CS861: Theoretical Foundations of Machine Learning Lecture 26 - 03/11/2023
University of Wisconsin–Madison, Fall 2023

Lecture 26: Online Convex Optimization (continued)

Lecturer: Kirthevasan Kandasamy *Scribed by: Haoyue Bai & Deep Patel*

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.*

We will return to `contextual bandits` next week. Right now, we will continue studying `online convex optimization`: A learner is given a convex set, $\Omega \subseteq \mathbb{R}^d$, and at each rount-$t$, the learner chooses an action $w_T \in \Omega$. Simultaneously, the environment picks a convex loss function $f_t : \Omega \to \mathbb{R}$ so that the player incurs the loss $f_t(w_t)$. We also assume that the player gets to observe all f_t's. So far, we have looked at the framework of Follow-The-Regularized-Leader (FTRL) framework that helps the learner choose an action $w_t \in \Omega$ such that its regret w.r.t. the best action in hindsight is minimized by introducing an appropriate additive regularizer term to the losses incurred.

We concluded the last lecture by looking at three examples in order to motivate how to pick these regularizers. In particular, we saw that Follow-The-Leader (FTL) has bad regret ($R_T \in \mathcal{O}(T)$) when the learner uses linear losses whereas FTL does well when using quadratic losses ($R_T \in \mathcal{O}(\log T)$). Next, we saw that if we were to use linear losses with quadratic regularizers, FTRL actually has better regret compared to FTL-with-linear-losses (i.e., $R_T \in \mathcal{O}(\sqrt{T})$). However, it is not clear what type of regularizer to use for a given loss function. We will now formalize this process of choosing an appropriate regularizer given a convex loss function so as to ensure good regret bounds under the FTRL framework. Specifically, in this lecture, we will first present a primer on some properties of convex functions, and then study the framework of Follow-The-Regularized-Leader (FTRL) with convex losses and strongly convex regularizers.

1 A primer on properties of convex functions

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 1.1. *Some facts about sub-gradients:*

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

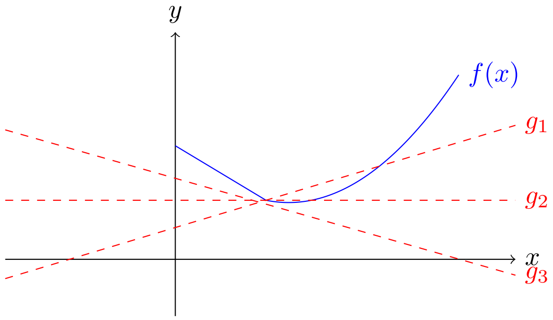

Figure 1: The blue curve above depicts a convex function, f, whereas the red lines denote the first-order linear underestimators to this function f. Additionally, these linear underestimators also serve as three of the uncountably-infinite possible subgradients at the non-different point $((0.8, 0.52))$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

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

(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 $||\cdot||_2$.

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

Remark 1.2. *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 $w^* = \underset{w\in\Omega}{\arg\min} f(w)$*, 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.) $(||\cdot||_1, ||\cdot||_\infty)$

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

(iii.) More generally, given that Hölder's inequality (Lemma 1 below) holds, 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 FTRL with convex losses and strongly-convex regularizers

We will not state our main theorem for FTRL with convex losses and strongly-convex regularizers.

Theorem 2.1. *Suppose f_t is convex for all t and $\Lambda(w) = \frac{1}{\eta}\lambda(w)$ where $\eta > 0$ and λ is 1-strongly convex with respect to some norm $||\cdot||$. Let $||\cdot||*$ be the dual-norm of $||\cdot||$, and let $g_t \in \partial f(w_t)$, where w_t was chosen by FTRL. Then,*

$$\begin{aligned} R_T(FTRL, \underline{f}) &\triangleq \sum_{t=1}^{T} f_t(w_t) - \min_{w\in\Omega} \sum_{t=1}^{T} f_t(w) \\ &\leq \frac{1}{\eta}\left(\max_{w\in\Omega} \lambda(w) - \min_{w\in\Omega} \lambda(w)\right) + \eta \sum_{t=1}^{T} ||g_t||_*^2 \end{aligned}$$

Corollary 1. *Suppose $\max\limits_{w\in\Omega} \lambda(w) - \min\limits_{w\in\Omega} \lambda(w) \leq B$ and $||g_t||_* \leq G\ \forall t$. Then, choosing $\eta = \sqrt{\frac{B}{TG^2}}$, we have*

$$R_T \leq \frac{B}{\eta} + \eta T G^2 \in \mathcal{O}(G\sqrt{BT}).$$

Remark 2.1. *Note that the condition $||g_t||_* \leq G\ \forall t$ here means that f_t is G-Lipschitz in $||\cdot||_*$-norm.*

We will do the proof of Theorem 1 after looking at some examples below.

Example 2 (Linear Losses)**.** Let $\Omega = \{w \mid ||w||_2 \leq 1\}$ and $f_t(w) = w^T \ell_2$ where $||\ell_t||_2 \leq 1$. We will apply FTRL result with $\lambda(w) = \frac{1}{2}||w||_2^2$ which is 1-strongly convex in $||\cdot||_2$. We will compute the best action on round-t as follows:

$$\begin{aligned} w_t &= \arg\min_{w\in\Omega} \sum_{s=1}^{t-1} f_s(w) + \Lambda(w) \\ &= \arg\min_{w\in\Omega} w^T \left(\sum_{s=1}^{t-1} \ell_s(w)\right) + \frac{1}{2\eta}||w||_2^2 \\ &= \arg\min_{w\in\Omega} ||w||_2^2 + 2\eta w^T \left(\sum_{s=1}^{t-1} \ell_s\right) + \eta^2 \left(\sum_{s=1}^{t-1} \ell_s(w)\right)^2 \\ &= \arg\min_{w\in\Omega} ||w + \eta \sum_{s=1}^{t-1} \ell_s(w)||_2 \end{aligned}$$

We should choose $w_t = \text{proj}_\Omega \left(-\eta \sum_{s=1}^{t-1} \ell_s(w)\right)$. This can be implemented via the following update in $\mathcal{O}(1)$ time on each round t:

$$\begin{aligned} u_0 &\triangleq 0 \\ u_t &\longleftarrow u_{t-1} - \eta \ell_{t-1} \\ w_t &\longleftarrow \arg\min_{w\in\Omega} ||w - u_t||_2 \end{aligned}$$

This gives us the following regret bound:

$$\begin{aligned}
R_T(FTRL, \underline{\ell}) &\leq \frac{1}{\eta}\left(\max_{w\in\Omega} \frac{1}{2}||w||_2^2 - \min_{w\in\Omega} \frac{1}{2}||w||_2^2\right) + \eta\sum_{t=1}^{T} ||\ell_t||_2^2 \\
&= \frac{1}{2\eta}(1-0) + \eta\sum_{t=1}^{T} ||\ell_t||_2^2 \ \ (\because 0 \leq ||w||_2 \leq 1 \ \forall \ w \in \Omega) \\
&\leq \frac{1}{2\eta} + \eta \cdot T \ \ (\because ||\ell_t||_2 \leq 1 \ \forall \ t) \\
&\in \mathcal{O}(\sqrt{T}) \quad \left(\text{if } \eta = \frac{1}{\sqrt{T}}\right)
\end{aligned}$$

Example 3 (Experts Problem)**.** $\Omega = \Delta^K = \{p \in \mathbb{R}_+^K, 1^T p = 1\}$, $f_t(p) = \ell^T p, \ell_t \in [0,1]^K$. Say $K \geq 2$.

Let's try FTRL with $\lambda(w) = \frac{1}{2}||w||_2^2$. Doing the same calculations as in the Example above, we get the following regret bound:

$$\begin{aligned}
&R_T(FTRL, \underline{\ell}) \leq \frac{B}{\eta} + \eta\sum_{t=1}^{T} ||\ell_t||_2^2 \\
&\quad \text{Note that } B = \max_{w\in\Omega} \lambda(w) - \min_{w\in\Omega} \lambda(w) = \frac{1}{2}\left(1 - \frac{1}{K}\right) \leq \frac{1}{2} \ (\because K \geq 2) \\
&\quad\quad \text{and that } ||\ell_t||_2^2 \leq K \ (\because \ell_t \in [0,1]^K) \\
&\therefore R_T(FTRL, \underline{\ell}) \leq \frac{1}{2\eta} + \eta K T \\
&\therefore R_T(FTRL, \underline{\ell}) \in \mathcal{O}(\sqrt{KT}) \quad \left(\text{for } \eta = \sqrt{\frac{1}{KT}}\right)
\end{aligned}$$

Remark 2.2. *We observe the following when comparing the regret bounds derived in Example 2 and Example 3 above with that of* `Hedge` *as derived in previous lectures:*

- *Recall that we have the following regret bound for* `Hedge` *algorithm:*

$$R_T \in \mathcal{O}(\sqrt{T \log K})$$

- *So, it seems that we are not accurately capturing the geometry of the problem here. That is, ℓ_2-norm hypercube scales with K, whereas, say, ℓ_∞-norm for $[0,1]^K$ would remain a constant. So, we would want to use a regularizer that is strongly convex in some norm other than ℓ_2-norm; say ℓ_1-norm.*
- *You will be proving in the upcoming homework that using $\lambda(p) = -H(p)$, which is strongly convex in the ℓ_1-norm, yields a better regret bound here.*

$$B = \max_{w\in\Omega} \lambda(w) - \min_{w\in\Omega} \lambda(w) \leq \log K$$

$$R_T \leq \frac{\log K}{\eta} + \eta \cdot T \cdot 1 \in \mathcal{O}(\sqrt{T \log K})$$

Remark 2.3. *Thus, if we know/anticipate that $\{\nabla f_t\}_{t\geq 1}$ are small in some dual-norm $||\cdot||_*$, then it would be a good idea to run FTRL with a regularizer Λ which is strongly convex w.r.t. the corresponding norm*[2] $(||\cdot||_*)_* = ||\cdot||$.

[2] dual of the dual-norm, which is the norm itself