LECTURE #3

APRIL 6, 2020

LECTURER: KFIR LEVY SCRIBE: ILYA OSADCHIY

1 Preliminaries

1.1 Dual norm

Definition 1.1. Let $\|\cdot\|$ be a norm. Its dual norm is

$$\|y\|_* \triangleq \max_{\|x\|\leq 1} x^T y$$

Dual norm is denoted by $*$ either as subscript or as superscript.

Example 1.2. Dual of $\|\cdot\|_2$ is $\|\cdot\|_2$.

Example 1.3. For a matrix A we define $\|x\|_A \triangleq \sqrt{x^T Ax}$. Then $\|x\|_A^* = \|x\|_{A^{-1}}$.

Example 1.4. For $p \geq 1$ we define $\|x\|_p \triangleq \left(\sum_i x_i^p\right)^{1/p}$. Then $\|x\|_p^* = \|x\|_q$ for q such that $\frac{1}{p} + \frac{1}{q} = 1$.

Lemma 1.5. *Generalized Cauchy–Schwarz inequality*

$$x^T z \leq \|x\| \|y\|_*$$

1.2 Strong convexity

Definition 1.6. Let $\mathcal{K}$ be a convex compact set and $\|\cdot\|$ be a general norm. Let function $\mathcal{R} : \mathcal{K} \to \mathbb{R}$.
$\mathcal{R}$ is μ-strongly convex if

$$\mathcal{R}(y) \geq \mathcal{R}(x) + \nabla\mathcal{R}(x)^T (y - x) + \frac{\mu}{2}\|x - y\|^2$$

1.3 Bregman divergence

Definition 1.7. Let $\mathcal{R}$ be a convex and differentiable function. Its Bregman divergence is

$$B_{\mathcal{R}}(x, y) \triangleq \mathcal{R}(x) - \mathcal{R}(y) - \nabla\mathcal{R}(y)^T (x - y)$$

Intuition: a distance function induced by a convex function.
For a linear function $B_{\mathcal{R}} \equiv 0$.
Adding a linear term to a function doesn't change its Bregman divergence.
If $\mathcal{R}(\cdot)$ is 1-strongly convex w.r.t. $\|\cdot\|$ then $B_{\mathcal{R}}(x, y) \geq \frac{1}{2}\|x - y\|^2$, and $B_{\mathcal{R}}(x, y)$ is 1-strongly convex in x.
If $\mathcal{R}$ is convex, $B_{\mathcal{R}}(x, y) \geq 0,\ \forall x, y$.

Example 1.8. $\mathcal{R}(x) = ax + b \implies B_{\mathcal{R}}(x, y) = 0$.

Example 1.9. $\mathcal{R}(x) = \frac{1}{2}\|x\|_2^2 \implies B_{\mathcal{R}}(x, y) = \frac{1}{2}\|x - y\|_2^2$.

Example 1.10. We denote simplex: $\Delta \triangleq \left\{x \in \mathbb{R}^N, \sum_{i=1}^N x_i = 1, \forall i : x_i \geq 0\right\}$.

Let $\mathcal{R} : \Delta \to \mathbb{R}$ be negative entropy:

$$\mathcal{R}(p) = \sum_{i=1}^{N} p(i) \log p(i)$$

Then its Bregman divergence is relative entropy:

$$\forall p, q \in \Delta : \quad B_{\mathcal{R}}(p, q) = \sum_{i=1}^{N} p(i) \log \frac{p(i)}{q(i)}$$

Also, $\mathcal{R}(p)$ is 1-strongly convex w.r.t. $\|\cdot\|_1$ on Δ.

Example 1.11. Let $\{c_i\}$ be constants, and let $\mathcal{R} : \Delta \to \mathbb{R}$ be barrier funciton:

$$\mathcal{R}(p) = \sum_{i=1}^{N} c_i \log \frac{1}{p(i)}$$

Then its Bregman divergence is

$$\forall p, q \in \Delta : \quad B_{\mathcal{R}}(p, q) = \sum_{i=1}^{N} c_i \left(\log \frac{q(i)}{p(i)} + \frac{p(i) - q(i)}{q(i)} \right)$$

1.4 Optimality in convex optimization

Lemma 1.12. *Let $\mathcal{K} \succeq \mathbb{R}^d$ be a convex compact set, and let $f : \mathcal{K} \to \mathbb{R}$ be a convex function.*

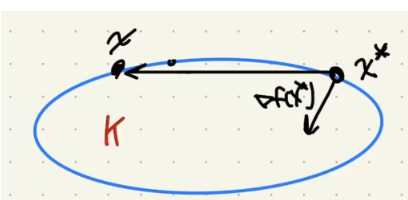
Denote

$$x^* = \arg\min_{x \in \mathcal{K}} f(x)$$

Then

$$\forall x \in \mathcal{K} : \quad \nabla f(x^*)^T (x - x^*) \geq 0$$

The intuition is that angle between the gradient at optimal point and any other point in the set is not greater than 90 deg, otherwise there would exist a feasible descent direction contradicting the optimality of x^*.



1.5 Derivative of Bregman divergence

Lemma 1.13. *Differentiating Bregman divergence by the first argument gives:*

$$\nabla_x B_{\mathcal{R}}(x, y) = \nabla \mathcal{R}(x) - \nabla \mathcal{R}(y)$$

1.6 Three point inequality

Lemma 1.14. *The following inequality holds:*

$$B_{\mathcal{R}}(x, y) + B_{\mathcal{R}}(y, z) \leq B_{\mathcal{R}}(x, z) + (\nabla \mathcal{R}(z) - \nabla \mathcal{R}(y))^T (x - y)$$

2 Online Mirror Descent

2.1 Motivation

There are two meta-algorithms (templates) for online learning with sublinear regret bounds. The algorithms that we've seen earlier (Hedge, OGD) can be obtained as private cases of these templates. These two meta-algorithms are:

- Online Mirror Descent (OMD)
- Follow the Regularized Leader (FTRL)

In the following we look into the OMD meta-algorithm. We start by looking at an imaginary game with different setting than the regular OCO protocol: now the loss function at each round is known to the player before he makes the decision. For every $t \in [T]$:

- The adversary reveals the loss function $f_t(\cdot)$.
- The player picks $x_t \in \mathcal{K}$ and incurs loss $f_t(x_t)$.

This setting is much easier than OCO. One possible algorithm in this setting is Best Response:

$$x_t = \arg\min_{x \in \mathcal{K}} f_t(x)$$

Theorem 2.1. *Best Response ensures* $\mathrm{Reg}_T \leq 0$.

Proof. By definition

$$f_t(x_t) \leq f_t(x), \quad \forall x \in \mathcal{K}$$
$$\mathrm{Reg}_T = \min_{x \in \mathcal{K}} \sum_{t=1}^{T} (f_t(x_t) - f_t(x)) \leq 0$$

□

Comments:

- Best Response doesn't require convexity of the loss functions.
- In literature Best Response may refer to a similar algorithm in the standard OCO setting.

2.2 The algorithm

We now introduce an algorithm that is inspired by Best Response, but works in the standard OCO setting. Reminder, the OCO protocol is: For every $t \in [T]$:

- The player picks $x_t \in \mathcal{K}$.
- The adversary reveals the loss function $f_t(\cdot)$.
- The player incurs loss $f_t(x_t)$ and receives $f_t(\cdot)$ as feedback.

Let's assume that we have a function (called regularization function) $\mathcal{R} : \mathcal{K} \to \mathbb{R}$, such that $\mathcal{R}$ is 1-strongly convex w.r.t.a norm $\|\cdot\|$. And let $B_{\mathcal{R}}(x, y)$ denote Bregman divergence induced by $\mathcal{R}$.

Algorithm 1: Online Mirror Descent

Parameters: $\mathcal{R} : \mathcal{K} \to \mathbb{R}$, $\eta > 0$

$x_1 = \arg\min_{x \in \mathcal{K}} \mathcal{R}(x)$;

for $t \in [T]$ **do**

  $x_{t+1} = \arg\min_{x \in \mathcal{K}} \left(f_t(x) + \frac{1}{\eta} B_{\mathcal{R}}(x, x_t) \right)$;

end

Unlike Best Response, this algorithm uses the last observed loss function and not the next one. It uses proximity to the last point to ensure stability of the solution and incorporate information from earlier rounds.

Remarks on linearizing the loss function:

- If losses are linear the rule becomes

$$x_{t+1} = \arg\min_{x \in \mathcal{K}} \left(g_t^T x + \frac{1}{\eta} B_{\mathcal{R}}(x, x_t) \right)$$

- If losses are general convex functions, we can linearize them:

$$g_t = \nabla f_t(x), \quad x_{t+1} = \arg\min_{x \in \mathcal{K}} \left(g_t^T x + \frac{1}{\eta} B_{\mathcal{R}}(x, x_t) \right)$$

- Such linearization simplifies the optimization problem.
- This is a pessimistic approximation of the loss:

$$\begin{aligned} &\tilde{f}_t(x) = f_t(x_t) + \nabla f_t(x_t)\,(x - x_t) \\ &\forall x : \; f_t(x_t) - f_t(x) \leq \tilde{f}_t(x_t) - \tilde{f}_t(x_t) = \nabla f_t(x_t)^T (x_t - x) \end{aligned}$$

- The original loss or another approximation of it may be used instead.

Design choices in the OMD algorithm:

- Regularization function $\mathcal{R}$ and the corresponding $B_{\mathcal{R}}$.
- Approximation of the losses.
- Learning rate η.

2.3 Regret bound

We will denote:

$$D \triangleq \sqrt{\max_x B_{\mathcal{R}}\,(x, y)}, \quad y = \arg\min_{y \in \mathcal{K}} \mathcal{R}(y)$$

$$G \triangleq \begin{cases} \max_{t \in [T]} \|g_t\|_*, & \text{linear loss} \\ \\ \max_{t \in [T], x \in \mathcal{K}} \|\nabla f_t(x)\|_*, & \text{general convex loss} \end{cases}$$

Theorem 2.2 (OMD regret bound)**.** *Setting* $\eta = \frac{D}{G\sqrt{T}}$ *guarantees* $\mathrm{Reg}_T \leq GD\sqrt{T}$.

Note that such G and D may be better than what we've previously had for Hedge and OGD, because we can choose $\mathcal{R}$ and the norm.

2.4 Private cases

2.4.1 OGD

We will set $\mathcal{R}(x) = \frac{1}{2}\|x\|_2^2$, and use linearized loss.

Then OMD update rule becomes:

$$\begin{aligned}
x_{t+1} &= \arg\min_{x\in\mathcal{K}} \left(g_t^T x + \frac{1}{2\eta}\|x - x_t\|_2^2 \right) \\
&= \arg\min_{x\in\mathcal{K}} \left(\frac{1}{2\eta}\|x - x_t + \eta g_t\|_2^2 + \text{const} \right) \\
&= \arg\min_{x\in\mathcal{K}} \left(\|x - (x_t - \eta g_t)\|_2 \right) \\
&= \Pi_{\mathcal{K}}^{\|\cdot\|_2}(x_t - \eta g_t)
\end{aligned}$$

Therefore the update rule in this case is same as OGD.

2.4.2 Hedge

We look at the experts setting.

The player decisions lie in the simplex: $p \in \mathcal{K} = \Delta = \left\{ p \in \mathbb{R}^N, \sum_{i=1}^N p(i) = 1, \forall i : p(i) \geq 0 \right\}$.

The loss is a combination of per-expert losses: $f_t p = \mathcal{L}_t^T p$, $\mathcal{L}_t = (\mathcal{L}_t(1), \mathcal{L}_t(2), \ldots, \mathcal{L}_t(N))$, $\mathcal{L}_t(i) \in [0, 1]$

We will use negative entropy as regularization: $\mathcal{R}(x) = \sum_{i=1}^N p(i)\log p(i)$.

Then OMD update rule becomes:

$$p_{t+1} = \arg\min_{p\in\Delta} \left(\mathcal{L}_t^T p + \frac{1}{\eta}\sum_{i=1}^N p(i)\log\frac{p(i)}{p_t(i)} \right)$$

We will solve it ignoring the inequality constraints of Δ, they will be satisfied anyway. The problem becomes

$$\arg\min_{\sum_{i=1}^N p(i)=1} \left(\mathcal{L}_t^T p + \frac{1}{\eta}\sum_{i=1}^N p(i)\log\frac{p(i)}{p_t(i)} \right)$$

Its Lagrangian is:

$$L(p,\lambda) = \mathcal{L}_t^T p + \frac{1}{\eta}\sum_{i=1}^N p(i)\log\frac{p(i)}{p_t(i)} + \lambda\left(\sum_{i=1}^N p(i) - 1 \right)$$

$$\frac{\partial L(p,\lambda)}{\partial p(i)} = \mathcal{L}_t(i) + \frac{1}{\eta}\left(\log\frac{p(i)}{p_t(i)} + 1 \right) + \lambda = 0$$

$$\log\frac{p(i)}{p_t(i)} = -\eta\mathcal{L}_t(i) - \eta(1+\lambda)$$

$$p(i) = C p_t(i) e^{-\eta\mathcal{L}_t(i)}, \qquad C = e^{-\eta(1+\lambda)}$$

We need to set λ to satisfy the constraints, which is equivalent to setting C to a positive number that satisfies the sum constraint; the non-negativeness constraint is satisfied because C, $p_t(i)$ and $e^{-\eta\mathcal{L}_t(i)}$ are all non-negative.

We can also write the decision rule as:

$$p(i) \propto p_t(i) e^{-\eta\mathcal{L}_t(i)}$$

And if we apply it recursively we get:

$$
\begin{aligned}
p(i) &\propto p_t(i) e^{-\eta \mathcal{L}_t(i)} \\
&\propto p_{t-1}(i) e^{-\eta (\mathcal{L}_{t-1}(i) + \mathcal{L}_t(i))} \\
&\propto \dots \\
&\propto p_1(i) e^{-\eta \mathcal{L}_{1:t}(i)}
\end{aligned}
$$

Recall that p_1 must minimize $\mathcal{R}(p)$, which in this case means $p_1(i) = \frac{1}{N} \quad \forall i$. And this gives us the decision rule that we had in the Hedge algorithm: $p_{t+1}(i) \propto e^{-\eta \mathcal{L}_{1:t}(i)}$.