CSE 522: Algorithms and Uncertainty Spring 2017

Lecture 9 — May 1, 2017

Lecturer: Anna R. Karlin

1 Follow the regularized leader

As with offline optimization, there are numerous algorithms with convergence rates that depend on properties of the function to be optimized. We will now develop another online convex optimization algorithm that achieves better bounds in some cases, and unifies a number of disparate results in the field.

Here are some key references: [Shalev-Shwartz 2007] [Abernethy, Hazan and Rakhlin, 2008]

The following algorithm unifies many of the ideas that had been used prior for online convex optimization.

Definition 1.1. The **Follow the Regularized Leader (FTRL)** strategy is to set

$$\mathbf{w}_{t+1} = \mathsf{argmin}_{\mathbf{w}\in S}\left[\eta\sum_{s=1}^{t}\ell_s(\mathbf{w}) + R(\mathbf{w})\right],$$

where $R(\cdot)$ is a **strongly convex** regularizer (see definition below).

Many of you are familiar with regularization as a technique for avoiding overfitting. The idea is that models that overfit the data are too complicated, e.g. have too many parameters. One way to penalize this complexity is to add a "regularization" term to the loss function that penalizes more complex models.

Here the regularization term is playing a similar role. It's going to help ensure stability, so that these points that we play don't move around too fast.

1.1 Digression on some basic convex optimization related definitions

1.1.1 Strong convexity and smoothness

Definition 1.2. Let $f : S \to \mathbb{R}$ be a convex function, where S is a convex set. Then $f(\cdot)$ is β-**strongly convex** w.r.t. norm $||\cdot||$ if for all $\mathbf{u}, \mathbf{v} \in S$

$$f(\mathbf{v}) \geq f(\mathbf{u}) + \nabla f(\mathbf{u})\cdot(\mathbf{v}-\mathbf{u}) + \frac{\beta}{2}||\mathbf{u}-\mathbf{v}||^2.$$

In other words, the function grows faster than linearly in every direction.

Definition 1.3. Suppose that R is a convex function, Then the **Bregman divergence** $B_f(\mathbf{x}||\mathbf{y})$ is defined as follows

$$B_f(\mathbf{x}||\mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})\cdot(\mathbf{x}-\mathbf{y})$$

The Bregman divergence from $\mathbf{x}$ to $\mathbf{y}$ is the difference between $f(\mathbf{x})$ and its linear approximation via the first-order Taylor expansion of f around $\mathbf{y}$. Since the function is convex, this is always nonnegative. It is not symmetric, but behaves in many ways like a distance function.

Remark 1.4. We can rewrite the definition of β-strongly convex as follows:

$$B_f(\mathbf{u}||\mathbf{v}) \geq \frac{\beta}{2}||\mathbf{u}-\mathbf{v}||^2. \tag{1}$$

The following are immediate consequences of the definition of a strongly convex function.

Claim 1.5. *Suppose that $f(\cdot)$ is β strongly convex. Then*

- *If g is convex, then $f+g$ is β strongly convex.*
- *The function cf is $c\beta$ strongly convex.*
- *If $\mathbf{u}$ is a minimizer of f, then for all $\mathbf{w}$*

$$f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{\lambda}{2}||\mathbf{w}-\mathbf{u}||^2.$$

Definition 1.6. Let $f : S \to \mathbb{R}$ be a convex function, where S is a convex set. Then $f(\cdot)$ is **α-smooth** w.r.t. norm $||\cdot||$ if for all $\mathbf{u}, \mathbf{v} \in S$

$$f(\mathbf{v}) \leq f(\mathbf{u}) + \nabla f(\mathbf{u}) \cdot (\mathbf{v}-\mathbf{u}) + \frac{\alpha}{2}||\mathbf{u}-\mathbf{v}||^2.$$

Claim 1.7. *1. If $f(\cdot)$ is β-strongly convex with respect to the Euclidean norm, then*

$$\nabla^2 R \succeq \beta I.$$

2. If $f(\cdot)$ is α-smooth, then

$$\nabla^2 R \preceq \alpha I.$$

1.2 Digression on dual norms

We all know the importance of the Cauchy-Schwartz inequality for the 2-norm. There is a classic generation known as Holder's inequality.

$$(\mathbf{x}, \mathbf{y}) \leq ||\mathbf{x}||_p ||\mathbf{y}||_q$$

where

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Recall that

$$||\mathbf{x}||_p = \left(\sum_i |x_i|^p \right)^{1/p}.$$

Motivates the following general definition.

Definition 1.8. Let $||\cdot||$ denote a norm.[1] Then the **dual norm**

$$||\mathbf{y}||_* := \max_{||\mathbf{x}||\leq 1} (\mathbf{x}, \mathbf{y}).$$

For example, the dual norm to the 2-norm is the 2-norm. From Holder's inequality it is easy to show that the dual norm to the p-norm is the q norm where $1/p + 1/q = 1$. For example, the dual norm to the 1-norm is the infinity norm. Consider the matrix norm $||x||_A$, where A is p.s.d. defined by

$$||\mathbf{x}||_A := \sqrt{\mathbf{x}^T A \mathbf{x}}.$$

Then the dual norm is $||\mathbf{x}||_{A^{-1}}$.

Proposition 1.9. *By the definition of the dual norm, for any norm $||\cdot||$,*

$$(\mathbf{x}, \mathbf{y}) \leq ||x|| \cdot ||y||_*.$$

Definition 1.10. A function $f : S \to \mathbb{R}$ is L-Lipschitz with respect to norm $||\cdot||$ if for any $\mathbf{u}, \mathbf{v} \in S$,

$$f(\mathbf{u}) - f(\mathbf{v}) \leq L||\mathbf{u} - \mathbf{v}||.$$

Claim 1.11. *Fix a convex continuous function $\ell : S \to \mathbb{R}$. Assume S is full dimensional. Then $\ell(\cdot)$ is L-Lipschitz over S with respect to norm $||\cdot||$ iff for all $\mathbf{w}$ in the interior of S and $\mathbf{z} \in \partial\ell(\mathbf{w})$, we have $||\mathbf{z}||_* \leq L$, where $||\cdot||_*$ is the dual norm.*

Proof. Suppose that ℓ is Lipschitz. Pick $\mathbf{w}$ and $\mathbf{z}$ as in statement. let $\mathbf{u}$ such that

$$\mathbf{u} - \mathbf{w} = \text{argmax}_{\mathbf{v} \mid ||\mathbf{v}||=1}(\mathbf{v}, \mathbf{z}).$$

Then

$$(\mathbf{u} - \mathbf{w}, \mathbf{z}) = ||\mathbf{z}||_*.$$

By the definition of a subgradient

$$||\mathbf{z}||_* = (\mathbf{z}, \mathbf{u} - \mathbf{w}) \leq \ell(\mathbf{u}) - \ell(\mathbf{w}) \leq L||\mathbf{u} - \mathbf{w}|| = L.$$

For the other direction, since $\mathbf{z}$ is a subgradient at $\mathbf{w}$, we have

$$\ell(\mathbf{w}) - \ell(\mathbf{u}) \leq (\mathbf{z}, \mathbf{w} - \mathbf{u}) \leq ||\mathbf{z}||_*||\mathbf{w} - \mathbf{u}|| \leq L||\mathbf{w} - \mathbf{u}||.$$

Therefore ℓ is L-Lipschitz. □

1.3 Back to FTRL

Recall the definition

[1] We are talking about norms $n(\cdot)$ over $\mathbb{R}^n$. Such norms are required to satisfy the following for any constant c, and two vectors $\mathbf{u}$ and $\mathbf{v}$: (a) $n(\mathbf{v}) \geq 0$, (b) $n(c\mathbf{v}) = |c|n(\mathbf{v})$, (c) triangle inequality, (e) $n(\mathbf{0}) = 0$.

Definition 1.12. The **Follow the Regularized Leader (FTRL)** strategy is to set

$$\mathbf{w}_{t+1} = \mathsf{argmin}_{\mathbf{w}\in S}\left[\eta\sum_{s=1}^{t}\ell_s(\mathbf{w}) + R(\mathbf{w})\right],$$

where $R(\cdot)$ is a **strongly convex** regularizer.

Lemma 1.13. *Suppose that $R(\cdot)$ is a β strongly convex regularizer with respect to some norm $||\cdot||$. Let*

$$G_t := ||\nabla\ell_t(\mathbf{w}_t)||_*.$$

Then

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \leq G_t||\mathbf{w}_t - \mathbf{w}_{t+1}|| \leq \frac{\eta G_t^2}{\beta}.$$

Proof. Define

$$F_t(\mathbf{w}) = \sum_{i=1}^{t-1}\ell_i(\mathbf{w}) + \frac{R(\mathbf{w})}{\eta}.$$

By definition

$$\mathbf{w}_t = \mathsf{argmin}_{\mathbf{w}\in S}F_t(\mathbf{w}).$$

Since F_t is β/η strongly convex, then by Lemma 1.5, we have

$$F_t(\mathbf{w}_{t+1}) \geq F_t(\mathbf{w}_t) + \frac{\beta}{2\eta}||\mathbf{w}_t - \mathbf{w}_{t+1}||^2$$
$$F_{t+1}(\mathbf{w}_t) \geq F_{t+1}(\mathbf{w}_{t+1}) + \frac{\beta}{2\eta}||\mathbf{w}_t - \mathbf{w}_{t+1}||^2$$

Adding these up we get

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \geq \frac{\beta}{\eta}||\mathbf{w}_t - \mathbf{w}_{t+1}||^2.$$

We also have

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) \leq (\nabla\ell_t, \mathbf{w}_t - \mathbf{w}_{t+1}) \leq ||\nabla\ell_t(\mathbf{w}_t)||_*||\mathbf{w}_t - \mathbf{w}_{t+1}|| = G_t||\mathbf{w}_t - \mathbf{w}_{t+1}||,$$

so

$$||\mathbf{w}_t - \mathbf{w}_{t+1}|| \leq \frac{G_t\eta}{\beta}$$

and we get the lemma. □

We now use this in the BTL-FTL lemma, which you recall implied that

$$\sum_{t=1}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u})) \leq \sum_{t=1}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}))$$

We can similarly show that for FTRL

Lemma 1.14.

$$\sum_{t=1}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u})) \leq \frac{R(\mathbf{u}) - R(\mathbf{w}_1)}{\eta} + \sum_{t=1}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1})$$

Proof. FTRL is exactly like FTL with an extra round where in step 0, the loss function is $R(\mathbf{x})/\eta$. Applying the BTL lemma (with $\ell_0 = R/\eta$ for FTRL). Therefore, we have

$$-\sum_{t=0}^{T}\ell_t(\mathbf{u}) \leq -\sum_{t=0}^{T}\ell_t(\mathbf{w}_{t+1}).$$

Adding $\sum_{t=0}^{T}\ell_t(\mathbf{w}_t)$ to both sides we get

$$\sum_{t=0}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u})) \leq \sum_{t=0}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}).$$

Thus, with $f_0 := R/\eta$, we have

$$\sum_{t=1}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u})) + f_0(\mathbf{w}_0) - f_0(\mathbf{u}) \leq \sum_{t=1}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) + f_0(\mathbf{w}_0) - f_0(\mathbf{w}_1)$$

or equivalently.

$$\sum_{t=1}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u})) \leq \frac{R(\mathbf{u}) - R(\mathbf{w}_1)}{\eta} + \sum_{t=1}^{T}(\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1})$$

□

Combining the previous two lemmas, we get the following corollary.

Corollary 1.15. *Suppose that the regularizer $R(\cdot)$ is β-strongly convex and $\ell_1, \ell_2, \ldots$ are arbitrary convex losses with respect to some norm. Then*

$$\text{Regret}_T(\mathbf{u}) \leq \frac{R(\mathbf{u}) - \min_{w\in S} R(\mathbf{w})}{\eta} + \sum_{t=1}^{T}\frac{\eta||\nabla\ell_t(\mathbf{w}_t)||_*^2}{\beta},$$

for all $\mathbf{u} \in S$.

If ℓ_t is L_t Lipschitz for each t, then by Corollary 1.11, we can also say that

$$\text{Regret}_T(\mathbf{u}) \leq \frac{R(\mathbf{u}) - \min_{w\in S} R(\mathbf{w})}{\eta} + \sum_{t=1}^{T}\frac{\eta L_t^2}{\beta},$$

for all $\mathbf{u} \in S$.

Proof. The first part follows immediately from plugging the result of Lemma 1.13 into Lemma 1.14.

The second part follows by observing that we can use the fact that boundedness of gradients of ℓ_t w.r.t. $||\cdot||_*$ equivalent to Lipschitzness of ℓ_t w.r.t. $||\cdot||$, as per Claim 1.11. For concreteness here: $||\nabla \ell_t(\mathbf{w}_t)||_*$ is a lower bound on the Lipschitz constant of ℓ. Indeed, let $\mathbf{u}$ such that

$$\mathbf{u} - \mathbf{w}_t = \text{argmax}_{\mathbf{v}|||\mathbf{v}||=1}(\mathbf{v}, \nabla \ell_t(\mathbf{w}_t)) = ||\nabla \ell_t(\mathbf{w}_t)||_*.$$

Then

$$||\nabla \ell_t(\mathbf{w}_t)||_* = (\nabla \ell_t(\mathbf{w}_t), \mathbf{u} - \mathbf{w}_t) \leq \ell_t(\mathbf{u}) - \ell_t(\mathbf{w}_t) \leq L_t||\mathbf{u} - \mathbf{w}_t|| = L_t.$$

□

Corollary 1.16. *Suppose that* $\max_t L_t \leq L$ *and* $\sqrt{\sup_{\mathbf{u},\mathbf{w}\in S} R(\mathbf{u}) - R(\mathbf{w})} = D$. *Then optimizing over* η *gives*

$$R_T(\mathbf{u}) \leq \frac{2LD}{\sqrt{\beta}}\sqrt{T}.$$

Remark 1.17. This gives us some clue as to how we want to choose the regularizer and the norm. There are two factors that we trade off, the diameter D of the underlying space, and the choice of a norm for which the losses have small Lipschitz constant.

2 Applications of FTRL

2.1 Euclidean regularization = Online gradient descent (for linear optimization)

Choose

$$R(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||_2^2.$$

Suppose that

$$\ell_t(\mathbf{w}) = (\mathbf{w}, \mathbf{z}_t).$$

Then when you solve for

$$\nabla \left(\sum_{i=1}^{t} (\mathbf{w}, \mathbf{z}_i) + \frac{1}{2\eta}||\mathbf{w}||_2^2 \right) = 0$$

to get $\mathbf{w}_{t+1}$, you get

$$\mathbf{w}_{t+1} = -\eta \sum_{i=1}^{t} \mathbf{z}_i = \mathbf{w}_t - \eta \mathbf{z}_t = \mathbf{w}_t - \eta \nabla \ell_t(\mathbf{w}_t).$$

If $\mathbf{w}_{t+1}$ is not in S, then one can show (you should check this) that

$$\mathsf{argmin}_{\mathbf{w}\in S} \left[\eta \sum_{i=1}^{t} (\mathbf{w}, \mathbf{z}_i) + \frac{1}{2}||\mathbf{w}||_2^2 \right],$$

is obtained by finding the closest point in S to $\mathbf{w}_{t+1}$ (closest in Euclidean distance). Thus, FTRL with Euclidean regularization is Online Gradient Descent.

Corollary 2.1. *Consider OGD with loss functions that are linear, specifically* $\ell_t(\mathbf{w}) = (\mathbf{w}, \mathbf{z}_t)$*, and suppose that*

$$S = \{\mathbf{w} \mid ||\mathbf{w}||_2 \leq B\} \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T} ||\mathbf{z}_t||_2^2 \leq L^2.$$

Then for all $\mathbf{u}$

$$\text{Regret}_T(\mathbf{u}) \leq \sqrt{2}LB\sqrt{T}.$$

Proof. By Corollary 1.16, with $R(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||_2^2$, we have $\beta = 1$ and $D = \sqrt{\max_{\mathbf{u},\mathbf{v}} R(\mathbf{u}) - R(\mathbf{v})} = B/\sqrt{2}$. Also, since $\ell_t(\cdot)$ is Lipschitz with $L_t = ||\mathbf{z}_t||_2$, and $\sum_{t=1}^{T} ||\mathbf{z}_t||_2^2 \leq L^2 T$, this corollary follows from Corollary 1.16. □

Remark 2.2. Consider OGD applied to the experts problem with losses in $[0, 1]$. This means that S is the probability simplex in n dimensions, and $\ell_t(\mathbf{w}) = (\ell_t, \mathbf{w})$, where $\ell_t \in [0, 1]^n$. Then $L = \sqrt{n}$ and $B = 1/\sqrt{2}$. Thus, the regret bound becomes $\sqrt{2nT}$.

Next we generalize the result to to allow any sequence of Lipschitz functions rather than linear functions with bounded norm.

2.2 Using the loss gradient

For all practical purposes, we only ever need to deal with linear functions, since we can linearize the convex losses.

$$\begin{aligned}\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) &\leq \nabla\ell_t(\mathbf{w}_t) \cdot \mathbf{w}_t - \nabla\ell_t(\mathbf{w}_t) \cdot \mathbf{u} \\ &= \nabla_t \cdot \mathbf{w}_t - \nabla_t \cdot \mathbf{u},\end{aligned}$$

where

$$\nabla_t := \nabla\ell_t(\mathbf{w}_t).$$

Let $R(\cdot)$ be a **strongly convex** regularizer and $\eta > 0$ is a scale parameter. The **Follow the Regularized Leader (FTRL) with linearized loss** strategy becomes the following:

$$\begin{aligned}\mathbf{w}_{t+1} &= \mathsf{argmin}_{\mathbf{w}\in S}\left[\sum_{s=1}^{t} \eta\ \nabla_s \cdot \mathbf{w} + R(\mathbf{w})\right], \\ &= \mathsf{argmin}_{\mathbf{w}\in S}\left[-\theta_{t+1} \cdot \mathbf{w} + R(\mathbf{w})\right],\end{aligned}$$

where

$$-\theta_{t+1} := \eta \sum_{s=1}^{t} \nabla_s.$$

2.2.1 FTRL = OGD for general convex functions

Again choose

$$R(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||_2^2.$$

Then as above, this algorithm specializes to OGD.

Lemma 2.3. *Consider OGD with general convex loss functions and suppose that*

$$S = \{\mathbf{w} \mid ||\mathbf{w}||_2 \leq B\} \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T} ||\nabla \ell_t(\mathbf{w}_t)||_2^2 \leq L^2.$$

Then for all $\mathbf{u}$

$$\text{Regret}_T(\mathbf{u}) \leq \sqrt{2}LB\sqrt{T}.$$

Proof. Same as before. □