6.883: Online Methods in Machine Learning

Alexander Rakhlin

LECTURE 21 AND 22

1. ONLINE LINEAR OPTIMIZATION: DETERMINISTIC METHODS

For the experts problem (and its close relative B_1/B_∞ linear game) we have developed a deterministic method (Exponential Weights) and a randomized method (Follow the Perturbed Leader). Both the deterministic and randomized methods extend to other online prediction problems where the loss is linear in the decision of the learner and linear in the outcome. We have seen that the geometry of the two sets (decisions and outcomes) plays a crucial role, both for computational tractability and for attainable regret guarantees. (Recall the online shortest path problem with its flow polytope, the online ranking problem with the Birkhoff polytope or the semidefinite representation)

In this lecture, we present techniques for deriving deterministic methods that take advantage of the geometry of the decision/outcome sets. We illustrate several closely related techniques on the B_2/B_2 analysis, since the solution is very clean and immediately suggests generalization to other norms.

1.1 B_2/B_2 online linear optimization

To restate the problem, let B_2 be the unit Euclidean ball. The protocol is

For $t = 1, \dots, n$
  Predict $\widehat{y}_t \in B_2$
  Observe costs $z_t \in B_2$

with regret defined as the difference

$$\sum_{t=1}^{n} \langle \widehat{y}_t, z_t \rangle - \min_{v \in B_2} \sum_{t=1}^{n} \langle v, z_t \rangle . \tag{1}$$

Recall that

$$-\min_{v \in B_2} \sum_{t=1}^{n} \langle v, z_t \rangle = \max_{v \in B_2} \sum_{t=1}^{n} \langle v, -z_t \rangle = \|L_n\| \tag{2}$$

with $L_n = \sum_{t=1}^{n} z_t$.

1.2 How to relax

A relaxation at step n is an upper bound on the benchmark term. The randomized methods developed in the previous lectures dealt directly with (2). Today, we will get further upper

bounds. The first inequality that comes to mind in relaxing (2) is triangle inequality $\|L_n\| \le \|L_{n-1}\| + 1$. However, this relaxation is too loose (why?). A tighter relaxation is

$$\|L_n\| = \sqrt{\|L_n\|^2} = \sqrt{\|L_{n-1}\|^2 + 2\langle L_{n-1}, z_n\rangle + \|z_n\|} \le \sqrt{\|L_{n-1}\|^2 + 2\langle L_{n-1}, z_n\rangle + 1} \tag{3}$$

or we may equivalently write the above as

$$\inf_{\eta} \left\{ \frac{1}{2\eta} + \frac{\eta}{2}\left(\|L_{n-1}\|^2 + 2\langle L_{n-1}, z_n\rangle + 1\right)\right\}. \tag{4}$$

The second relaxation follows from a useful and trivial inequality that for $a, b > 0$,

$$2\sqrt{a \cdot b} = \min_{\eta} \left\{ a\eta + b\eta^{-1} \right\}. \tag{5}$$

The two relaxations (3) and (4) are identical, but the second one is more amenable to extensions (in particular, the convex conjugate form defined later in this lecture).

1.3 B_2/B_2 analysis based on (3)

We take

$$\mathbf{Rel}(z_{1:n}) = \sqrt{\|L_{n-1}\|^2 + 2\langle L_{n-1}, z_n\rangle + 1}. \tag{6}$$

for $L_s = \sum_{t=1}^{s} z_t$. The minimax problem for the last step is

$$\min_{\widehat{y}_n \in B_2} \max_{z_n \in B_2} \left\{ \langle \widehat{y}_n, z_n\rangle + \sqrt{\|L_{n-1}\|^2 + 2\langle L_{n-1}, z_n\rangle + 1}\right\} \tag{7}$$

One can actually solve this expression by considering the direction L_{n-1} and any other orthogonal direction. To save time, let us just state the strategy:

$$\widehat{y}_n = -\frac{L_{n-1}}{\sqrt{\|L_{n-1}\|^2 + 1}}. \tag{8}$$

Observe that

$$-\frac{\alpha}{\sqrt{A}} + \sqrt{A + 2\alpha}$$

is maximized at $\alpha = 0$. Hence, the optimal response z_n is orthogonal to L_{n-1}. This gives an upper bound of

$$\sqrt{\|L_{n-1}\|^2 + 1} \tag{9}$$

on the minimax value (7). Further upper bounding gives

$$\sqrt{\|L_{n-1}\|^2 + 1} \le \sqrt{\|L_{n-2}\|^2 + 2\langle L_{n-2}, z_{n-1}\rangle + 2} \triangleq \mathbf{Rel}(z_{1:n-1}). \tag{10}$$

We established

$$\mathbf{Rel}(z_{1:t}) = \sqrt{\|L_{t-1}\|^2 + 2\langle L_{t-1}, z_t\rangle + (n-t+1)}$$

and an admissible strategy for this relaxation is

$$\widehat{y}_t = -\frac{L_{t-1}}{\sqrt{\|L_{t-1}\|^2 + (n-t+1)}}. \tag{11}$$

Noting that $L_t = \sum_{s=1}^{t} z_s$, the solution is an interesting form of "gradient descent" with adaptive step size. Finally, the regret bound is

$$\mathbf{Rel}(\varnothing) = \sqrt{n}.$$

Why did we mention "gradient descent"? The connection, as explained in class, is simple. An algorithm for online linear optimization is also an algorithm for convex optimization, where z_t's are the gradients of the function we are optimizing.

1.4 B_2/B_2 analysis based on (4)

$$\mathbf{Rel}(z_{1:n}) = \inf_{\eta} \left\{ \frac{1}{2\eta} + \frac{\eta}{2} \left(\|L_{n-1}\|^2 + 2\langle L_{n-1}, z_n \rangle + 1 \right) \right\}. \tag{12}$$

The minimax problem for the last step is

$$\min_{\widehat{y}_n \in B_2} \max_{z_n \in B_2} \left\{ \langle \widehat{y}_n, z_n \rangle + \inf_{\eta} \left\{ \frac{\eta}{2} \left(\|L_{n-1}\|^2 + 2\langle L_{n-1}, z_n \rangle + 1 \right) + \frac{1}{2\eta} \right\} \right\} \tag{13}$$

which is upper bounded by

$$\inf_{\eta} \min_{\widehat{y}_n \in B_2} \max_{z_n \in B_2} \left\{ \langle \widehat{y}_n, z_n \rangle + \frac{\eta}{2} \left(\|L_{n-1}\|^2 + 2\langle L_{n-1}, z_n \rangle + 1 \right) + \frac{1}{2\eta} \right\} \tag{14}$$

which simplifies to (can you show this?)

$$\inf_{\eta} \min_{\widehat{y}_n \in B_2} \left\{ \|\widehat{y}_n + \eta L_{n-1}\| + \frac{\eta}{2} \left(\|L_{n-1}\|^2 + 1 \right) + \frac{1}{2\eta} \right\} \tag{15}$$

One may check (by taking derivatives) that the optimal choice of η occurs at a value that ensures $\eta L_{n-1} \in B_2$ and the optimal choice for $\widehat{y}_n$ is

$$\widehat{y}_n = -\eta L_{n-1}.$$

Continuing in this fashion,

$$\mathbf{Rel}(z_{1:t}) = \inf_{\eta} \left\{ \frac{1}{2\eta} + \frac{\eta}{2} \left(\|L_{t-1}\|^2 + 2\langle L_{t-1}, z_t \rangle + (n-t+1) \right) \right\}, \tag{16}$$

with

$$\mathbf{Rel}(\varnothing) = \inf_{\eta} \left\{ \frac{1}{2\eta} + \frac{\eta n}{2} \right\} = \sqrt{n}$$

if we choose $\eta = 1/\sqrt{n}$.

2. RELAXATIONS BASED ON CONVEX CONJUGACY

For a real-valued function $\phi : \mathcal{A} \to \mathbb{R} \cup \{\infty\}$, the **convex conjugate** $\phi^* : \mathcal{B} \to \mathbb{R} \cup \{\infty\}$ is defined by

$$\phi^*(b) = \sup_{a \in \mathcal{A}} \langle b, a \rangle - \phi(a). \tag{17}$$

Technically, $\mathcal{A}$ and $\mathcal{B}$ are dual spaces, but let's not worry about these details. We will take ϕ convex. Oftentimes, the supremum in the above definition is taken to be unconstrained.

As in the beginning of the course, we say that ϕ is σ-strongly convex with respect to a norm $\|\cdot\|$ if

$$\phi(a) \geq \phi(b) + \langle \nabla\phi(b), a - b\rangle + \frac{\sigma}{2}\|a - b\|^2 \tag{18}$$

Then ϕ^* is smooth with respect to the dual norm:

$$\phi^*(a) \leq \phi^*(b) + \langle \nabla\phi^*(b), a - b\rangle + \frac{1}{2\sigma}\|a - b\|_*^2 \tag{19}$$

Examples are in order (see [SS07] for more)

- The function $\phi(a) = \frac{1}{2}\|a\|^2$, half Euclidean norm squared, is conjugate to itself: $\phi^*(b) = \frac{1}{2}\|b\|^2$. This function is strongly convex with respect to Euclidean norm.
- The entropy function

$$\phi(a) = \sum_{i=1}^{N} a_i \log a_i + \log N \tag{20}$$

  over the probability simplex in N dimensions is strongly convex with respect to the ℓ_1 norm $\|\cdot\|_1$. The conjugate is

$$\phi^*(b) = \log\left(\frac{1}{N}\sum_{i=1}^{N} e^{b_i}\right) \tag{21}$$

Recall that the starting relaxation is defined as an upper bound on

$$-\inf_{v\in K}\langle v, L_n\rangle = \sup_{v\in K}\langle v, -L_n\rangle . \tag{22}$$

Convex conjugacy gives a principled way of choosing a good relaxation for the given set K. We aim to find a function ϕ that is strongly convex over the set K. To gain intuition, suppose $\phi(0) = 0$, the minimum of ϕ, and K is a ball with respect to a norm. If ϕ is 2-strongly convex with respect to this norm, it means that the value of function is at least 1 at the boundary of K. If we ensure that ϕ is not too large over K, then ϕ "respects" the geometry of the set.

Write, for any $\eta > 0$,

$$\sup_{v\in K}\langle v, -L_n\rangle = \sup_{v\in K}\left\{\langle v, -L_n\rangle - \frac{1}{\eta}\phi(v) + \frac{1}{\eta}\phi(v)\right\} \tag{23}$$

which is upper bounded, for any $\eta > 0$, by

$$\frac{1}{\eta}\sup_{v\in K}\{\langle v, -\eta L_n\rangle - \phi(v)\} + \frac{1}{\eta}\sup_{v\in K}\phi(v) = \frac{1}{\eta}\phi^*\left(-\eta L_n\right) + \frac{1}{\eta}R \tag{24}$$

for $R = \sup_{v\in K}\phi(v)$. This step can be seen as relaxing the indicator $I_K(v)$ constraint for $v \in K$ by a convex function ϕ that "mimics" the shape of K.

By smoothness, we have

$$\phi^*\left(-\eta L_n\right) \leq \phi^*\left(-\eta L_{n-1}\right) + \left\langle \nabla\phi^*\left(-\eta L_{n-1}\right), z_n\right\rangle + \frac{1}{2\sigma}\left\|z_n\right\|_*^2 \tag{25}$$

and so

$$\sup_{v\in K}\langle v, -L_n\rangle \leq \inf_{\eta}\left\{\frac{1}{\eta}\left(\phi^*\left(-\eta L_{n-1}\right) + \left\langle \nabla\phi^*\left(-\eta L_{n-1}\right), z_n\right\rangle + \frac{1}{2\sigma}G^2\right) + \frac{1}{\eta}R\right\} \tag{26}$$

where $\left\|z_n\right\|_*^2 \leq G^2$ and $R = \sup_{v\in K}\phi(v)$.

This last upper bound can be taken as a relaxation, and it gives the Dual Averaging solution. Relaxation in (26) should be compared to (12). The latter is a particular case of the former when $\phi(a) = \frac{1}{2}\left\|a\right\|^2$. For a general set K, the method is also known as Follow the Regularized Leader:

$$\widehat{y}_t = \operatorname*{argmin}_{v\in K}\ \langle v, \eta L_{t-1}\rangle + \phi(v) \tag{27}$$

If $\tilde{y}_t$ is the unconstrained minimum in (27), we see that

$$\nabla\phi(\tilde{y}_t) = -\eta L_{t-1}. \tag{28}$$

Under some conditions (the conjugate needs to be unconstrained), the functional inverse of $\nabla\phi(\cdot)$ is $\nabla\phi^*$, and so

$$\tilde{y}_t = \nabla\phi^*(-\eta L_{t-1}), \tag{29}$$

and $\widehat{y}_t$ is just a projection of $\tilde{y}_t$ onto the set K. This projection needs to be done with respect to the geometry induced by ϕ. More precisely, one defines a Bregman divergence

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

the difference between the value of ϕ at u and its first-order approximation. The projection with respect to ϕ is then

$$\widehat{y}_t = \operatorname*{argmin}_{y\in K}\ D_\phi(y, \tilde{y}_t). \tag{30}$$

It can be shown that (30) with (29) is equivalent to (27). Sometimes, this algorithm is called lazy projection because it keeps a running sum of L_t and projects onto the set K at every step to calculate the prediction $\widehat{y}_t$. A closely related method of Mirror Descent (discussed below) keeps track of the previous projection, rather than L_{t-1}, as a sufficient statistic.

2.1 Example: Exponential Weights

Take the entropy function (20) ϕ and its dual ϕ^*. The gradient

$$\left[\nabla\phi(a)\right]_i = \log(a_i) + 1$$

and the gradient of the dual is

$$\left[\nabla\phi^*(b)\right]_i = \frac{e^{b_i}}{\sum_i e^{b_i}}$$

which can be immediately recognized as the Exponential Weights algorithm.

3. MIRROR DESCENT

As mentioned above, Mirror Descent (MD) keeps track of $\widehat{g}_t$ rather than L_t as the sufficient statistic. In other words, relaxation $\mathbf{Rel}(z_{1:t})$ can be written as $\mathbf{Rel}(g_t(z_{1:t-1}), z_t)$.

The MD algorithm is

$$g_{t+1} = \operatorname*{argmin}_{v \in K} \ \langle v, z_t \rangle + \eta^{-1} D_\phi(v, g_t) \tag{31}$$

for some η that will be calculated later. Let $\tilde{g}_{t+1}$ be the unrestricted minimum in (31). By taking derivatives of the objective and setting to zero (use the definition of Bregman divergence), we see that

$$-\eta z_t = \nabla\phi(\tilde{g}_{t+1}) - \nabla\phi(g_t).$$

That is, Mirror Descent can be viewed as: map g_t via $\nabla\phi$ to the dual space, then make a gradient step, then map back via the inverse map $\nabla\phi^*$, and then project onto K. The difference with respect to FTRL is the interleaved projections, but that's about it.

Let's see if we can recover the algorithm through relaxations. We set

$$\max_{v \in K} \langle v, -L_n \rangle \le \max_{v \in K} \left\{ \langle v, -L_n \rangle + \eta^{-1} D_\phi(v, g_n) \right\} \triangleq \mathbf{Rel}(g_n, z_n) \tag{32}$$

The last step is

$$\min_{\widehat{g}_n} \max_{z_n} \left\{ \langle \widehat{g}_n, z_n \rangle + \max_{v \in K} \left\{ \langle v, -L_n \rangle + \eta^{-1} D_\phi(v, g_n) \right\} \right\} \tag{33}$$

$$= \min_{\widehat{g}_n} \max_{z_n} \max_{v \in K} \left\{ \langle v, -L_{n-1} \rangle + \langle \widehat{g}_n - v, z_n \rangle + \eta^{-1} D_\phi(v, g_n) \right\} \tag{34}$$

The optimality condition for (31) says that

$$\langle g_n - v, \ -\eta z_n - \nabla\phi(g_n) + \nabla\phi(g_{n-1}) \rangle \ge 0 \tag{35}$$

(that is, negative gradient direction has positive inner product with $g_n - v$ for any $v \in K$). Rearranging,

$$\langle g_n - v, \ \eta z_n \rangle \le \langle v - g_n, \ \nabla\phi(g_{n-1}) - \nabla\phi(g_n) \rangle . \tag{36}$$

We now use an elementary property of Bregman divergences:

$$D_\phi(a, b) + D_\phi(b, c) = D_\phi(a, c) + \langle a - b, \nabla\phi(c) - \nabla\phi(b) \rangle \tag{37}$$

Choosing $v = a$, $g_n = b$, $g_{n-1} = c$, and dividing by η, (36) becomes

$$\langle g_n - v, \ z_n \rangle \le \eta^{-1} D_\phi(v, g_{n-1}) - \eta^{-1} D_\phi(v, g_n) + \eta^{-1} D_\phi(g_n, g_{n-1}). \tag{38}$$

Finally (this requires some work),

$$\eta^{-1} D_\phi(g_n, g_{n-1}) = \eta^{-1} D_{\phi^*}(\nabla\phi(g_n), \nabla\phi(g_{n-1})) \le \eta^{-1} \frac{1}{2\sigma} \|\eta z_n\|_*^2 \le \frac{\eta}{2\sigma} G^2$$

where we used the fact that ϕ is strongly convex while ϕ^* is smooth with respect to the dual norm.

Putting everything together, (34) is equal to

$$\min_{\widehat{y}_n} \max_{z_n} \max_{v \in K} \left\{ \langle v, -L_{n-1} \rangle + \langle \widehat{y}_n - g_n, z_n \rangle + \langle g_n - v, z_n \rangle + \eta^{-1} D_\phi(v, g_n) \right\} \tag{39}$$

which is upper bounded, in view of the above calculations, by

$$\min_{\widehat{y}_n} \max_{z_n} \max_{v \in K} \left\{ \langle v, -L_{n-1} \rangle + \langle \widehat{y}_n - g_n, z_n \rangle + \eta^{-1} D_\phi(v, g_{n-1}) \right\} + \frac{\eta}{2\sigma} G^2 \tag{40}$$

Now the terms decouple and we get

$$\min_{\widehat{y}_n} \max_{z_n} \langle \widehat{y}_n - g_n, z_n \rangle + \max_{v \in K} \left\{ \langle v, -L_{n-1} \rangle + \eta^{-1} D_\phi(v, g_{n-1}) \right\} + \frac{\eta}{2\sigma} G^2 \tag{41}$$

The optimal strategy for $\widehat{y}_n$ is to set $\widehat{y}_n = g_n$. This proves the recursion for the last step. One can see that

$$\mathbf{Rel}(z_{1:t}) = \mathbf{Rel}(g_t(z_{1:t-1}), z_t) = \max_{v \in K} \left\{ \langle v, -L_t \rangle + \eta^{-1} D_\phi(v, g_t) \right\} + \frac{(n-t+1)\eta}{2\sigma} G^2.$$

In particular,

$$\mathbf{Rel}(\varnothing) = \eta^{-1} R^2 + \frac{G^2 \eta n}{2\sigma}$$

for $R^2 = \max_{v \in K} D_\phi(v, g_0)$. It remains to tune η to conclude that

Lemma 1. *Mirror Descent with a σ-strongly convex function ϕ (with respect to a norm $\|\cdot\|$ over K) with $\eta = (R/G)\sqrt{\frac{2\sigma}{n}}$, $R^2 = \max_{v \in K} D_\phi(v, g_0)$, guarantees that*

$$\sum_{t=1}^{n} \langle \widehat{y}_t, z_t \rangle - \min_{v \in K} \sum_{t=1}^{n} \langle v, z_t \rangle \le \sqrt{2} \frac{RG}{\sqrt{\sigma}} \sqrt{n} \tag{42}$$

for any sequence $z_1, \ldots, z_n$.

4. ADDING EXTRA KNOWLEDGE

It is sometimes argued that online algorithms are overly pessimistic. While the benchmark term does capture some of the prior knowledge of the practitioner, the result still protects against all sequences. Below, we provide one of the ways in which extra knowledge can be infused into the method. Imagine a linear optimization problem where on each step we first observe a guess M_t of the next z_t. This guess may come from external sources, or as an extra statistic believed to be important for this time step. How can we use this information, yet still be protective of the worst-case behavior? The algorithm below addresses this problem. We shall call it "Optimistic Mirror Descent" because it uses the extra information M_t, and the regret becomes tighter if the information is relevant, and matches the usual guarantee if the information is irrelevant. The method has been used already in a number of applications, including faster convergence in zero-sum games, coarse correlated equilibria in multiplayer normal form games, maximum flow problems, etc. The algorithm is a procedure that keeps track of two sequences:

$$g_{t+1} = \operatorname*{argmin}_{v \in K} \eta \langle v, z_t \rangle + D_\phi(v, g_t) \tag{43}$$

$$\widehat{y}_{t+1} = \operatorname*{argmin}_{v \in K} \eta \langle v, M_{t+1} \rangle + D_\phi(v, g_{t+1}) \tag{44}$$

Note that if $M_t = 0$, we recover Mirror Descent.

Lemma 2. *If K is convex set and ϕ is 1-strongly convex on K with respect to $\|\cdot\|$. Then Optimistic Mirror Descent yields, for any sequence $z_1, \ldots, z_n$,*

$$\sum_{t=1}^{n} \langle \widehat{y}_t, z_t \rangle - \min_{v \in K} \sum_{t=1}^{n} \langle v, z_t \rangle \leq cR \sqrt{\sum_{t=1}^{n} \|z_t - M_t\|_*^2 + 1} \tag{45}$$

where $R^2 = \max_{v \in K} \phi(v) - \min_{v \in K} \phi(v)$ and c is a constant.

The lemma above needs to choose η in a time-varying manner, and we refer to [RS12] for this.

The extra information can significantly decrease the regret bound by subtracting off the predictable part of the process, just as in Statistics we regress a variable and subtract a seasonal trend until only the unpredictable noise remains.

References

[RS12] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. *arXiv preprint arXiv:1208.3728*, 2012.

[SS07] Shai Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, 2007.