Algorithms and Uncertainty, Summer 2021 — Lecture 24 (5 pages)

Basics of Online Convex Optimization, Part 2

Thomas Kesselheim — Last Update: July 8, 2021

Recall the basic setting of Online Convex Optimization. We are optimizing over a convex set $S \subseteq \mathbb{R}^d$. There is an initially unknown sequence of cost functions $f_1, \ldots, f_T \colon S \to \mathbb{R}$. In step t, our algorithm chooses a point $\mathbf{w}^{(t)} \in S$ and only then gets to know f_t and incurs cost $f_t(\mathbf{w}^{(t)})$.

We assume that each function f_t is convex, that is

$$f_t(\mathbf{u}) \geq f_t(\mathbf{v}) + \langle \nabla f_t(\mathbf{v}), (\mathbf{u} - \mathbf{v}) \rangle \qquad \text{for all } \mathbf{u}, \mathbf{v} \in S \enspace .$$

Today, we will analyze the algorithm *Follow-the-Regularized-Leader*: Choose $\mathbf{w}^{(t)}$ so that

$$R(\mathbf{w}^{(t)}) + \sum_{t'=1}^{t-1} f_{t'}(\mathbf{w}^{(t)})$$

is minimized. We would like to bound its regret

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

In order to derive bounds, we require the cost functions $f_1, \ldots, f_T$ as well as the regularizer R to fulfill several properties beyond convexity.

1 Norms and Lipschitz Conditions

Our first assumption is that the cost functions $f_1, \ldots, f_T$ have to be bounded. More precisely, their rate of change has to be bounded. Note that in the experts setting, we also required $0 \leq \ell_i^{(t)} \leq \rho$ for all i and t; we now generalize this assumption.

We assume that there is a norm $\|\cdot\|$ defined on the set S. This means nothing but that every point in S has a "length". Based on this norm, we assume that each function f_t fulfills a Lipschitz condition. We require that for all $\mathbf{u}, \mathbf{v} \in S$

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

Example 24.1. *Standard examples of norms are the ℓ_1, ℓ_2, and ℓ_∞ norm. Generally, the ℓ_p norm is defined by $\|\mathbf{v}\|_p = \sqrt[p]{\sum_{i=1}^d |v_i|^p}$.*

We captured the experts setting by $S = \{\mathbf{v} \mid v_i \geq 0 \text{ for all } i, \sum_{i=1}^d v_i = 1\}$ and each function $f_t(\mathbf{v}) = \sum_{i=1}^d \ell_i^{(t)} v_i$. Consider now the case that $\ell_i^{(t)} \in [0, \rho]$ for all i and t.

With the ℓ_1 norm, we have

$$\|\mathbf{u} - \mathbf{v}\|_1 = \sum_{i=1}^{d} |u_i - v_i| \enspace ,$$

and therefore

$$f_t(\mathbf{u}) - f_t(\mathbf{v}) = \sum_{i=1}^{d} \ell_i^{(t)} (u_i - v_i) \leq \rho \|\mathbf{u} - \mathbf{v}\|_1 \enspace .$$

So, if we are considering the ℓ_1 norm, then we could choose $L = \rho$.

To get a bound for the ℓ_2 norm, we can use that $\|\mathbf{v}\|_1 \leq \sqrt{d} \|\mathbf{v}\|_2$ for all $\mathbf{v} \in \mathbb{R}^d$. So, in this case, we could choose $L = \sqrt{d}\rho$.

2 Strongly Convex Regularizers

The second condition concerns the regularizing function R. We assume it to be strongly convex.

Definition 24.2. *Let $\sigma \geq 0$. A differentiable function F is σ-strongly convex if for all $\mathbf{u}, \mathbf{v}$*

$$F(\mathbf{u}) \geq F(\mathbf{v}) + \langle \nabla F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \frac{\sigma}{2} \|\mathbf{u} - \mathbf{v}\|^2$$

So, strong convexity requires the function F to not only stay above its tangent but also move away from it (see Figure 1 for an illustration).

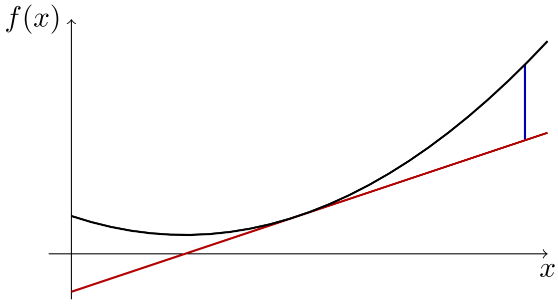

Figure 1: A strongly convex function moves away from the tangent (tangent drawn in red, distance drawn in blue).

Example 24.3. *The function R with $R(\mathbf{v}) = \frac{1}{2\eta} \sum_{i=1}^d v_i^2$ is $\frac{1}{\eta}$-strongly convex with respect to the ℓ_2-norm.*

We have to determine the gradient ∇R, which is the vector of all partial derivatives. We get $\frac{\partial R}{\partial v_i} = \frac{1}{\eta} v_i$ and so

$$\langle \nabla R(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle = \sum_{i=1}^d \left(\frac{\partial R}{\partial v_i}(\mathbf{v}) \right) (u_i - v_i) = \frac{1}{\eta} \sum_{i=1}^d v_i (u_i - v_i)$$

Overall, this gives us for all $\mathbf{u}$ and $\mathbf{v}$

$$\begin{aligned} \langle \nabla R(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \frac{1}{2\eta} \|\mathbf{u} - \mathbf{v}\|_2^2 &= \frac{1}{\eta} \sum_{i=1}^d v_i (u_i - v_i) + \frac{1}{2\eta} \sum_{i=1}^d (u_i - v_i)^2 \\ &= \frac{1}{2\eta} \sum_{i=1}^d u_i^2 - \frac{1}{2\eta} \sum_{i=1}^d v_i^2 = R(\mathbf{u}) - R(\mathbf{v}) \ . \end{aligned}$$

This matches exactly the definition of $\frac{1}{\eta}$-strongly convex.

There is one important observation: If we add a σ-strongly convex functions to a sum of convex functions, then the overall sum is again σ-strongly convex.

Observation 24.4. *If R is σ-strongly convex and $f_1, f_2, \ldots$ are convex then $R + \sum_t f_t$ is σ-strongly convex.*

So, if the regularizer is σ-strongly convex, Follow-the-Regularized-Leader minimizes a σ-strongly convex function in every step.

3 Analysis of Follow-the-Regularized-Leader

Having introduced the Lipschitz condition and strong convexity, we are prepared to state the regret guarantee from Follow-the-Regularized-Leader.

Theorem 24.5. *If the regularizer R is σ-strongly convex and each f_t fulfills the Lipschitz condition with parameter L, the regret of Follow-the-Regularized-Leader is bounded by*

$$\text{Regret}^{(T)} \leq \max_{\mathbf{u} \in S} R(\mathbf{u}) - R(\mathbf{w}^{(1)}) + T\frac{L^2}{\sigma} \ .$$

Before we proceed to the proof of this theorem, let us first derive a bound for the experts setting with Euclidean regularization.

Example 24.6. *If we use Euclidean regularization $R(\mathbf{v}) = \frac{1}{2\eta}\sum_{i=1}^{d} v_i^2$ in the experts setting, then $\max_{\mathbf{u}\in S} R(\mathbf{u}) = \frac{1}{2\eta}$, $R(\mathbf{w}^{(1)}) \geq 0$. Furthermore, using the ℓ_2-norm, we have $L = \sqrt{d}\rho$ and $\sigma = \frac{1}{\eta}$. This gives us a regret bound of*

$$\frac{1}{2\eta} + T(d\rho^2)\eta \ .$$

So, setting $\eta = \frac{1}{\sqrt{2dT}\rho}$, this guarantee becomes $\sqrt{2dT}\rho$. So, this is another no-regret algorithm for the experts setting.

To prove Theorem 24.5, we can use the bound that we derived last lecture.

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

Observe that, with this bound, it is enough to show the following lemma.

Lemma 24.7. *If the regularizer is σ-strongly convex and f_t fulfills the Lipschitz condition with parameter L, then $f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^{(t+1)}) \leq \frac{L^2}{\sigma}$ for all t.*

To prove this lemma, we will need another one, which is the core insight of why we want our functions to be strongly convex: If so, every point that is far from the minimum also has a much higher function value.

Lemma 24.8. *Let $F \colon S \to \mathbb{R}$ be a σ-strongly convex differentiable function over S with respect to a norm $\|\cdot\|$. Let $\mathbf{w} \in \arg\min_{\mathbf{v}\in S} F(\mathbf{v})$. Then, for all $\mathbf{u} \in S$*

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

Proof. Consider any $\mathbf{u} \in S$. Note that $\langle \nabla F(\mathbf{w}), \mathbf{u} - \mathbf{w}\rangle$ is the directional derivative of F and $\mathbf{w}$ in the direction from $\mathbf{w}$ to $\mathbf{u}$. It indicates by how much F changes when we move from $\mathbf{w}$ a bit towards $\mathbf{u}$ and is given by $\lim_{\epsilon\to 0} \frac{F(\mathbf{w}+\epsilon(\mathbf{u}-\mathbf{w}))-F(\mathbf{w})}{\epsilon}$.

We assume that $\mathbf{w}$ is a (global) minimum, so $F(\mathbf{w} + \epsilon(\mathbf{u} - \mathbf{w})) \geq F(\mathbf{w})$ for all $\epsilon \in [0, 1]$. (Note that $\mathbf{w} + \epsilon(\mathbf{u} - \mathbf{w})) \in S$ for $\epsilon \in [0, 1]$ because of convexity of S.) So, this also means $\langle \nabla F(\mathbf{w}), \mathbf{u} - \mathbf{w}\rangle \geq 0$.

So, by strong convexity, we have

$$F(\mathbf{u}) - F(\mathbf{w}) \geq \langle \nabla F(\mathbf{w}), \mathbf{u} - \mathbf{w}\rangle + \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2 \geq \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2 \ .$$ □

Now, we can finally prove Lemma 24.7 and this way complete our analysis of Follow-the-Regularized-Leader.

Proof of Lemma 24.7. For all t, let $F_t(\mathbf{v}) = R(\mathbf{v}) + \sum_{t'=1}^{t-1} f_{t'}(\mathbf{v})$. Note that by Observation 24.4, F_t is σ-strongly convex. Furthermore, $\mathbf{w}^{(t)}$ is exactly a vector that minimizes F_t. Therefore, by Lemma 24.8, we have

$$F_t(\mathbf{w}^{(t+1)}) - F_t(\mathbf{w}^{(t)}) \geq \frac{\sigma}{2} \|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|^2 \ .$$

We can apply the same argument on F_{t+1}, which is minimized by $\mathbf{w}^{(t+1)}$, to get

$$F_{t+1}(\mathbf{w}^{(t)}) - F_{t+1}(\mathbf{w}^{(t+1)}) \geq \frac{\sigma}{2} \|\mathbf{w}^{(t)} - \mathbf{w}^{(t+1)}\|^2 \ .$$

By taking the sum of these two inequalities, we get

$$\Big(F_{t+1}(\mathbf{w}^{(t)}) - F_t(\mathbf{w}^{(t)})\Big) - \Big(F_{t+1}(\mathbf{w}^{(t+1)}) - F_t(\mathbf{w}^{(t+1)})\Big) \geq \sigma \|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|^2 \ .$$

So, by the definition of F_t and F_{t+1},

$$f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^{(t+1)}) \geq \sigma \|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|^2 \ .$$

The Lipschitz condition of f_t gives us

$$f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^{(t+1)}) \leq L \|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\| \ ,$$

so in combination

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\| \leq \frac{L}{\sigma} \ ,$$

and therefore

$$f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^{(t+1)}) \leq \frac{L^2}{\sigma} \ .$$

□

4 Bound for Entropical Regularization

One of our frequent examples was Entropical regularization $R(\mathbf{w}) = \frac{1}{\eta} \sum_{i=1}^d w_i \ln w_i$ on $S = \{\mathbf{v} \in \mathbb{R}^d \mid v_i \geq 0 \text{ for all } i, \sum_{i=1}^d v_i = 1\}$. If the functions f_t are linear, this corresponds exactly to the multiplicative-weights algorithm in the experts setting.

One can show that R is $\frac{1}{\eta}$-strongly convex with respect to the ℓ_1 norm. In the experts setting, we would have $L = \rho$. As $-\frac{\ln d}{\eta} \leq R(\mathbf{w}) \leq 0$, Theorem 24.5 gives us a bound of $\frac{\ln d}{\eta} + T\eta\rho^2$. Setting $\eta = \frac{1}{\rho}\sqrt{\frac{\ln d}{T}}$, this is $2\rho\sqrt{T \ln d}$, which is exactly the guarantee that we derived before.

5 Equivalent Formulations of Follow-the-Regularized-Leader

A potential downside of Follow-the-Regularized-Leader is that it seemingly requires to do a complicated optimization task in every round, namely to find the best point so far. Quite surprisingly, this is not as difficult as it sounds. As mentioned before, the case of Entropical regularization in the experts setting corresponds to the multiplicative-weights algorithm, which easily derives $\mathbf{w}^{(t+1)}$ from $\mathbf{w}^{(t)}$. This holds in a much more general sense.

5.1 Euclidean Regularization with Linear Functions

Consider $S = \mathbb{R}^d$ and linear functions f_t. That is, $f_t(\mathbf{v}) = \sum_{i=1}^d \ell_i^{(t)} v_i$. Then Euclidean regularization tells us to choose $\mathbf{w}^{(t)}$ so as to minimize

$$\sum_{t'=1}^{t-1} \sum_{i=1}^{d} \ell_i^{(t')} w_i + \frac{1}{2\eta} \sum_{i=1}^{d} (w_i)^2 \ .$$

The partial derivative by w_i is

$$\sum_{t'=1}^{t-1} \ell_i^{(t')} + \frac{1}{\eta} w_i \ .$$

In order for $\mathbf{w}^{(t)}$ to minimize the function, $w_i^{(t)}$ has to be a zero of this partial derivative. Therefore, $w_i^{(t)} = -\eta \sum_{t'=1}^{t-1} \ell_i^{(t')}$ for all i, so $\mathbf{w}^{(t)} = -\eta \sum_{t'=1}^{t-1} \ell^{(t')}$. Written recursively, this is $\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \eta \ell^{(t-1)}$. That is, from $\mathbf{w}^{(t-1)}$ to $\mathbf{w}^{(t)}$, we move by $\eta \ell^{(t-1)}$. As $\ell^{(t-1)}$ is the gradient of f_{t-1} at any point, this algorithm is also called *online gradient descent*.

5.2 Online Mirror Descent

With any regularizers, Follow-the-Regularized-Leader is equivalent to an algorithm called *online mirror descent*, which works as follows. It uses vectors $\theta^{(1)}, \ldots, \theta^{(T)} \in \mathbb{R}^d$. Initially $\theta^{(1)} = 0$. In step t,

- Choose new point $\mathbf{w}^{(t)} = g(\theta^{(t)})$ by “mirroring”
- Update vector θ by $\theta^{(t+1)} = \theta^{(t)} - \nabla f_t(\mathbf{w}^{(t)})$

When setting $g(\theta) = -\eta\theta$, this is exactly online gradient descent.