CSE 522: Algorithms and Uncertainty Spring 2017

Lecture 10 — May 5, 2017

Lecturer: Anna R. Karlin

Acknowledgement

Lectures 8-10 are drawn largely from the monographs of Elad Hazan and Shai Shalev-Schwartz on online optimization. The references in these notes are extremely incomplete. For full references see these two monographs as well as the notes on online optimization by Sebastien Bubeck.

1 Application: Exponentiated gradient

[Kivinen and Warmuth, 1997] We next consider instantiating FTRL using the entropic regularizer (i.e., the negative entropy function) with predictions confined to the probability simplex. Concretely for $\mathbf{p} \in \Delta_n$ where Δ_n is the probability simplex, we take our regularizer to be

$$R(\mathbf{p}) = \sum_{i=1}^{d} p_i \ln p_i.$$

To analyze this, we need a few preliminaries.

The entropic regularizer is strongly convex w.r.t. ℓ_1 norm

This regularizer is 1-strongly convex with respect to the ℓ_1 norm. To prove this, we need to show that

$$R(\mathbf{p}) - R(\mathbf{q}) - \nabla R(\mathbf{q}) \cdot (\mathbf{p} - \mathbf{q}) \geq \frac{1}{2}||\mathbf{p} - \mathbf{q}||_1^2. \tag{1}$$

The left hand side of the above is the **Bregman divergence** of the negative entropy function and is equal to (1)

$$\sum_i p_i \ln p_i - \sum_i q_i \ln q_i - \sum_i (\ln q_i + 1)(p_i - q_i) = \sum_i p_i \ln \left(\frac{p_i}{q_i}\right)$$

This is also called the **relative entropy** or KL-divergence between the distributions $\mathbf{p}$ and $\mathbf{q}$. It is a measure of the distance (though not symmetric) between two probability distributions.

Note that in the above definition we have

$$0 \log \frac{0}{0} = 0 \log \frac{0}{q} = 0 \quad \text{and} \quad p \log \frac{p}{0} = \infty.$$

Remark 1.1. We all know that the entropy of a random variable measures the uncertainty that exists as to the value of that random variable; the average amount of information received when the value of the random variable is observed, and also the average amount of "surprise" one receives upon learning the value of the r.v.. We also know that if we want to send messages distributed according to $\mathbf{p}$, then asymptotically the optimal encoding has expected length equal to the entropy.

Now, suppose that we think the distribution is $\mathbf{q}$, and we design the code optimally for that distribution, but we then use this encoding to send messages distributed according to $\mathbf{p}$. Then the relative entropy is the number of additional bits I will need to send relative to what I would have had to send if the code had been optimized for $\mathbf{p}$ to begin with. In terms of information, the relative entropy is a measure of the information gained when you revise your beliefs from your prior distribution $\mathbf{q}$ to your posterior distribution $\mathbf{q}$.

Thus, (1) is the same as

$$\sum_i p_i \ln\left(\frac{p_i}{q_i}\right) \geq \frac{1}{2}\left(\sum_{i=1}^{d} |p_i - q_i|\right)^2.$$

This is a famous inequality known as called **Pinsker's Inequality**. It says that

$$D(\mathbf{p}||\mathbf{q}) \geq \frac{||\mathbf{p}-\mathbf{q}||_1^2}{2}.$$

Notice that the right hand side is related to the total variation distance between the two distributions:

$$TV(\mathbf{p},\mathbf{q}) = \sup_A |P(A) - Q(A)| = \frac{1}{2}||\mathbf{p}-\mathbf{q}||_1.$$

Here is a proof for $n = 2$. Assume wlog that $p \geq q$. Then LHS of Pinsker is

$$\begin{aligned} D(p||q) &= \int_q^p \left(\frac{p}{x} - \frac{1-p}{1-x}\right) dx \\ &= \int_q^p \left(\frac{p-x}{x(1-x)}\right) dx \geq 4\int_q^p (p-x)dx = 2(p-q)^2. \end{aligned}$$

and the RHS is

$$1/2(p - q + (1-q) - (1-p))^2 = 1/2(4(p-q)^2) = 2(p-q)^2.$$

You can then handle the general case using a "chain rule" for relative entropy. See linked notes on relative entropy for details.

What FTRL does

Next we determine the probability vector $\mathbf{p}_{t+1}$ that FTRL plays at time $t+1$. One way is via Lagrange multipliers. Alternatively, we can simply observe that given

$$\boldsymbol{\theta} := -\sum_{s=1}^{t} \nabla_s,$$

where $\nabla_s = \nabla \ell_s(\mathbf{p}_s)$, in step $t+1$, FTRL will play

$$\mathsf{argmin}_{\mathbf{p}\in\Delta} - \eta\boldsymbol{\theta}\cdot\mathbf{p} + \sum_i p_i \ln p_i = \mathsf{argmin}_{\mathbf{p}\in\Delta} \sum_i p_i \ln\left(p_i e^{-\eta\theta_i}\right).$$

(Δ is the probability simplex.) Minimizing the above quantity is the same as minimizing

$$\sum_i p_i \ln\left(\frac{p_i Z}{e^{\eta\theta_i}}\right), \quad \text{where} \quad Z = \sum_j e^{\eta\theta_j}.$$

This is the same as minimizing

$$D(\mathbf{p}||\mathbf{q}) \quad \text{where} \quad q_i = \frac{e^{\eta\theta_i}}{Z}.$$

Relative entropy is minimized at $\mathbf{p} = \mathbf{q}$ since for that choice it is 0, and it is always nonnegative. (Any Bregman divergence is non-negative.)

Summary of what FTRL does

Since

$$\boldsymbol{\theta}_{t+1} = -\sum_{s=1}^{t} \nabla_s,$$

where $\nabla_s = \nabla \ell_s(\mathbf{p}_s)$, we get

$$p_{t+1,i} = \frac{p_{t,i} e^{-\eta \nabla \ell_t(\mathbf{p}_t)_i}}{\sum_{j=1}^{d} p_{t,j} e^{-\eta \nabla \ell_t(\mathbf{p}_t)_j}}.$$

When losses are linear, this is precisely Hedge.

Performance of exponentiated gradient

Applying the earlier theorem, we have

$$R_T(\mathbf{u}) \le 2GD\sqrt{T}$$

where $D = \sqrt{\max_{\mathbf{x},\mathbf{y}\in S}(R(\mathbf{x}) - R(\mathbf{y}))} = \sqrt{\log n}$, and

$$||\nabla \ell_t(\mathbf{p}_t)||_* = ||\nabla \ell_t(\mathbf{p}_t)||_\infty \le G.$$

As you all know $D \le \log n$ (with equality for the uniform distribution), but a quick proof uses the fact that $D(\mathbf{p}||\mathbf{q}) \ge 0$ with the substitution $\mathbf{q}$ equal to the uniform distribution.

In the experts settings, where $\ell_t(\mathbf{x}) = \ell_t \cdot \mathbf{x}$ and each component of ℓ_t is in $[0,1]$, we have

$$||\nabla \ell_t(\mathbf{p}_t)||_* \le ||\nabla \ell_t(\mathbf{p}_t)||_\infty \le 1.$$

Thus, we get the bound,

$$\text{Regret}_T \le 2\sqrt{T \log n}.$$

Thus bound is better than the bound we got using the Euclidean regularizer.

2 Another interpretation of FTRL

We now develop an alternative view of FTRL.

2.1 Digression on conjugate functions

There are two ways we can define a convex function: by its value at each point, or by the set of tangent planes.

Definition 2.1. For function f with domain S, the **conjugate function** or **Fenchel conjugate** of $f(\cdot)$ is defined as

$$f^*(\mathbf{y}) = \max_{\mathbf{w} \in S}(\mathbf{y} \cdot \mathbf{w} - f(\mathbf{w})).$$

It is easy to verify that if

$$\mathbf{x} = \text{argmax}_{\mathbf{w} \in S}(\mathbf{y} \cdot \mathbf{w} - f(\mathbf{w}))$$

then

$$\mathbf{y} \cdot \mathbf{x} - f^*(\mathbf{y})$$

is a supporting line of $f(\cdot)$ at $\mathbf{x}$, i.e. $\mathbf{y}$ is a subgradient at $\mathbf{x}$.

Remark 2.2. We will be interested in applying this to convex functions, but regardless the conjugate function is convex.

Examples

- If $f(x) = x \log x$ (where the domain is $\mathbb{R}^+$, with $f(0 = 0)$. Then $f^*(y) = e^{y-1}$.
- If $f(\mathbf{x}) = ||\mathbf{x}||$, then

$$f^*(\mathbf{y}) = \begin{cases} 0 & ||\mathbf{y}||_* \leq 1 \\ \infty & ||\mathbf{y}||_* > 1. \end{cases}$$

  To see this, observe that $f^*(\mathbf{y}) = \sup_{\mathbf{x}}(\mathbf{y}^T\mathbf{x} - ||x||)$. Consider two cases. If $||\mathbf{y}||_* \leq 1$, then $\mathbf{y}^T\mathbf{x} \leq ||\mathbf{x}||$ for all $\mathbf{x}$ of norm at most 1 with equality when $\mathbf{x} = 0$. Therefore the supremum above is 0.

  If $||\mathbf{y}||_* > 1$, then there is a $\mathbf{x}$ of norm at most 1 such that $\mathbf{x}^T\mathbf{y} > 1$. Thus

$$f^*(\mathbf{y}) \geq \mathbf{y}^T c\mathbf{x} - ||c\mathbf{x}|| = c(\mathbf{y}^T\mathbf{x} - ||\mathbf{x}||)$$

  which tends to infinity as $c \to \infty$.

The following facts are immediate from the definition of conjugate function.

- The conjugate function is always convex (maximum of affine functions).
- Fenchel-Young inequality:

$$f(\mathbf{x}) + f^*(\mathbf{y}) \leq \mathbf{x} \cdot \mathbf{y}.$$

- $f^{**}(x) \leq f(x)$.

You can find the following fact in any convex optimization book.

Claim 2.3. *If f is convex and closed (that is, the epigraph is a closed set), then $f^{**} = f$.*

Lemma 2.4. *Suppose that f is convex and closed. Then*

$$\mathbf{y} \in \partial f(\mathbf{x}) \iff \mathbf{x} \in \partial f^*(\mathbf{y}) \iff \mathbf{x} = \text{argmax}_z(\mathbf{y} \cdot \mathbf{z} - f(\mathbf{z})).$$

Proof. Suppose that $\mathbf{x} = \text{argmax}_z(\mathbf{y} \cdot \mathbf{z} - f(\mathbf{z}))$. Then

$$f^*(\mathbf{y}) = \mathbf{y} \cdot \mathbf{x} - f(\mathbf{x}).$$

Therefore,

$$f^*(\mathbf{w}) - f^*(\mathbf{y}) \geq (\mathbf{w}\mathbf{x} - f(\mathbf{x})) - (\mathbf{y}\mathbf{x} - f(\mathbf{x})) = \mathbf{x}(\mathbf{w} - \mathbf{y}).$$

Therefore, $\mathbf{x} \in \partial f^*(\mathbf{y})$. Similarly

$$f(\mathbf{z}) - f(\mathbf{x}) \geq \mathbf{y} \cdot \mathbf{z} - f^*(\mathbf{y}) - (\mathbf{y} \cdot \mathbf{x} - f^*(\mathbf{x})) = \mathbf{y} \cdot (\mathbf{z} - \mathbf{x})$$

so $\mathbf{y} \in \partial f(\mathbf{x})$.

On the other hand, $\mathbf{y} \in \partial f(\mathbf{x})$ implies that

$$f(\mathbf{z}) - f(\mathbf{x}) \geq \mathbf{y}(\mathbf{z} - \mathbf{x}) \quad \forall \mathbf{z} \in S$$

and therefore

$$\mathbf{y}\mathbf{x} - f(\mathbf{x}) \geq \mathbf{y} \cdot \mathbf{z} - f(\mathbf{z}) \iff \mathbf{x} \in \text{argmax}_{\mathbf{z} \in S}(\mathbf{y}\mathbf{z} - f(\mathbf{z})).$$

A similar argument applies for the remaining implication. □

Remark 2.5. The conjugate of a differentiable function f on $\mathbb{R}^n$ is called the Legendre transform of the function. In this case ∇f and ∇f^* are inverses.

$$\mathbf{z} = \nabla f^*(\nabla f(\mathbf{z})).$$

(And we can write the function $f^*(\mathbf{y}) = \mathbf{z} \cdot \nabla f(\mathbf{z}) - f(\mathbf{z})$, where $\mathbf{y} = \nabla f(\mathbf{z})$.)

Claim 2.6. *Let R be a β-strongly convex, closed function (from $\mathbb{R}^n$ to $\mathbb{R}$) w.r.t. norm $||\cdot||$. Then R^* is β^{-1}-smooth with respect to $||\cdot||_*$.*

Proof. Let

$$\mathbf{w}_\mathbf{y} = \text{argmax}_\mathbf{w}(\mathbf{y} \cdot \mathbf{w} - R(\mathbf{w})).$$

$$R^*(\mathbf{x}) - R^*(\mathbf{y}) \leq \max_\mathbf{w}(\mathbf{x} \cdot \mathbf{w} - R(\mathbf{w})) - (\mathbf{y} \cdot \mathbf{w}_\mathbf{y} - R(\mathbf{w}_\mathbf{y}))$$

By strong convexity of R,

$$-(R(\mathbf{w}) - R(\mathbf{w}_\mathbf{y})) \leq -\left(\nabla R(\mathbf{w}_\mathbf{y}) \cdot (\mathbf{w} - \mathbf{w}_\mathbf{y}) + \frac{\beta}{2}||\mathbf{w} - \mathbf{w}_\mathbf{y}||^2\right)$$

and $\nabla R(\mathbf{w_y}) = \mathbf{y}$, so

$$R^*(\mathbf{x}) - R^*(\mathbf{y}) \leq \max_{\mathbf{w}} \left(\mathbf{x} \cdot \mathbf{w} - \mathbf{y} \cdot \mathbf{w_y} - \left(\mathbf{y} \cdot (\mathbf{w} - \mathbf{w_y}) + ||\mathbf{w} - \mathbf{w_y}||^2 \frac{\beta}{2} \right) \right).$$

Letting $\mathbf{w} - \mathbf{w_y} := \mathbf{z}$, the right hand side is equal to

$$(\mathbf{x} - \mathbf{y}) \cdot \mathbf{w_y} + \max_{\mathbf{z}} \left[(\mathbf{x} - \mathbf{y}) \cdot \mathbf{z} - \frac{\beta}{2} ||\mathbf{z}||^2 \right].$$

The maximization is at $\mathbf{x} - \mathbf{y} = \beta \mathbf{z}$, and plugging back in yields

$$(\mathbf{x} - \mathbf{y}) \cdot \mathbf{w_y} + \frac{1}{\beta} ||\mathbf{x} - \mathbf{y}||^2 - \frac{1}{2\beta} ||\mathbf{x} - \mathbf{y}||^2.$$

Thus, since $\mathbf{w_y} = \nabla R^*(\mathbf{y})$, we conclude that

$$R^*(\mathbf{x}) - R^*(\mathbf{y}) \geq (\mathbf{x} - \mathbf{y}) \cdot \nabla R^*(\mathbf{y}) + \frac{1}{2\beta} ||\mathbf{x} - \mathbf{y}||^2.$$

In other words, R^* is β^{-1}-smooth. □

Also, the convex dual is everywhere differentiable.

2.2 Back to FTRL

So what do conjugate functions have to do with FTRL?

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathsf{argmin}_{\mathbf{w} \in S}(-\theta_{t+1} \cdot \mathbf{w} + R(\mathbf{w})) \\ &= \mathrm{argmax}_{\mathbf{w} \in S}(\theta_{t+1} \cdot \mathbf{w} - R(\mathbf{w})) \\ &= \nabla R^*(\theta_{t+1}), \end{aligned}$$

where

$$\theta_{t+1} = -\eta \sum_{s=1}^{t} \nabla \ell_s(\mathbf{w}_s).$$

Remark 2.7. If follows from 2.4 that $\theta_{t+1} = \nabla R(\mathbf{w}_{t+1})$.

3 FTRL = Online mirror descent [Nemirovsky and Yudin, 1983]

This algorithm is also called "online mirror descent"; it updates the gradient of the dual of regularizer at sum of gradients of past losses and can be described as follows

Parameters: Strongly convex regularizer R, and $\eta > 0$.

Initialization: $\mathbf{z}_1$ s.t. $\nabla R(\mathbf{z}_1) = 0$ and $\mathbf{x}_1 = \mathsf{argmin}_{w \in S} B_R(\mathbf{w} || \mathbf{z}_1)$

For each $t \geq 1$,

1. Play $\mathbf{x}_t$ and incur loss $\ell_t(\mathbf{x}_t)$
2. Observe loss gradient $\nabla \ell_t(\mathbf{x}_t)$.
3. Gradient step: Update $\mathbf{z}_t$.

$$\nabla R(\mathbf{z}_{t+1}) = \nabla R(\mathbf{z}_t) - \eta \nabla \ell_t(\mathbf{x}_t). \tag{2}$$

   Let

$$\mathbf{z}_{t+1} = \nabla R^* \left(\nabla R(\mathbf{z}_t) - \eta \nabla \ell_t(\mathbf{x}_t) \right).$$

4. Projection step:

$$\mathbf{x}_{t+1} = \mathsf{argmin}_{\mathbf{w} \in S} B_R(\mathbf{w} || \mathbf{z}_{t+1}).$$

3.1 Equivalence of (lazy) OMD and FTRL

Lemma 3.1.

$$\mathsf{argmin}_{\mathbf{w} \in S} B_R(\mathbf{w} || \mathbf{z}_t) = \mathsf{argmin}_{\mathbf{w} \in S} \left(-\eta \sum_{s=1}^{t-1} \nabla_s \cdot \mathbf{w} + R(\mathbf{w}) \right).$$

Proof. The unconstrained minimum $\mathbf{w}_t^*$ on the right hand side satisfies

$$\nabla R(\mathbf{w}_t^*) = -\eta \sum_{s=1}^{t-1} \nabla_s. \tag{3}$$

By definition (2), $\mathbf{z}_t$ also satisfies this. Since $R(\mathbf{w})$ is strictly convex, there is only one solution, therefore $\mathbf{z}_t = \mathbf{w}_t^*$.

Now consider the Bregman divergence. We have

$$\begin{aligned} B_R(\mathbf{w} || \mathbf{z}_t) &= R(\mathbf{w}) - R(\mathbf{z}_t) - \nabla R(\mathbf{z}_t) \cdot (\mathbf{w} - \mathbf{z}_t) \\ &= R(\mathbf{w}) - R(\mathbf{z}_t) + \eta \sum_{s=1}^{t-1} \nabla_s \cdot (\mathbf{w} - \mathbf{z}_t) \end{aligned}$$

by (3). Therefore, $B_R(\mathbf{w} || \mathbf{z}_t)$ is minimized at the minimum of

$$R(\mathbf{w}) + \eta \sum_{s=1}^{t-1} \nabla_s \cdot \mathbf{w}$$

in S, which is precisely $\mathbf{w}_t$ from FTRL. Therefore, $\mathbf{x}_t = \mathbf{w}_t$.

□

Remark 3.2. FTRL is not necessarily the same as the "agile", i.e., non-lazy, version, of OMD where the gradient step replaces (2) with

$$\nabla R(\mathbf{z}_{t+1}) = \nabla R(\mathbf{x}_t) - \eta \nabla \ell_t(\mathbf{x}_t).$$

Often when people talk about mirror descent, they mean the agile version.

3.2 Notes

The offline version of this algorithm was discovered in the 70s by Nemirovski and Yudin. Their goal was to minimize a convex function with some Lipschitz constant L with respect to some norm.

The mirror descent idea is perhaps motivated by thinking about the approximate change in function value $f(y) - f(x)$ when the algorithm takes a step, which can be approximated by $\nabla f(x) \cdot (y - x)$. This quantity is the dot product of a vector that lives in the "primal space", the space of points, where distances are measured using one norm, and the latter in the "dual space", the space of gradients, where distances are measured using the dual norm.

If you think about gradient descent, it follows directions of steepest descent, directions opposite to gradient. This leads to an iteration of the type $y_{t+1} = y_t - \eta \nabla f(y_t)$. In a sense this is an apples and oranges combination since the objects y_t and $\nabla f(y_t)$ live in these two different spaces. This becomes strange if different norms are being used. It's fine when everything is being measured in terms of the ℓ_2 norm.

In MD, we keep two vectors w_t, θ_t, one in the primal space and one in the dual space. In each iteration we compute $\nabla f(t)$, obtaining a dual vector and the update $\theta_{t+1} = \theta_t - \eta \nabla f(w_t)$ in the dual space. We map between primal and dual points using a *mirror map*. If we think of this map as the gradient of a convex function $g^*(\cdot)$, then the inverse map is the gradient of the conjugate function.

4 OMD version of Exponentiated Gradient Descent

Say $R(\mathbf{x}) = \mathbf{x} \log \mathbf{x} = \sum_i x_i \log x_i$.Then

$$\nabla R(\mathbf{x})_i = 1 + \log x_i.$$

Hence, OMD becomes the following:

- $\log(\mathbf{z}_{t+1}) = \log(\mathbf{z}_t) - \eta \nabla_t$. Thus,
$$\mathbf{z}_{t+1} = e^{-\eta \nabla_t} \mathbf{z}_t.$$
- $\mathbf{x}_{t+1} = \mathsf{argmin}_{\mathbf{w} \in S} B_R(\mathbf{w} || \mathbf{z}_t)$. Projection according to negative entropy is the same as scaling by the ℓ_1 norm, which gives exactly what we had above, since
$$B_R(\mathbf{p} || \mathbf{z}) = \mathbf{p} \ln \left(\frac{\mathbf{p}}{\mathbf{z}} \right).$$