18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Jonathan Weed

Lecture 23
Dec. 2, 2015

1. POTENTIAL BASED APPROACHABILITY

Last lecture, we saw Blackwell's celebrated Approachability Theorem, which establishes a procedure by which a player can ensure that the average (vector) payoff in a repeated game approaches a convex set. The central idea was to construct a hyperplane separating the convex set from the point $\bar{\ell}_{t-1}$, the average loss so far. By projecting perpendicular to this hyperplane, we obtained a scalar-valued problem to which von Neumann's minimax theorem could be applied. The set S is approachable as long as we can always find a "silver bullet," a choice of action a_t for which the loss vector ℓ_t lies on the side of the hyperplane containing S. (See Figure 1.)

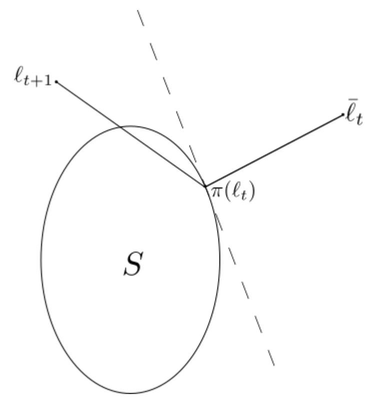

Figure 1: Blackwell approachability

Concretely, Blackwell's Theorem also implied the existence of a regret-minimizing algorithm for expert advice. Indeed, if we define the vector loss ℓ_t by $(\ell_t)_i = \ell(a_t, z_t) - \ell(e_i, z_t)$, then the average regret at time t is equivalent to the sup-norm distance between the average loss $\bar{\ell}_t$ and the negative orthant. Approaching the negative orthant therefore corresponds to achieving sublinear regret.

However, this reduction yielded suboptimal rates. To bound average regret, we replaced the sup-norm distance by the Euclidean distance, which led to an extra factor of $\sqrt{k}$ appearing in our bound. In the sequel, we develop a more sophisticated version of approachability that allows us to adapt to the geometry of our problem. (Much of what follows resembles out development of the mirror descent algorithm, though the two approaches differ in crucial details.)

1.1 Potential functions

We recall the setup of mirror descent, first described in Lecture 13. Mirror descent achieved accelerated rates by employing a potential function which was strongly convex with respect to the given norm. In this case, we seek what is in some sense the opposite: a function whose gradient does not change too quickly. In particular, we make the following definition.

Definition: A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a *potential* for $S \in \mathbb{R}$ if it satisfies the following properties:

- Φ is convex.
- $\Phi(x) \le 0$ for $x \in S$.
- $\Phi(y) = 0$ for $y \in \partial S$.
- $\Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle \le \frac{h}{2}\|x - y\|^2$, where by abuse of notation we use $\nabla\Phi(x)$ to denote a subgradient of Φ at x.

Given such a function, we recall two associated notions from the mirror descent algorithm. The Bregman divergence associated to Φ is given by

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle \,.$$

Likewise, the associated Bregman projection is

$$\pi(x) = \operatorname*{argmin}_{y \in S} D_\Phi(y, x)\,.$$

We aim to use the function Φ as a stand-in for the Euclidean distance that we employed in our proof of Blackwell's theorem. To that end, the following lemma establishes several properties that will allow us to generalize the notion of a separating hyperplane.

Lemma: For any convex, closed set S and $z \in S$, $x \in S^C$, the following properties hold.

- $\langle z - \pi(x), \nabla\Phi(x)\rangle \le 0$,
- $\langle x - \pi(x), \nabla\Phi(x)\rangle \ge \Phi(x)$.

In particular, if Φ is positive on S^C, then $H := \{y \mid \langle y - \Phi(x), \nabla\Phi(x)\rangle = 0\}$ is a separating hyperplane.

Our proof requires the following proposition, whose proof appears in our analysis of the mirror descent algorithm and is omitted here.

Proposition: For all $z \in S$, it holds

$$\langle \nabla\Phi(\pi(x)) - \nabla\Phi(x), \pi(x) - z\rangle \le 0\,.$$

Proof of Lemma. Denote by π the projection $\pi(x)$. The first claim follows upon expanding the expression on the left-hand side as follows

$$\langle z - \pi, \nabla\Phi(x)\rangle = \langle z - \pi, \nabla\Phi(x) - \nabla\Phi(\pi)\rangle + \langle z - \pi, \nabla\Phi(\pi)\rangle.$$

The above Proposition implies that the first term is nonpositive. Since the function Φ is convex, we obtain

$$0 \geq \Phi(z) \geq \Phi(\pi) + \langle z - \pi, \nabla\Phi(\pi)\rangle \,.$$

Since π lies on the boundary of S, by assumption $\Phi(\pi) = 0$ and the claim follows.

For the second claim, we again use convexity:

$$\Phi(\pi) \geq \Phi(x) + \langle \pi - x, \nabla\Phi(x)\rangle \,.$$

Since $\Phi(\pi) = 0$, the claim follows. □

1.2 Potential based approachability

With the definitions in place, the algorithm for approachability is essentially the same as it before we introduced the potential function. As before, we will use a projection defined by the hyperplane $H = \{y \mid \langle y - \pi(\bar\ell_{t-1}), \nabla\Phi(\bar\ell_{t-1}\rangle = 0\}$ and von Neumann's minmax theorem to find a "silver bullet" a_t^* such that $\ell_t = \ell(a_t^*, z_t)$ satisfies

$$\langle \ell_t - \pi_t, \nabla\Phi(\bar\ell_{t-1})\rangle \leq 0 \,.$$

All that remains to do is to analyze this procedure's performance. We have the following theorem.

Theorem: If $\|\ell(a,z)\| \leq R$ holds for all $z \in \mathcal{A}, z \in \mathcal{Z}$ and all assumptions above are satisfied, then

$$\Phi(\bar\ell_n) \leq \frac{4R^2 h \log n}{n} \,.$$

Proof. The definition of the potential Φ required that Φ be upper bounded by a quadratic function. The proof below is a simple application of that bound.

As before, we note the identity

$$\bar\ell_t = \bar\ell_{t-1} + \frac{\ell_t - \bar\ell_{t-1}}{t} \,.$$

This expression and the definition of Φ imply.

$$\Phi(\bar\ell_t) \leq \Phi(\bar\ell_{t-1}) + \frac{1}{t}\langle \ell_t - \bar\ell_{t-1}, \nabla\Phi(\bar\ell_{t-1})\rangle + \frac{h}{2t^2}\|\ell_t - \bar\ell_{t-1}\|^2 \,.$$

The last term is the easiest to control. By assumption, ℓ_t and $\bar\ell_{t-1}$ are contained in a ball of radius R, so $\|\ell_t - \bar\ell_{t-1}\|^2 \leq 4R^2$.

To bound the second term, write

$$\frac{1}{t}\langle \ell_t - \bar\ell_{t-1}, \nabla\Phi(\bar\ell_{t-1})\rangle = \frac{1}{t}\langle \ell_t - \pi_t, \nabla\Phi(\bar\ell_{t-1})\rangle + \frac{1}{t}\langle \pi_t - \bar\ell_{t-1}, \nabla\Phi(\bar\ell_{t-1})\rangle \,.$$

The first term is nonpositive by assumption, since this is how the algorithm constructs the silver bullet. By the above Lemma, the inner product in the second term is at most $-\Phi(\bar\ell_{t-1})$.

We obtain

$$\Phi(\bar\ell_t) \leq \left(\frac{t-1}{t}\right)\Phi(\bar\ell_{t-1}) + \frac{2hR^2}{t^2} \,.$$

Defining $u_t = t\Phi(\bar{\ell}_t)$ and rearranging, we obtain the recurrence

$$u_t \le u_{t-1} + \frac{2hR^2}{t}\,,$$

So

$$u_n = \sum_{t=1}^{n} u_t - u_{t-1} \le 2hR^2 \sum_{t=1}^{n} \frac{1}{t} \le 4hR^2 \log n\,.$$

Applying the definition of u_n proves the claim. □

1.3 Application to regret minimization

We now show that potential based approachability provides an improved bound on regret minimization. Our ultimate goal is to replace the bound $\sqrt{nk}$ (which we proved last lecture) by $\sqrt{n \log k}$ (which we know to be the optimal bound for prediction with expert advice). We will be able to achieve this goal up to logarithmic terms in n. (A more careful analysis of the potential defined below does actually yields an optimal rate.)

Recall that $\frac{R_n}{n} = d_\infty(\bar{\ell}_n, O_K^-)$, where R_n is the cumulative regret after n rounds and O_K^- is the negative orthant. It is not hard to see that $d_\infty = \|x_+\|_\infty$, where x_+ is the positive part of the vector x.

We define the following potential function:

$$\Phi(x) = \frac{1}{\eta} \log \left(\frac{1}{K} \sum_{j=1}^{K} e^{\eta (x_j)_+} \right)\,.$$

The function Φ is a kind of "soft max" of the positive entries of x. (Note that this definition does not agree with the use of the term soft max in the literature—the difference is the presence of the factor $\frac{1}{K}$.) The terminology soft max is justified by noting that

$$\|x_+\|_\infty = \max_j (x_j)_+ \le \max_j \frac{1}{\eta} \log \frac{1}{K} e^{\eta (x_j)_+} + \frac{\log K}{\eta} \le \Phi(x) + \frac{\log K}{\eta}\,.$$

The potential function therefore serves as an upper bound on the sup distance, up to an additive logarithmic factor.

The function Φ defined in this way is clearly convex and zero on the negative orthant. To verify that it is a potential, it remains to show that Φ can be bounded by a quadratic.

Away from the negative orthant, Φ is twice differentiable and we can compute the Hessian explicitly:

$$\nabla^2 \Phi(x) = \eta \operatorname{diag}(\nabla \Phi(x)) - \eta \nabla \Phi \nabla \Phi^\top\,.$$

For any vector u such that $\|u\|_2 = 1$, we therefore have

$$u^\top \nabla^2 \Phi(x) u = \eta \sum_{j=1}^{K} u_j^2 (\nabla \Phi(x))_j - \eta (u^\top \nabla \Phi(x))^2 \le \eta \sum_{j=1}^{K} u_j^2 (\nabla \Phi(x))_j \le \eta\,,$$

since $\|u\|_2 = 1$ and $\|\nabla \Phi(x)\|_1 \le 1$.

We conclude that $\nabla^2 \Phi(x) \preceq \eta I$, which for nonnegative x and y implies the bound

$$\Phi(y) - \Phi(x) - \langle \nabla \Phi(x), y - x \rangle \le \frac{\eta}{2} \|y - x\|^2\,.$$

In fact, this bound holds everywhere. Therefore Φ is a valid potential function for the negative orthant, with $h = \eta$.

The above theorem then implies that we can ensure

$$\frac{R_n}{n} \leq \Phi(\bar{\ell}_n) + \frac{\log K}{\eta} \leq \frac{4R^2 \eta \log n}{n} + \frac{\log K}{\eta}\,.$$

To optimize this bound, we pick $\eta = \frac{1}{2R}\sqrt{\frac{n \log K}{\log n}}$ and obtain the bound

$$R_n \leq 4R\sqrt{n \log n \log K}\,.$$

As alluded to earlier, a more careful analysis can remove the $\log n$ term. Indeed, for this particular choice of Φ, we can modify the above Lemma to obtain the sharper bound

$$\langle x - \pi(x), \nabla\Phi(x)\rangle \geq 2\Phi(x)\,.$$

When we substitute this expression into the above proof, we obtain the recurrence relation

$$\Phi(\bar{\ell}_t) \leq \frac{t-2}{t}\Phi(\bar{\ell}_{t-1}) + \frac{c}{t^2}\,.$$

This small change is enough to prevent the appearance of $\log n$ in the final bound.