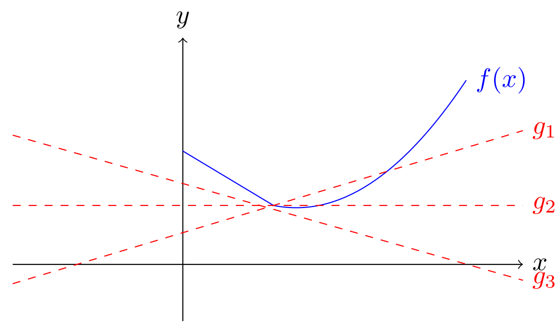

Figure 1: The blue curve above depicts a convex function, f, whereas the red lines denote the first-order linear underestimators to this function f. Additionally, these linear underestimators also serve as three of the uncountably-infinite possible subgradients at the non-different point $((0.8, 0.52))$ on function f.

(ii.) The negative entropy $f(w) = \sum_{i=1}^{d} w(i) \log(w(i))$ is 1-strongly convex in $||\cdot||_1$, when $\Omega = \Delta^d$.

Proof of (ii)

As f is differentiable, $\partial f(\omega) = \{\nabla f(\omega)\}$. Therefore, we need to show, for all $\omega, \omega' \in \Delta([d])$, we have

$$f(\omega') \geq f(\omega) + \nabla f(\omega)^\top(\omega' - \omega) + \frac{1}{2}\|\omega' - \omega\|_1^2.$$

$$\iff f(\omega') - f(\omega) - \nabla f(\omega)^\top(\omega' - \omega) \geq \frac{1}{2}\|\omega' - \omega\|_1^2.$$

Note that

$$\frac{\partial f \omega(i)}{\partial \omega(i)} = 1 + \log(p(i)).$$

Therefore,

$$\text{LHS} = \sum_{i=1}^{K} \omega'(i) \log(\omega'(i)) - \sum_{i=1}^{K} \omega(i) \log(\omega(i)) - \sum_{i=1}^{K}(1 + \log(\omega(i)))(\omega'(i) - \omega(i))$$

$$= \sum_{i=1}^{K} \omega'(i) \log\left(\frac{\omega'(i)}{\omega(i)}\right)$$

$$= KL(\omega', \omega) \geq \frac{1}{2}\|\omega' - \omega\|_1^2.$$

The last step follows by Pinsker's inequality,

$$KL(P,Q) \geq 2TV(P,Q)^2 = 2\left(\frac{1}{2}\|P - Q\|_1\right)^2.$$

□

Remark Some remarks and properties of strongly-convex functions:

(i.) If f is strongly convex in $||\cdot||_2$, then this is equivalent to saying that $f(w) - \frac{\alpha}{2}||w||_2^2$ is convex. In other words, f is 'at least as convex as a quadratic function'.

(ii.) If f is α-strongly convex and f_2 is convex, then $\beta f_1 + f_2$ is $(\beta\alpha)$-strongly convex $\forall \beta > 0$.

(iii.) Let $w^* = \arg\min_{w \in \Omega} f(w)$, where f is α-strongly convex. Then $f(w) \geq f(w^*) + \frac{\alpha}{2}||w - w^*||^2$. The proof uses the definition of strong convexity and the fact that $0 \in \partial f(w^*)$.

Definition 4 (Dual norm)**.** *Given a norm $||\cdot||$, its dual norm $||\cdot||_*$ is defined as:*

$$||w||_* = \max_{||u||\leq 1} u^T w$$

Example 2. Some examples of dual-norm pairs:

(i.) $(||\cdot||_2, ||\cdot||_2)$

(ii.) $(||\cdot||_1, ||\cdot||_\infty)$

(iii.) More generally, the following are also dual-norm pairs when considering ℓ_α-norms ($\alpha > 0$):

$$(||\cdot||_p, ||\cdot||_q), \text{ where } p, q > 0 \text{ and } \frac{1}{p} + \frac{1}{q} = 1$$

Lemma 1 (Hölder's inequality)**.** *$\forall a, b \in \mathbb{R}^d$, $a^T b \leq ||a|| \cdot ||b||_*$.*

2 Examples and Unified Framework

We will first present two examples to show what is online convex optimization. The online linear classification, and the expert problem.

Example 3 (online linear classification)**.** Let $\Theta \left\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\right\}$. On each round, the learner chooses some $\theta_t \in \Theta$. Simultaneously, the environment picks an instance $\{x_t, y_t\} \in \mathcal{X} \times \mathcal{Y}$ where the domain $\mathcal{X} \in \mathbb{R}^d, \mathcal{Y} = \{+1, -1\}$. Then, the learner incurs the hinge loss $\ell_t(\theta_t) = \max\{0, 1 - y_t\theta_t^\top x_t\}$. Finally, the learner observes the instance $\{x_t, y_t\}$, and hence knows the loss for all $\theta \in \Theta$. The regret is defined as follows

$$R_T\left(\pi, \{x_t, y_t\}_{t=1}^T\right) = \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta)$$

Example 4 (The Expert Problem)**.** Given K arms, and denote $\Delta^K = \{p \in \mathbb{R}_+^K : p^\top \mathbf{1} = 1\}$. On each round t, the learner chooses some $p_t \in \Delta^K$. Simultaneously, the environment picks a loss vector $\ell_t \in [0, 1]^K$. Then, the learner incurs the loss $p_t^\top \ell_t$. Finally, the learner observes the loss vector ℓ_t, and hence knows the loss for all $p \in \Delta^K$. The regret is defined as follows

$$R_T\left(\pi, \underline{\ell}\right) = \sum_{t=1}^T p_t^\top \ell_t - \min_{a\in[K]} \sum_{t=1}^T \ell_t(a) = \sum_{t=1}^T p_t^\top \ell_t - \min_{p\in\Delta^K} \sum_{t=1}^T p^\top \ell_t$$

where $\min_{p\in\Delta^K} \sum_{t=1}^T p^\top \ell_t = \min_{a\in[K]} \sum_{t=1}^T \ell_t(a)$ is easy to see if we take derivative w.r.t. each coordinates of p in $\sum_{t=1}^T p^\top \ell_t$.

We will now present a unified framework for *online convex optimization.*

Definition 5 (Online convex optimization)**.** *Consider the following frame. A learner is given a weight space $\Omega \subset \mathbb{R}^d$. On each round t , the learner chooses a weight vector $w_t \in \Omega$. Simultaneously, the environment chooses a loss function $f_t : w \to \mathbb{R}$, a mapping from weight space to real line. Then the linear incurs the loss $f_t(w_t)$. Finally, the learner observes the loss function f_t, and hence knows the value of $f_t(w)$ for all $w \in \Omega$.*

In the above framework, if (1) the weight space Ω is convex and compact, and (2) the loss function f_t at every round is convex, the framework is called online convex optimization.

Given a horizon T. The goal is to minimize the regret against the best-fixed weight vector in Ω w.r.t. the policy π of choosing the weight vector at each round.

$$R_T\left(\pi, \underline{f}\right) = \sum_{t=1}^T f_t(w_t) - \min_{w\in\Omega} \sum_{t=1}^T f_t(w)$$

In example 3, the ℓ_2-ball is convex and compact, and the hinge loss is convex. In 4, Δ^K is convex and compact, and the loss $p_t^\top \ell_t$ is a linear function of p_t and thus convex.

3 Follow the Regularized Leader

A most straightforward policy is **Follow the Leader**(FTL). The weight w_t is chosen by

$$w_t \in \arg\min_{w\in\Omega} \sum_{s=1}^{t-1} f_s(w)$$

which is the best weight vector based on the observed loss function. However, this is often a bad idea, as the chosen weight could fluctuate from round to round. Therefore, we will stabilize the FTL by adding a regularized term $\Lambda(w)$

$$w_t \in \arg\min_{w\in\Omega} \left\{ \sum_{s=1}^{t-1} f_s(w) + \Lambda(w) \right\}$$

We call the above policy with the regularized term **Follow the Regularized Leader**(FTRL), and we will give its regret upper bound.

Theorem 5 (Regret Upper Bound for FTRL)**.** *For any $u \in \Omega$, FTRL satisfies*

$$\begin{aligned} R_T(\text{FTRL}, \underline{f}) &\leq \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(u) \\ &\leq \sum_{t=1}^{T} (f_t(w_t) - f_t(w_{t+1})) + \Lambda(u) - \min_{w\in\Omega} \Lambda(w) \end{aligned}$$

N.B. We have not assumed convexity of Ω, f_t, or Λ in the theorem.

Proof The first inequality is by the definition of regret. For the proof of the second inequality, we denote

$$F_t(w) = \sum_{s=1}^{t} f_s(w) + \Lambda(w)$$

and let

$$\Phi_t = \min_{w\in\Omega} F_t(w) = F_t(w_{t+1})$$

Consider $\Phi_{t-1} - \Phi_t$, and we have

$$\begin{aligned} \Phi_{t-1} - \Phi_t &= F_{t-1}(w_t) - F_t(w_{t+1}) \\ &= F_{t-1}(w_t) - (F_{t-1}(w_{t+1}) + f_t(w_{t+1})) \\ &= (F_{t-1}(w_t) - F_{t-1}(w_{t+1})) - f_t(w_{t+1})) \\ &\leq -f_t(w_{t+1}) \end{aligned}$$

since $F_{t-1}(w_t) \leq F_{t-1}(w_{t+1})$, Then we will have

$$\Phi_{t-1} - \Phi_t + f_t(w_t) \leq f_t(w_t) - f_t(w_{t+1})$$

by adding $f_t(w_t)$ to both sides of the equation. Then we sum both sides from $t = 1, .., T$, and we will have

$$\Phi_0 - \Phi_T + \sum_{t=1}^{T} f_t(w_t) \leq \sum_{t=1}^{T} (f_t(w_t) - f_t(w_{t+1}))$$

We can compute the values of Φ_T, Φ_0 as follows:

$$\Phi_T = \min_{w\in\Omega}\left(\sum_{s=1}^{T} f_s(w) + \Lambda(w)\right) \leq \sum_{s=1}^{T} f_s(u) + \Lambda(u)$$

$$\Phi_0 = \min_{w\in\Omega} \Lambda(w)$$

Therefore, we have

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{s=1}^{T} f_s(u) - \Lambda(u) + \min_{w\in\Omega}\Lambda(w) \leq \sum_{t=1}^{T}(f_t(w_t) - f_t(w_{t+1}))$$

and thus

$$\begin{aligned} R_T(\text{FTRL}, \underline{f}) &\leq \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(u) \\ &\leq \sum_{t=1}^{T}(f_t(w_t) - f_t(w_{t+1})) + \Lambda(u) - \min_{w\in\Omega}\Lambda(w) \end{aligned}$$

□

Remark:

- The above theorem implies that for follow the leader (FTL),

$$R_T(\text{FTRL}, \underline{f}) \leq \sum_{t=1}^{T}(f_t(w_t) - f_t(w_{t+1}))\,.$$

- If w_t fluctuates frequently, the regret of FTRL/FTL will be bad.
- The purpose of the regularized term $\Lambda(w)$ is to stabilize the chosen weight w_t.

4 Examples Analysis: How a regularizer is Chosen

To motivate how a regularizer is chosen, we will consider 3 examples for FTL with $\Omega = [0, 1]$ and $f_t : [0, 1] \to [0, 1]$

4.1 Example 1: FTL with linear losses

First, Let $\Omega = [0, 1]$. Then we define $f_t(w)$ $\forall w \in \Omega$:

$$f_t(w) = \begin{cases} \frac{1}{2}w & \text{if } t = 1 \\ w & \text{if } t \text{ is odd, } t > 1 \\ 1 - w & \text{if } t \text{ is even} \end{cases}$$

We have:

$$F_t(w) = \sum_{s=1}^{t} f_s(w) = \begin{cases} \frac{1}{2}w + \frac{t-1}{2} & \text{if } t \text{ is odd} \\ -\frac{1}{2}w + \frac{t}{2} & \text{if } t \text{ is even} \end{cases}$$

Hence, we have the following:

$$w_t = \arg\min_{w\in[0,1]} F_{t-1}(w) = \begin{cases} 0 & \text{if } t \text{ is even} \\ 1 & \text{if } t \text{ is odd} \end{cases}$$

Therefore, we obtain the Upper Bound from the Thm 3:

$$R_T \leq \sum_{t=1}^{T} f_t(w_t) - f_t(w_{t-1}) = \sum_{t \text{ s.t } t \text{ is odd}} (1-0) + \sum_{t \text{ s.t } t \text{ is even}} (1-0) \simeq T.$$

The bound given by the theorem is $O(T)$. Moreover, it is not hard to see that the actual regret is also large. The total loss of FTL is at least $T-1$. The best action in hindsight will have loss at most $\frac{T}{2}$. Therefore, we have **Regret** $\geq \frac{T}{2} - 1$, and we could see that the Bound on R_T is pretty tight. The linear losses are bad use case for FTL.

4.2 Example 2: FTL with quadratic losses

Let $\Omega = [0,1]$, and we define $f_t(w), \forall w \in \Omega$ as following:

$$f_t(w) = \begin{cases} w^2 & \text{if } w \text{ is odd} \\ (1-w)^2 & \text{if } w \text{ is even} \end{cases}$$

Similar to the previous example, the best action for a given round oscillates between 0 and 1. However, we will see that the regret is not large.

First note that the sum of losses can be written as:

$$F_t(w) = \begin{cases} \frac{t+1}{2} w^2 + \frac{t-1}{2}(1-w)^2 & \text{if } t \text{ is odd} \\ \frac{t}{2}\left(w^2 + (1-w)^2\right) & \text{if } t \text{ is even} \end{cases}$$

Hence we have,

$$w_t = \arg\min_{w\in[0,1]} F_{t-1}(w) = \begin{cases} \frac{1}{2} & \text{if } t \text{ is odd} \\ \frac{1}{2} - \frac{1}{2t} & \text{if } t \text{ is evens} \end{cases}$$

We see that the choices made by FTL do not oscillate much, with $w_t \to \frac{1}{2}$ as $t \to \infty$. We have the following upper bound:

$$\begin{aligned} R_T &\leq \sum_{t=1}^{T} f_t(w_t) - f_t(w_{t+1}) \\ &= \sum_{t \text{ s.t } t \text{ is odd}} \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2} - \frac{1}{2(t+1)}\right)^2 + \sum_{t \text{ s.t } t \text{ is even}} \left(\frac{1}{2} + \frac{1}{2t}\right)^2 - \left(\frac{1}{2}\right)^2 \\ &= \sum_{t=1}^{T} \mathcal{O}\left(\frac{1}{t}\right) + \mathcal{O}\left(\frac{1}{t^2}\right) \\ &\in \mathcal{O}(\log T) \end{aligned}$$

Acknowledgement

These notes are based on scribed lecture materials prepared in Fall 2023 by Xindi Lin, Tony Chang Wang, and Haoyue Bai,Deep Patel.