CS861: Theoretical Foundations of Machine Learning Lecture 24 - 10/30/2024
University of Wisconsin–Madison, Fall 2024

Lecture 24: Online Convex Optimization (Contd.)

Lecturer: Kirthevasan Kandasamy Scribed by: Yupeng Zhang, Sathya Kamesh Bhethanabhotla

In this lecture we will continue where we left off in the previous lecture, by looking at the examples of FTL (follow the leader algorithm).

Example 2: FTL with quadratic losses (cont'd)

$$\text{Example 1:} \quad \Omega = [0,1], \quad f_t(\omega) = \begin{cases} \frac{1}{2}\omega & \text{if } t = 1, \\ \omega & \text{if } t \text{ is odd}, t > 1, \\ 1-\omega & \text{if } t \text{ is even.} \end{cases}$$

$$\text{Example 2:} \quad \Omega = [0,1], \quad f_t(\omega) = \begin{cases} \omega^2 & \text{if } t \text{ is odd}, \\ (1-\omega)^2 & \text{if } t \text{ is even.} \end{cases}$$

Like in Example 1, *the best action for a given round* i.e. $\text{argmin}_\omega f_t(\omega)$ fluctuates from 0 to 1. However, the regret is not large since $\text{argmin}_\omega F_t(\omega)$ does not fluctuate.

Question: Let us consider linear losses again, but with FTRL. What type of regularizer should we use?

Example 3: FTRL with Linear Losses

We will revisit the linear losses in the first example (see the previous lecture note),

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

and

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

but this time we add a regularizer to stabilize the fluctuations. Since quadratic losses achieved small regret, let us try regularizer $\Lambda(w) = \frac{1}{\eta}(w - \frac{1}{2})^2$ (η will be chosen later). We define f_t same as in example 1, namely: $\forall w \in \Omega = [0,1]$:

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

Then we have $F_t(w)$ with parameter η to be specified later:

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

Hence we got:

$$w_t = \arg\min_{w\in[0,1]} F_{t-1}(w) = \begin{cases} \frac{1}{2}+\frac{\eta}{4} & \text{if } t \text{ is odd} \\ \frac{1}{2}-\frac{\eta}{4} & \text{if } t \text{ is even} \end{cases}$$

Then we have the following upper bound on the regret. Define

$$B := \max_{w\in[0,1]} \frac{1}{\eta}(w-\frac{1}{2})^2 - \min_{w\in[0,1]} \frac{1}{\eta}(w-\frac{1}{2})^2 = \frac{1}{4\eta}$$

We have:

$$\begin{aligned} R_T &\leq \sum_{t=1}^{T} f_t(w_t) - f_t(w_{t+1}) + B \\ &= \sum_{t \text{ odd}} \left(\frac{1}{2}+\frac{\eta}{4}\right) - \left(\frac{1}{2}-\frac{\eta}{4}\right) + \sum_{t \text{ even}} \left(\frac{1}{2}+\frac{\eta}{4}\right) - \left(\frac{1}{2}-\frac{\eta}{4}\right) + B \\ &= \sum_{t=1}^{T} \frac{\eta}{2} + \frac{1}{4\eta} = \frac{\eta T}{2} + \frac{1}{4\eta} \end{aligned}$$

Next, we choose optimal $\eta = \frac{1}{\sqrt{T}}$. Based on the regret's upper bound we just showed, we have:

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

Take-aways from the Examples

Some key insights from the examples above:

- Linear functions have bad behaviour in FTL due to the instability of the chosen w_t
- Strong convexity in the loss function helps stabilize the algorithm
- We should add a "nice" regularizer to stabilize oscillations ("nice" means strong convexity here)
- The choice of regularization parameter η is crucial for achieving optimal regret bounds
- With proper regularization, we can achieve $\mathcal{O}(\sqrt{T})$ regret even with linear losses

1 FTRL with convex losses and strongly-convex regularizers

We can now state our main theorem for FTRL with convex losses and strongly-convex regularizers.

Theorem 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,*

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

Proof: Recall the following bound for FTRL. For all $u \in \Omega$,

$$\sum_{t=1}^{T} f_t(\omega_t) - \sum_{t=1}^{T} f_t(u) \leq \lambda(u) - \min_{\omega \in \Omega} \lambda(\omega) + \sum_{t=1}^{T} (f_t(\omega_t) - f_t(\omega_{t+1})).$$

We will apply this theorem with $u \leftarrow \omega_* \in \operatorname{argmin}_{\omega \in \Omega} \sum_{t=1}^{T} f_t(\omega)$. We have,

$$\begin{aligned} R_T(\pi, f) &\triangleq \sum_{t=1}^{T} f_t(\omega_t) - \sum_{t=1}^{T} f_t(\omega_*) \\ &\leq \frac{1}{\eta} \left(\lambda(\omega_*) - \min_{\omega \in \Omega} \lambda(\omega) \right) + \sum_{t=1}^{T} (f_t(\omega_t) - f_t(\omega_{t+1})) \end{aligned}$$

It is sufficient to show $(f_t(\omega_t) - f_t(\omega_{t+1})) \leq \eta \|g_t\|_*^2$. By convexity, as $g_t \in \partial f_t(\omega_t)$, we have

$$f_t(\omega_{t+1}) \geq f_t(\omega_t) + g_t^\top (\omega_{t+1} - \omega_t)$$

Hence, by Hölder's inequality, we have

$$f_t(\omega_t) - f_t(\omega_{t+1}) \leq g_t^\top (\omega_t - \omega_{t+1}) \leq \|\omega_{t+1} - \omega_t\| \|g_t\|_*$$

Now, denote $F_t(\omega) = \sum_{s=1}^{t} f_s(\omega) + \frac{1}{\eta} \lambda(\omega)$. We have that F_t is $\frac{1}{\eta}$-strongly convex, as λ is 1-strongly convex and f_t's are convex. Note:
(i) F_t is $\frac{1}{\eta}$-strongly convex.
(ii) If $\omega_* = \operatorname{argmin}_{\omega \in \Omega} f(\omega)$, where f is α-strongly convex, then $f(\omega) \geq f(\omega_*) + \frac{\alpha}{2} \|\omega - \omega_*\|_2^2$.
(iii) $f_t(\omega_t) - f_t(\omega_{t+1}) \leq \|\omega_{t+1} - \omega_t\| \|g_t\|_*$

Recall, in FTRL, we have $\omega_t = \operatorname{argmin}_\omega F_{t-1}(\omega)$. Therefore, ω_{t+1} minimizes F_t and ω_t minimizes F_{t-1}. Using (i), (ii) we have,

$$\begin{aligned} F_{t-1}(\omega_{t+1}) - F_{t-1}(\omega_t) &\geq \frac{1}{2\eta} \|\omega_t - \omega_{t+1}\|^2, \\ F_t(\omega_t) - F_t(\omega_{t+1}) &\geq \frac{1}{2\eta} \|\omega_t - \omega_{t+1}\|^2. \end{aligned}$$

Summing both sides we have, $f_t(\omega_t) - f_t(\omega_{t+1}) \geq \frac{1}{\eta} \|\omega_t - \omega_{t+1}\|^2$.
(i) $f_t(\omega_t) - f_t(\omega_{t+1}) \leq \|\omega_{t+1} - \omega_t\| \|g_t\|_*$.
(ii) $f_t(\omega_t) - f_t(\omega_{t+1}) \geq \frac{1}{\eta} \|\omega_t - \omega_{t+1}\|^2$.

Therefore,

$$\begin{aligned} (i), (ii) &\Rightarrow \|\omega_t - \omega_{t+1}\|^2 \leq \eta (f_t(\omega_t) - f_t(\omega_{t+1})) \leq \eta \|\omega_{t+1} - \omega_t\| \|g_t\|_* \qquad \text{(iii)} \\ &\Rightarrow \|\omega_t - \omega_{t+1}\| \leq \eta \|g_t\|_* \\ (i), (iii) &\Rightarrow f_t(\omega_t) - f_t(\omega_{t+1}) \leq \eta \|g_t\|_*^2 \end{aligned}$$

□

Now that we have proved the theorem, we can state a corollary for Theorem 1 that is a more useful form of the result.

Corollary 1. *Suppose* $\max_{w \in \Omega} \lambda(w) - \min_{w \in \Omega} \lambda(w) \leq B$ *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 The corollary gives a good intuition about the rate when the regularizer is bounded by some quantity. Note that the condition $||g_t||_* \leq G \; \forall t$ here means that f_t is G-Lipschitz in $|| \cdot ||_*$-norm.

Let us now look at examples with some strongly convex regularizers.

Example 2 (Linear Losses)**.** Let $\Omega = \{w \mid ||w||_2 \leq 1\}$ and $f_t(w) = w^T \ell_2$ where $||\ell_t||_2 \leq 1$ (element-wise). 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:

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

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 iterative scheme in $\mathcal{O}(1)$ time:

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

So the Regret satisfies the following bound:

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

Example 3 (Online Gradient Descent)**.** Let f_t be differentiable[1] $\forall t$ and Ω be a compact, convex set. We can apply `FTRL` with any regularizer that is 1-strongly convex in some norm $|| \cdot ||$ with the following rule:

$$w_t \in \arg\min_{w \in \Omega} \sum_{s=1}^{t-1} f_s(w) + \frac{1}{2\eta}||w||_2^2$$

Here we can notice that though this scheme gives us good regret rates, computing a new gradient $\nabla f_t(w)$ in each iteration gives us a complexity that grows linearly in t. However we would ideally like a constant

[1] We don't actually need this assumption. We are using it for simplicity in this class.

cost method. So, we will take a different perspective to circumvent this issue. We will start by rewriting the regret as follows:

$$
\begin{aligned}
R_T(\pi, \{f_t\}_{t=1}^T) &= \sum_{t=1}^T f_t(w_t) - \min_{w\in\Omega} \sum_{t=1}^T f_t(w) \\
&= \max_{w\in\Omega} \left(\sum_{t=1}^T [f_t(w_t) - f_t(w)] \right) \\
&\leq \max_{w\in\Omega} \left(\sum_{t=1}^T \nabla f_t^T(w_t)(w_t - w) \right) \\
&\big(\because f_t \text{ is convex } \iff f_t(w) \geq f_t(w_t) + (w - w_t)^T \nabla f_t(w_t)\ \forall w \in \Omega\big) \\
&= \sum_{t=1}^T w_t^T \nabla f_t(w_t) - \min_{w\in\Omega} \sum_{t=1}^T w^T \nabla f_t(w_t) \\
&= R_T\left(\pi, \underbrace{\{\nabla f_t(w_t)\}_{t=1}^T}_{\text{abuse of notation}^2}\right)
\end{aligned}
$$

We can see that these are Linear Losses with $\ell_t = \nabla f_t(\omega_t)$ We can now apply FTRL on the linear losses $\tilde{f}_t(w) \triangleq w^T \nabla f_t(w_t)$ with $\lambda(w) = \frac{1}{2}||w||_2^2$ as shown below:

$$
\begin{aligned}
w_t &= \arg\min_{w\in\Omega} \left(w^T \left(\sum_{s=1}^{t-1} \nabla f_s(w_s) \right) + \frac{1}{2\eta} ||w||_2^2 \right) \\
&= \arg\min_{w\in\Omega} ||w + \eta \sum_{s=1}^{t-1} \nabla f_s(w_s)||_2 \ \text{ (by completing the squares)}
\end{aligned}
$$

Hence, w_t will be the ℓ_2-projection of $-\eta \sum_{s=1}^{t-1} \nabla f_s(w_s)$ to Ω, which can be implemented in $\mathcal{O}(1)$-time[3] at each round-t as follows:

$$
\begin{aligned}
u_t &\longleftarrow u_{t-1} - \eta \nabla f_{t-1}(w_{t-1}) \\
w_t &\longleftarrow \arg\min_{w\in\Omega} ||w - u_t||_2
\end{aligned} \tag{1}
$$

Now, we can show that

$$
\begin{aligned}
R_T(\pi, \{f_t\}_{t=1}^T) &\leq R_T(\pi, \{\nabla f_t(w_t)\}_{t=1}^T) \\
&\leq \frac{B}{\eta} + \eta T G^2 \ \text{ (By Theorem 1)} \\
&\in \mathcal{O}(G\sqrt{BT}) \ \left(\text{if } \eta = \sqrt{\frac{B}{TG^2}}\right)
\end{aligned} \tag{2}
$$

where $B = \max_{w\in\Omega} \lambda(w) - \min_{w\in\Omega} \lambda(w)$ and $||\nabla f_t(w_t)||_2 \leq G\ \forall t$.

Remark Some connections that we can make a note of:

- If we fix the function $f_t = f$ and we want to find its minimum $\omega_\star = \arg\min_{\omega\in\Omega} f(\omega)$, this is similar to

[2] We mean $w_t^T \nabla f_t(w_t)$ here

[3] We are not considering how this scales with the dimensionality, d, of $\Omega \subseteq \mathbb{R}^d$ at the moment

the standard Projected Gradient Descent (PGD) step:

$$\begin{aligned} u_t &\longleftarrow w_{t-1} - \eta \nabla f(w_{t-1}) \\ w_t &\longleftarrow \arg\min_{w \in \Omega} ||w - u_t||_2 \end{aligned}$$

We can also obtain the following guarantee for the PGD step:

$$\begin{aligned} \min_{w_t} f(w_t) - f(w_*) &\leq \frac{1}{T}\left(\sum_{t=1}^{T} f(w_t) - f(w_*)\right) \quad (\because \min \leq \text{ avg.}) \\ &\in \mathcal{O}\left(G\sqrt{\frac{B}{T}}\right) \end{aligned}$$

Note that this need not necessarily be an optimal bound. We are simply showing *an* application of Theorem 1 to a convex optimization problem.

- In machine learning, update rule defined in Equation 1 is similar to the (projected) Stochastic Gradient Descent (SGD) update where f_t is the loss for instance (x_t, y_t)

Example 4 (Experts Problem - Revisited)**.** Here we have $\Omega = \Delta^K = \{p \in \mathbb{R}_+^K, 1^T p = 1\}$, and $f_t(p) = \ell^T p, \ell_t \in [0,1]^K$. Lets consider $K \geq 2$.

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:

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

When comparing the regret bounds derived in the Example above with that of `Hedge` as derived in previous lectures, we can see that `Hedge` has a tighter bound:

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

The issue is 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; for instance, the ℓ_1-norm.

We can try the negative entropy as a regularizer,

$$\lambda(p) = -H(p) = \sum_{i=1}^{K} p(i) \log(p(i))$$

Recall that $\lambda(p)$ is 1-strongly convex in $|| \cdot ||_1$ We will continue this approach in the next lecture.

Acknowledgements

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