Lecture 16: FTRL and Online Mirror Descent

Akshay Krishnamurthy
akshay@cs.umass.edu

November 2, 2017

1 Recap

Last time we saw two online learning algorithms. First we saw the Weighted Majority algorithm, which is also called Hedge, Exponential Weights, or as we'll see today, Exponential Gradient. We proved that in the experts setting this algorithm achieves $O(\sqrt{T \log(d)})$ regret when there are T rounds and d experts.

We also started discussion of the FTRL algorithm for the online convex optimization setting, which of course generalizes the experts setting. Here the algorithm is

$$w_t \leftarrow \operatorname*{argmin}_{w \in S} R(w) + \sum_{i=1}^{t-1} f_i(w)$$

where the convex loss functions are f_t, R is some regularizer and w_ts are the actions that the learner plays. Today we will study this algorithm in detail.

2 FTRL

Before turning to the general case let us just study the quadratic regularizer $R(w) = \frac{1}{2\eta}\|w\|_2^2$ with linear loss functions:

$$w_t \in \operatorname*{argmin}_{w \in S} \frac{1}{2\eta}\|w\|_2^2 + \sum_{i=1}^{t-1}\langle w, \ell_i\rangle = \operatorname*{argmin}_{w \in S} \frac{1}{2\eta}\|w\|_2^2 - \langle w, \theta_t\rangle$$

where $\theta_t = -\sum_{i=1}^{t-1} \ell_i$ is the sum of the loss functions so far. When $S = \mathbb{R}^d$ FTRL has a closed form

$$w_t = \eta\theta_t$$

which we can equivalently write as $w_t = w_{t-1} - \eta\ell_{t-1}$ which is the *Online Gradient Descent* algorithm.

If $S \neq \mathbb{R}^d$ then we are doing a lazy projection step where

$$w_t = \Pi_S(\eta\theta_t),$$

This is called a lazy projection since we don't project the iterates θ_t but only project when we need to make a prediction. This is also called *Nesterov's Dual Averaging* algorithm.

Theorem 1. *FTRL with quadratic regularizer $\frac{1}{2\eta}\|w\|_2^2$ and linear losses $f_t(w) = \langle w, \ell_t\rangle$ satisfies*

$$\textit{Regret}(T, u) \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\|\ell_t\|_2^2$$

If $\|u\|_2 \leq B$ and $\|\ell_t\| \leq L$ then setting $\eta = \frac{B}{L\sqrt{T}}$ gives $O(BL\sqrt{T})$ regret.

Proof. We first apply the Be-the-Leader lemma, where we imagine that the first loss is just R, this gives

$$R(w_0) - R(u) + \sum_{t=1}^{T} \langle w_t, \ell_t \rangle - \langle u, \ell_t \rangle \le R(w_0) - R(w_1) + \sum_{t=1}^{T} \langle w_t - w_{t+1}, \ell_t \rangle$$

Re-arranging and using non-negativity of R we obtain the first term

$$\text{Regret}(T, u) \le R(u) + \sum_{t=1}^{T} \langle w_t - w_{t+1}, \ell_t \rangle$$

Now for a single term

$$\langle w_t - w_{t+1}, \ell_t \rangle \le \|w_t - w_{t+1}\|_2 \|\ell_t\|_2 = \|\Pi_S(\eta\theta_t) - \Pi_S(\eta\theta_{t+1})\| \|\ell_t\| \le \|\eta\theta_t - \eta\theta_{t+1}\| \|\ell_t\| = \eta\|\ell_t\|_2^2$$

□

The above bound works for any linear losses, but what about convex losses? The trick here is to linearize the convex loss function and the main question is what linear function we should use? Given f_t, we need to find a linear function that we then pass to our linear FTRL algorithm. The trick is to use $\ell_t \in \partial f_t(w_t)$

Proposition 2 (Linearization)**.** *For a convex loss function f_t, with $\ell_t \in \partial f_t(w_t)$ we have*

$$f_t(w_t) - f_t(u) \le \langle w_t - u, \ell_t \rangle$$

Proof. The proof is just by the definition of convexity, we know that $f_t(u) \ge f_t(w_t) + \langle \ell_t, u - w_t \rangle$ which is just a re-arranging of the claim. □

The intuition here is that, at least statistically, linear losses are the hardest to optimize, since the have no curvature. Looking at the quadratic example from last time, we saw that curvature seemed to help us, so this is not super surprising.

Corollary 3 (FTRL Regret for experts setting)**.** *FTRL with quadratic regularizer for the experts setting has regret*

$$\textit{Regret}(T) \le \sqrt{dT}$$

Proof. Since $S = \Delta(d)$ we know that $\|u\|_2 = 1$ while $\|\ell_t\|_2 = \sqrt{d}$, since we actually have that $\ell_t \in [0,1]^d$. Optimizing for η proves the result. □

This is actually terrible since we saw that weighted majority achieves $\sqrt{T \log(d)}$ regret in the experts setting. How can we achieve this with FTRL?

3 Online Mirror Descent

The key is to use a different regularizer. To study the more general case we will need some more technical machinery. We will study the *online mirror descent* algorithm, which is equivalent to FTRL, but we'll see a different perspective as well.

$$w_t \in \operatorname*{argmin}_{w \in S} R(w) - \langle w, \theta_t \rangle, \qquad \theta_t = -\eta \sum_{i=1}^{t-1} \ell_i, \qquad \ell_i \in \partial f_i(w_i)$$

Here I also moved the learning rate to the loss term, but clearly this doesn't change anything.

1. OMD on losses f_t just FTRL on linearization $\langle w, z_t \rangle$.

2. OGD is just OMD with quadratic regularizer.

Fenchel Conjugates. To understand OMD we need to introduce three concepts. The first is fenchel conjugates. For a function ψ that need not be convex

$$\psi^\star(\theta) \triangleq \sup_{w \in \mathbb{R}^d} \langle w, \theta \rangle - \psi(w)$$

The idea here is that $\psi^\star$ describes the supporting hyperplanes to the function ψ. In the one-dimensional case, $-\psi^\star(\theta)$ is the x-intercept of the supporting hyperplane to ψ with slope θ. The main properties about Fenchel Conjugates are

1. $\psi^\star$ is always convex. It is a pointwise maximum of linear functions
2. $\psi^\star(\theta) + \psi(w) \geq \langle w, \theta \rangle$ which is called the Fenchel-Young inequality.
3. If $R(w) = \alpha\psi(w)$ then $R^\star(\theta) = \alpha\psi^\star(\theta/\alpha)$ for $\alpha > 0$.
4. If ψ is differentiable then $\nabla\psi^\star(\theta) = \operatorname{argmax}_w \langle w, \theta \rangle - \psi(w)$. This follows since the gradient of a maximum is the gradient of the function acheiving the maximum, which in this case is just the w.

Thus in the unconstrained case, we can write OMD in another way

$$w_t = \nabla R^\star(\theta_t), \qquad \theta_t = \theta_{t-1} - \eta\ell_{t-1},$$

and we also know that w_t and θ_t are linked since $\theta_t = \nabla R(w_t)$. This is where the word *mirror* comes from: the θ_t are updated in a gradient descent style and the updates are mirrored to the "primal" space using Fenchal duality.

Bregman Divergences. For any continuously differentiable convex function ψ, define

$$D_\psi(u||v) = \psi(u) - \psi(v) - \langle \nabla\psi(v), u - v \rangle$$

which is the difference between $\psi(u)$ and the first order approximation relative to v. Convexity ensures that $D_\psi \geq 0$ but note that it is not in general symmetric. The idea for a better analysis of FTRL/OMD is to use a different regularizer and instead of the ℓ_2 norm $\|w_t - w_{t+1}\|_2$ that we saw in the proof above, we will use the Bregman divergence.

Before turning to the real analysis, an characterization for the constrained case that is closer to the second one above is.

1. Choose $\tilde{w}_{t+1}$ so that $\nabla R(\tilde{w}_{t+1}) = \nabla R(\tilde{w}_t) - \eta\ell_t$ (Or inductively so that $\nabla R(\tilde{w}_{t+1}) = \theta_t$)
2. Choose $w_{t+1} = \operatorname{argmin}_{w \in S} D_R(w||\tilde{w}_{t+1})$

In the unconstrained case this is equivalent to the second one since the unconstrained argmin is just $\tilde{w}_{t+1}$. Actually it doesn't really matter whether you use $\tilde{w}_t$ in the first line or not. This version is called the Lazy version, and if you use w_t there it is called the Agile version (at least according to Hazan).

Lemma 4. *Lazy OMD and FTRL produce identical predictions when the loss functions are linear*

Proof. We need to prove that

$$\operatorname*{argmin}_{w \in S} D_R(w||\tilde{w}_t) = \operatorname*{argmin}_{w \in S} R(w) - \eta\langle w, \theta_t \rangle$$

First, observe that inductively we have that $\nabla R(\tilde{w}_t) = \theta_t$ since that is where we are accumulating the gradients. This means that

$$\operatorname*{argmin}_{w \in S} D_R(w||\tilde{w}_t) = \operatorname*{argmin}_{w \in S} R(w) - R(\tilde{w}_t) - \langle \nabla R(\tilde{w}_t), w - \tilde{w}_t \rangle = \operatorname*{argmin}_{w \in S} R(w) - \langle w, \theta_t \rangle$$

which is exactly what the FTRL algorithm is doing. □

Theorem 5. *OMD with regularizer R obtains the regret bound*

$$\eta \mathit{Regret}(T,u) \leq R(u) - R(w_1) + \sum_{t=1}^{T} D_{R^\star}(\theta_{t+1}||\theta_t)$$

Proof. We work only with linear loss functions since the convex case can be handled by first using Proposition 2. We know that

$$R^\star(\theta_{T+1}) \geq \langle u, \theta_{T+1}\rangle - R(u) = \eta \sum_{t=1}^{T} \langle u, -\ell_t\rangle - R(u)$$

by Fenchel-Young inequality. Now for the upper bound

$$\begin{aligned} R^\star(\theta_{T+1}) &= R^\star(\theta_1) + \sum_{t=1}^{T} R^\star(\theta_{t+1}) - R^\star(\theta_t) \\ &= R^\star(\theta_1) + \sum_{t=1}^{T} \nabla R^\star(\theta_t)(\theta_{t+1} - \theta_t) + D_{R^\star}(\theta_{t+1}||\theta_t) \\ &= R^\star(\theta_1) + \eta \sum_{t=1}^{T} \langle w_t, -\ell_t\rangle + D_{R^\star}(\theta_{t+1}||\theta_t) \end{aligned}$$

For the first term, since $R^\star(\theta_1) = \max\langle w, \theta_1\rangle - R(w) = \max -R(w)$ and $w_1 = \nabla R^\star(\theta_1)$ is the argmax, we get that $R^\star(\theta_1) = R(w_1)$. Now, re-arrang things to obtain the result. □

Example 1 (Quadratic regularizer)**.** *The quadratic regularizer $R(w) = \frac{1}{2}\|w\|_2^2$ has conjugate $\psi^\star(\theta) = \frac{1}{2}\|\theta\|_2^2$ with $\nabla\psi^\star(\theta) = \theta$. The Bregman divergence term*

$$D_{R^\star}(u||v) = \frac{1}{2}\|u\|_2^2 - \frac{1}{2}\|v\|_2^2 - \langle v, u - v\rangle = \frac{1}{2}\|u - v\|_2^2,$$

which happens to be symmetric. This reproduces the OGD analysis from last lecture.

Strong convexity. To get a better bound for the experts setting we need to use a new regularizer and we need to understand its properties. Since we have $S = \Delta(d)$, it makes some sense to use entropic regularization.

$$R(w) = \sum_{j=1}^{d} w_j \log w_j, \qquad R^\star(\theta) = \log(\sum_{j=1}^{d} \exp(\theta_j)), \qquad \nabla R^\star(\theta)_j = \frac{\exp(\theta_j)}{\sum_{k=1}^{d} \exp(\theta_j)}$$

This lead to the algorithm

$$w_t \propto \exp(\theta_t) = \exp(-\eta \sum_{j=1}^{t-1} \ell_j),$$

which is just the weighted majority algorithm. Here we will refer to it as the Exponentiated Gradient. The last thing to verify is

$$D_R(u||v) = KL(u||v) = \sum_j u_j \log(u_j/v_j),$$

which we will use later.

To complete the analysis we need to relate the dual bregman divergence to a norm. Working backwords, if we show that $R^\star$ is α-strongly smooth then we can upper bound the Bregman term. It turns out this is equivalent to asking that R is $1/\alpha$-strongly convex with respect to the dual norm. More formally

Definition 6. *ψ is α-strongly convex with respect to a norm $\|\cdot\|$ if for all u, v*

$$D_\psi(u||v) \geq \frac{\alpha}{2}\|u - v\|^2$$

Analogously ψ is α-strongly smooth with respect to a norm $\|\cdot\|$ if for all u, v

$$D_\psi(u||v) \leq \frac{\alpha}{2}\|u - v\|^2$$

For a norm $\|\cdot\|$ the dual norm is $\|x\|_\star = \sup_{y:\|y\|\leq 1}\langle x, y\rangle$.

Lemma 7. *$\psi(w)$ is $1/\alpha$ strongly convex with respect to some norm $\|\cdot\|$ if and only if $\psi^\star(\theta)$ is α strongly smooth with respect to the dual norm $\|\cdot\|_\star$*

Proposition 8. *The entropic regularizer $R(w) = \sum_j w_j \log w_j$ is 1 strongly convex with respect to the L_1 norm. Thus $R^\star$ is 1-strongly smooth with respect to the L_∞ norm and the second term in the regret is at most $\frac{\eta^2}{2}\sum_t \|\ell_t\|_\infty^2$, leading to an $O(\sqrt{T\log(d)})$ regret bound.*

Proof. The main thing we need to prove is that the entropic regularizer is 1 strongly convex. Let us just show that the KL-divergence is 1-strongly convex for the simpler case where $d = 2$, thus we must show

$$KL(p, q) = p\log p/q + (1-p)\log(1-p)/(1-q) \geq \frac{1}{2}\left(|p-q| + |(1-p)-(1-q)|\right)^2 = 2(p-q)^2$$

Let us look at the difference between these two sides as a function of q

$$\begin{aligned} g(q) &= p\log p/q + (1-p)\log(1-p)/(1-q) - 2(p-q)^2 \\ g'(q) &= -\frac{p}{q} + \frac{1-p}{1-q} + 4(p-q) = (q-p)\left[\frac{1}{q(1-q)} - 4\right] \end{aligned}$$

Since $q \in [0, 1]$ we know that $q(1-q) \leq 1/4$ and hence the derivative is negative for $q \leq p$ and positive for $q \geq p$. This means that p minimizes this function and it is easy to see that $g(p) = 0$, which proves this result. The easy way to generalize this is to observe that the total varation distance $\|p - q\|_1 = \max_{S\subset[d]} 2|P_S - Q_S|$ and also that the KL only contracts if we consider the binary distribution with probabilities $P_S = \sum_{j\in S} p_j$ and Q_S defined analogously. □