Algorithms and Uncertainty, Winter 2018/19 Lecture 19 (5 pages)

Basics of Online Convex Optimization, Part 1

Instructor: Thomas Kesselheim

Today, we will get to know a much larger framework for online learning. Indeed, the experts setting will come back as a special case and also the multiplicative-weights algorithm. As a motivating example, we will consider linear regression. In its simplest case, one is given a number of pairs $(x^{(t)}, y^{(t)})$ of data points. One then computes a line, defined by a slope w_1 and a y-intercept w_2 so as to minimize the squared error $\sum_t \left(w_1 x^{(t)} + w_2 - y^{(t)}\right)^2$. One can then use this line to predict the y-label given an x-coordinate.

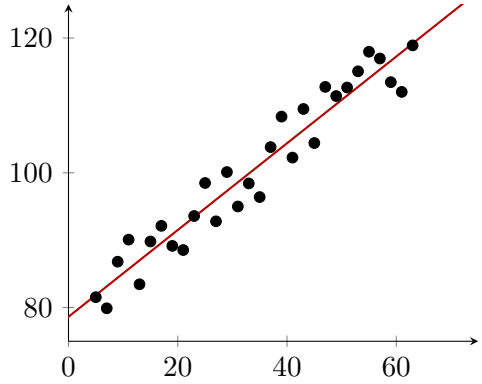

Figure 1: The red line is the regression line, which minimized the sum of the squared errors.

We turn this problem into an online problem as follows. We will see the data points $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots$ one after the other. Indeed, we will first only see $x^{(t)}$ and have to predict $y^{(t)}$ before we get to know its actual value. That is, we already have to predict while learning.

One could also picture a kind of "bandit" feedback for this model: Instead of getting to know the actual $y^{(t)}$, we only get to know how far we are off the actual value.

1 General Setup

We consider the following round-based problem. We will have to optimize a sequence of a priori unknown functions $f_1, \dots, f_T$. Each f_t maps from set S to the real numbers. The set $S \subseteq \mathbb{R}^d$ is a set of d-dimensional real vectors.

In each step $t \in \{1, \dots, T\}$, we choose a point $\mathbf{w}^{(t)} \in S$. Only afterwards, we get to know f_t and incur a cost of $f_t(\mathbf{w}^{(t)})$.

The *regret* of a sequence $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(T)}$ is defined as before as the amount by which our decisions are more expensive than the best single point in hindsight. That is,

$$\text{Regret}^{(T)} = \sum_{t=1}^{T} f_t(\mathbf{w}^{(t)}) - \min_{\mathbf{u} \in S} \sum_{t=1}^{T} f_t(\mathbf{u}) \ .$$

Example 19.1. *To capture our example of simple linear regression, we can set $d = 2$ and $S = \mathbb{R}^2$. A point $(w_1, w_2) \in S$ corresponds to the slope w_1 and the y-intercept w_2 of the*

regression line. A function f_t is the square of the error that we make on the t-th sample, depending on which w_1 and w_2 we use. So

$$f_t(w_1, w_2) = \left(w_1 x^{(t)} + w_2 - y^{(t)}\right)^2 \quad .$$

Note that the best single (w_1, w_2) in hindsight corresponds to exactly the optimal regression line.

If the set S is finite, we could run, for example, the Randomized Weighted Majority algorithm. In our regression example it is infinite. Instead we will assume that S and the functions are convex.

2 Convex Sets, Convex Functions, and Gradients

We assume that each function f_t is differentiable[1] and convex.

The typical example one should keep in mind is a quadratic function in one dimension (see Figure 2). One way to define convexity in this setting is to require that the function never falls below its tangents. This is expressed in terms of the derivative as follows. For all $u, v \in S$ we have to have

$$f(u) \geq f(v) + f'(v)(u - v) \quad .$$

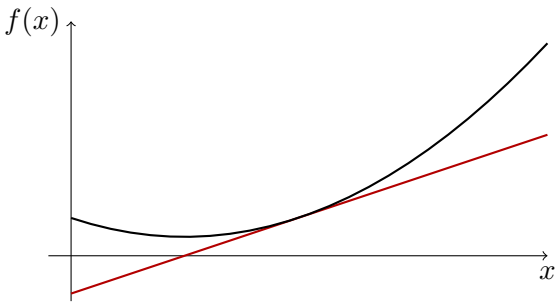

Figure 2: A typical convex function in one dimension, including a tangent in red.

In multiple dimensions, the idea is just the same. The function f now has a gradient ∇f, which is defined to be the vector of all partial derivatives; $(\nabla f(\mathbf{u}))_i = \frac{\partial f}{\partial u_i}(\mathbf{u})$. A function f is convex if never falls below the tangent hyperplane (see Figure 3). That is for all $\mathbf{u}, \mathbf{v}$

$$f(\mathbf{u}) \geq f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), (\mathbf{u} - \mathbf{v}) \rangle \quad . \tag{1}$$

Here $\langle \cdot, \cdot \rangle$ denotes the inner product, defined by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{d} x_i y_i$.

Example 19.2. *Another important—and familiar—example is the following. Set $S = \{\mathbf{v} \mid \sum_{i=1}^{d} v_i = 1\}$. The functions f_t are linear. That is, $f_t(\mathbf{v}) = \sum_{i=1}^{d} \ell_i^{(t)} v_i$ for some $\ell_i^{(t)} \in \mathbb{R}$. These functions are clearly convex. And we already know this setting: It's the experts setting with d experts and the vectors $\mathbf{v} \in S$ correspond to probability distributions over experts.*

[1] None of the results actually requires differentiability but the exposition gets a lot easier.

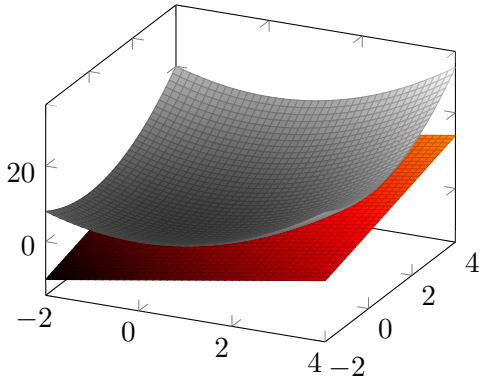

Figure 3: A convex function in two dimensions.

3 Follow-the-Leader

A very natural algorithm is the following *Follow the Leader*: In every step t, choose the point $\mathbf{w}^{(t)}$ that would have resulted in the cheapest cost up to now, that is, set $\mathbf{w}^{(t)}$ to $\mathbf{v}$ such that $\sum_{t'=1}^{t-1} f_{t'}(\mathbf{v})$ is minimal. The point $\mathbf{w}^{(1)}$ is arbitrary.

What we would actually want to do is to also include the function f_t in the sum because this determines the actual cost in step t. Unfortunately, we do not know it when choosing $\mathbf{w}^{(t)}$ but only when choosing $\mathbf{w}^{(t+1)}$. Our first observation is that we can bound the regret by the distances of $\mathbf{w}^{(t)}$ and $\mathbf{w}^{(t+1)}$.

Lemma 19.3. *For Follow-the-Leader, we have*

$$\text{Regret}^{(T)} \leq \sum_{t=1}^{T} (f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^{(t+1)})) \ .$$

Proof. We have to show that for all $T \geq 0$

$$\sum_{t=1}^{T} f_t(\mathbf{w}^{(t)}) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \sum_{t=1}^{T} (f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^{(t+1)})) \qquad \text{for all } u \in S \ ,$$

or equivalently

$$\sum_{t=1}^{T} f_t(\mathbf{u}) \geq \sum_{t=1}^{T} f_t(\mathbf{w}^{(t+1)}) \qquad \text{for all } u \in S \ .$$

We will show this bound by induction on T.

The statement is trivial for $T = 0$. For $T > 0$, we may assume by induction hypothesis that it already holds for $T - 1$. So, in particular, we can set $\mathbf{u} = \mathbf{w}^{(T+1)}$ to get

$$\sum_{t=1}^{T-1} f_t(\mathbf{w}^{(T+1)}) \geq \sum_{t=1}^{T-1} f_t(\mathbf{w}^{(t+1)}) \ .$$

By adding $f_T(\mathbf{w}^{(T+1)})$ to both sides, we get

$$\sum_{t=1}^{T} f_t(\mathbf{w}^{(T+1)}) \geq \sum_{t=1}^{T} f_t(\mathbf{w}^{(t+1)}) \ .$$

Recall the definition of $\mathbf{w}^{(T+1)}$. It is chosen such that $\sum_{t=1}^{T} f_t(\mathbf{w}^{(T+1)})$ is minimized, which means nothing but

$$\sum_{t=1}^{T} f_t(\mathbf{w}^{(T+1)}) \leq \sum_{t=1}^{T} f_t(\mathbf{u}) \qquad \text{for all } u \in S \ .$$

In combination, these two bounds show the claim for T.

□

Example 19.4. *Let $S = [-1, 1]$. We choose the functions as follows*

$$f_1(w) = \frac{w}{2} \qquad f_{2k}(w) = -w \qquad f_{2k+1}(w) = w \qquad \text{for all } k \in \mathbb{N} \ .$$

In odd steps $t \geq 3$, $\sum_{t'=1}^{t-1} f_t(w) = -\frac{w}{2}$; in even steps t, $\sum_{t'=1}^{t-1} f_t(w) = \frac{w}{2}$. Therefore, Follow-the-leader chooses w_1 arbitrarily, $w_2 = -1$, $w_3 = 1$, $w_4 = -1$, ... Therefore $f_t(w^{(t)}) = 1$ for all $t > 1$. Choosing, in contrast, $u = 0$, then for all t we get $f_t(u) = 0$. So, $\text{Regret}^{(T)} \geq T - 1$.

4 Follow-the-Regularized-Leader

The problem in Example 19.4 is that the optimal point keeps jumping from one extreme to the other; Follow-the-Leader is always "too late". Therefore, we modify the algorithm a tiny bit. We add a *regularization term*: We choose $\mathbf{w}^{(t)}$ as the $\mathbf{v}$ that minimizes $R(\mathbf{v}) + \sum_{t'=1}^{t-1} f_{t'}(\mathbf{v})$. The function $R\colon S \to \mathbb{R}$ is a suitable function that has higher values for more "extreme" values.

Example 19.5. *Typical choice of regularizers are*

- *Euclidean regularization*

$$R(\mathbf{v}) = \frac{1}{2\eta} \sum_{i=1}^{d} v_i^2 \ ,$$

- *Entropical regularization (for non-negative vectors)*

$$R(\mathbf{v}) = \frac{1}{\eta} \sum_{i=1}^{d} v_i \ln v_i \ ,$$

where $\eta > 0$ is a scaling factor, determining how strong the regularization works. Smaller values of η mean stronger regularization. In the case of Euclidean regularization, points closer to the origin are preferred. Entropical regularization prefers values between 0 and 1 to the boundary points.

Recall that $S = \{\mathbf{v} \in \mathbb{R}^d \mid v_i \geq 0 \text{ for all } i, \sum_{i=1}^{d} v_i = 1\}$ with linear functions f_i is exactly the experts setting. One can show that Entropical regularization makes us choose $\mathbf{w}^{(t)}$ exactly such that $w_i^{(t)}$ is proportional to $\exp(-\eta \sum_{t'=1}^{t-1} \ell_i^{(t')})$. This is exactly the multiplicative-weights update rule.

Example 19.6. *Let us see what happens in Example 19.4 with Euclidean regularization. In odd steps $t \geq 3$, $\sum_{t'=1}^{t-1} f_t(w) + R(w) = -\frac{w}{2} + \frac{1}{2\eta} w^2$; in even steps t, $\sum_{t'=1}^{t-1} f_t(w) + R(w) = \frac{w}{2} + \frac{1}{2\eta} w^2$. These are minimized by $w^{(t)} = \frac{\eta}{2}$ for odd t and $w^{(t)} = -\frac{\eta}{2}$ for even t. So, if η is small enough, we indeed keep close to the origin.*

We can extend the regret bound to Follow-the-Regularized-Leader.

Lemma 19.7. *For Follow-the-Regularized-Leader, we have*

$$\text{Regret}^{(T)} \leq \max_{\mathbf{u} \in S} R(\mathbf{u}) - R(\mathbf{w}^{(1)}) + \sum_{t=1}^{T} (f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^{(t+1)})) \ .$$

Proof. Follow-the-Regularized-Leader is nothing but Follow-the-Leader with a hypothetical "step zero", in which $f_0 = R$. So, Lemma 19.3 tells us that for all $\mathbf{u} \in S$

$$\sum_{t=0}^{T} (f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{u})) \leq \sum_{t=0}^{T} (f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^{(t+1)})) \ ,$$

which now means

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

Because this bound holds for all $\mathbf{u} \in S$, rearranging gives us

$$\text{Regret}^{(T)} = \max_{\mathbf{u} \in S} \sum_{t=1}^{T} (f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{u})) \leq \max_{\mathbf{u} \in S} R(\mathbf{u}) - R(\mathbf{w}^{(1)}) + \sum_{t=1}^{T} (f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^{(t+1)})) \ . \ \square$$

At first sight, this regret bound might look weaker than the one for Follow-the-Leader. The point is, however, that the regularization keeps the difference of $f_t(\mathbf{w}^{(t)})$ and $f_t(\mathbf{w}^{(t+1)})$ smaller if it is chosen in a suitable way.