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 $\|x\|_A \triangleq \sqrt{x^T A x}$. 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 K be a convex compact set and $\|\cdot\|$ be a general norm. Let function $\mathcal{R} : K \to \mathbb{R}$.
$\mathcal{R}$ 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

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

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

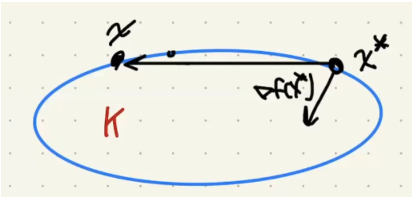
Denote

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

Then

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

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:*

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

2 Online Mirror Descent

2.1 Motivation

There are two meta-algorithms (templates) for online learning that can have 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 $x_t \in K$ and incurs loss $f_t(x_t)$.

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

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

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

Proof. By definition

$$f_t(x_t) \leq f_t(x), \quad \forall x \in K$$

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

□

Remark: Best Response doesn't require convexity of the loss functions.

2.2 The algorithm

linear, convex

2.3 Regret bound

2.4 Private cases

2.5 Proof of regret bound

3 Notation

We use the following mathematical notation in this writeup:

- d-dimensional Euclidean space is denoted $\mathbb{R}^d$.
- Vectors are denoted by boldface lower-case letters such as $\mathbf{x} \in \mathbb{R}^d$. Coordinates of vectors are denoted by regular brackets $\mathbf{x}(i)$
- Matrices are denoted by boldface upper-case letters such as $\mathbf{X} \in \mathbb{R}^{m \times n}$. Their coordinates by $\mathbf{X}(i, j)$.

- Functions are denoted by lower case letters $f : \mathbb{R}^d \mapsto \mathbb{R}$.
- The k-th differential of function f is denoted by $\nabla^k f \in \mathbb{R}^{d^k}$. The gradient is denoted without the superscript, as ∇f.
- We use the mathbb macro for sets, such as $\mathcal{K} \subseteq \mathbb{R}^d$.