Advanced Topics in ML and AGT **Fall Semester, 2017/18**

Lecture 3: November 6

Lecturer: Yishay Mansour *Scribe: Dana Cohen, Ofir Epstein, Roey Rozi*

3.1 Convex functions

Definition $f(x)$ is a convex function if $\forall x, y \in S, \forall \theta \in [0,1]$:

$$f(\theta x + (1-\theta)y) \ \le \ \theta f(x) + (1-\theta) f(y)$$

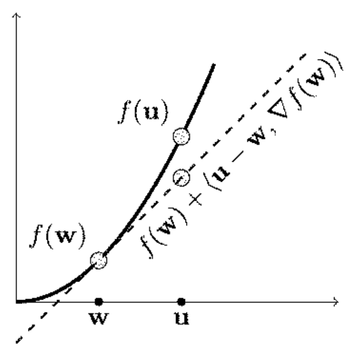

Figure 3.1: A convex function.

Theorem 3.1 *Let S be a convex domain. $f(x)$ is convex iff*

$$\forall x, y \in S: \ \ f(y) \ \ge \ f(x) \ + \ (\nabla f(x))^T (y - x)$$

Proof:
Let us begin by proving the theorem for a single dimensional function $f : \mathbb{R} \to \mathbb{R}$

First direction: f is convex $\Rightarrow$ $\forall x, y \in S: \ f(y) \ge f(x) + (\nabla f(x))^T (y - x)$
Using the definition of convexity on S:

$$\forall \theta \in [0,1], \forall x, y \in S: \ (1-\theta)x + \theta y = x + \theta(y - x) \in S$$

Using the definition of convexity on f:

$$\begin{aligned}
f(x+\theta(y-x)) &\leq (1-\theta)f(x)+\theta f(y)\\
f(x+\theta(y-x))-f(x) &\leq -\theta f(x)+\theta f(y)\\
(y-x)\frac{f(x+\theta(y-x))-f(x)}{\theta(y-x)}+f(x) &\leq f(y)\\
\lim\theta\to 0: \quad f(x)+(y-x)f'(x) &\leq f(y)
\end{aligned}$$

Second direction: $\forall x,y\in S: \ f(y)\geq f(x)+f'(x)(y-x)\Rightarrow f$ is convex.
Define: $z=(1-\theta)x+\theta y\in S$
Using the given inequality with z as well as x and y:

$$\begin{array}{lll}
(1) & f(x)\geq f(z)+f'(z)(x-z) \quad\Rightarrow & f(x)(1-\theta)\geq(f(z)+f'(z)(x-z))(1-\theta)\\
(2) & f(y)\geq f(z)+f'(z)(y-z) \quad\Rightarrow & f(y)\theta\geq(f(z)+f'(z)(y-z))\theta
\end{array}$$

Adding (1) and (2) and simplifying:

$$(1-\theta)f(x)+\theta f(y)\geq f(z)+f'(z)((1-\theta)x+\theta y-z)=f(z)+f'(z)(z-z)=f(z)$$

Proving that f is convex.

Finally, we will prove the theorem for $f:\mathbb{R}^d\to\mathbb{R}$:

For $x,y\in\mathbb{R}^d$ we will define a function $g:\mathbb{R}\to\mathbb{R}$:

$$g(\theta)=f((1-\theta)x+\theta y)$$

The function g is differentiable,

$$g'(\theta)=(\nabla f((1-\theta)x+\theta y))^T(y-x)$$

First direction: f is convex $\Rightarrow\forall x,y\in S: \ f(y)\geq f(x)+(\nabla f(x))^T(y-x)$.

f is convex iff it is convex on each of the lines between x and y.
The function g is defined on the line between x and $y\to g$ is convex.
From g's convexity:

$$g(1)\geq g(0)+g'(0)(1-0)=g(0)+g'(0)$$

Substituting in g's definition:

$$f(y)\geq f(x)+(\nabla f(x))^T(y-x)$$

Second direction: $\forall x,y\in S: \ f(y)\geq f(x)+(\nabla f(x))^T(y-x)\to f$ is convex
For any 2 points x and y, define:

$$(1-\theta)x+\theta y \ \ \textit{and} \ \ (1-\tilde{\theta})x+\tilde{\theta}y$$

Using the given inequality:

$$f((1-\theta)x+\theta y)\geq f((1-\tilde{\theta})x+\tilde{\theta}y)+(\nabla f((1-\tilde{\theta})x+\tilde{\theta}y))^T(y-x)(\theta-\tilde{\theta})$$

We can substitute g and g' and obtain:

$$g(\theta) \geq g(\tilde{\theta}) + g'(\tilde{\theta})(\theta - \tilde{\theta})$$

Because g is single dimensional we can use the single dimensional theorem, showing that g is convex.

g is defined and convex on all lines between x and $y \Rightarrow f$ is convex (f is convex iff it is convex on each of the lines between x and y).

$\square$

Corollary 3.2 *Let S be convex domain, $f : S \to \mathbb{R}$ a convex function. Then:*

$$\forall w \in S,\ \exists z\ s.t.\ \forall u \in S\ f(u) \geq f(w) + z^T(u - w)$$

We can prove this easily using theorem (3.1) and defining $z = \nabla f(w)$.

3.2 Sub-gradient

Definition z is a sub-gradient of f at w if

$$\forall u \in S:\ f(u) \geq f(w) + z^T(u - w)$$

The set of sub-gradients of f at w is denoted $\partial f(w)$.

Notice that if f is differentiable then $\partial f(w) = \{\nabla f(w)\}$

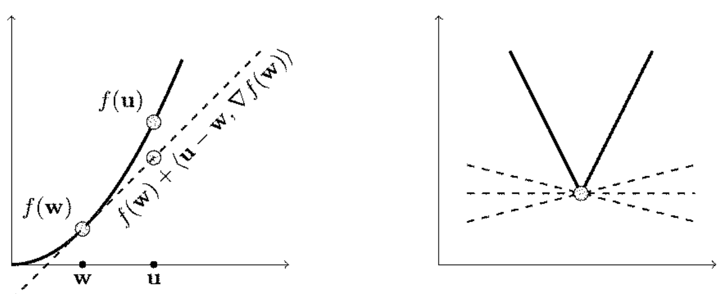

Figure 3.2: Sub-gradient of a differentiable function vs. non-differentiable function.

We shall look at: $f_1, ..., f_T$ (loss functions) - a sequence of convex function, $w_1, ..., w_T$ vectors and $z_t \in \partial f_t(w_t)$.

$$\forall u:\ \text{Regret}_{\text{online convex optimization}}(u) = \sum_{t=1}^{T} f_t(w_t) - f_t(u)$$

Using the definition of sub-gradient:

$$z^T w - z^T u = z^T(w - u) \geq f(w) - f(u)$$

And we can get an upper bound for the regret of the online convex optimization problem:

$$\forall u: \ \sum_{t=1}^{T} f_t(w_t) - f_t(u) \ \leq \ \sum_{t=1}^{T} w_t^T z_t - \sum_{t=1}^{T} u^T z_t \ = \ \text{Regret}_{\text{online linear optimization}}$$

In Lecture 2 we proved an upper bound on online linear optimization:

$$\text{Regret}_T(u) \leq \frac{1}{2\eta}||u||_2^2 + \eta \sum_{t=1}^{T} ||z_t||_2^2$$

Important note: Notice that z_t is dependent on w_t ($z_t \in \partial f_t(w_t)$). This is not an issue because in our online learning model w_t is defined by the past.

Corollary 3.3 *We can solve online convex optimization by using an algorithm that solves the online linear optimization problem. The upper bound of the regret is bound by the regret of online linear optimization.*

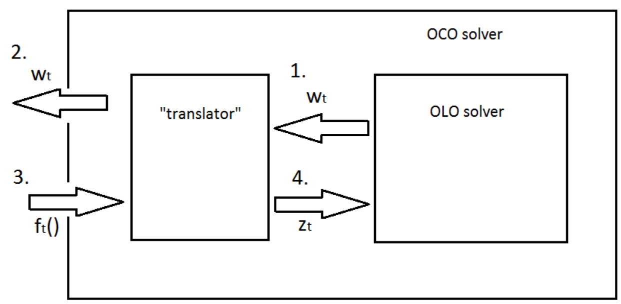

Figure 3.3: A scheme showing how to solve OCO using OLO as described in corollary 3.3.

3.2.1 Online linear optimization

Algorithm:

1. Define parameter: $\eta > 0$.
2. Init: $w_1 = 0$.
3. Update: $w_{t+1} = w_t - \eta z_t$ (for $z_t \in \partial f_t(w_t)$).

This algorithm is identical to the algorithm we saw for Follow The Regularized Leader, hence we get:

$$\text{Regret}_T(u) \leq \frac{1}{2\eta}||u||_2^2 + \eta \sum_{t=1}^{T} ||z_t||_2^2 \tag{3.1}$$

We would like to show a connection between z_t and the fact that the functions are Lipschitz. More precisely, we would like to show that if f_t is L-Lipshcitz then z_t is bounded by L.

3.2.2 Dual Norm

Definition Let $\|\cdot\|$ be a norm. The dual norm $\|\cdot\|_*$ is defined to be:

$$\|\cdot\|_* = \max_w \{w^T z : \|w\| \leq 1\}$$

Examples:

We will show some examples of common dual norms:

1. L_2 :

$$\|z\|_{*2} = \max_{\|w\|_2 \leq 1} w^T z \Rightarrow w = \frac{z}{\|z\|}$$
$$= \left(\frac{z}{\|z\|}\right)^T \cdot z = \|z\|_2$$

2. L_1 :

$$\|z\|_{*1} = \max_{\|w\|_1 \leq 1} w^T z = \max_i |z[i]| = \|z\|_\infty$$

3. L_∞ :

$$\|z\|_{*\infty} = \max_{\|w\|_\infty \leq 1} w^T \Rightarrow w[i] = sign(z[i])$$
$$= \sum_i |z[i]| = \|z\|_1$$

Lemma 3.4 *Let f be a convex function. $f(x)$ is L-Lipschitz for norm $\|\cdot\|$ iff:*

$$\forall w \in S,\ \ \forall z \in \partial f_t(w) : \ \ \|z\|_* \leq L$$

Proof:

$\Rightarrow$ **First direction:**
Let $f$be a L-Lipschitz function.
Choose $w \in S$ and $z \in \partial f_t(w)$.
Choose u such that $u - w = argmax_{\|w\|=1}$.
Therefore $\|z\|_* = (u-w)^T z \leq f(u) - f(w) \leq L\|u-w\| = L$
The first inequality is from the definition of the gradient, the second is from the Lipschitz property and the third is from the definition of u.

$\Leftarrow$ **Second direction:**
$f(u) - f(w) \leq (u-w)^T z \leq \|w-u\|\|z\|_*$. Since $t\|z\|_* \leq L$. This implies that $f(u) - f(w) \leq \|(u-w)\|L$, hence f is L-Lipschitz

□

Note: Therefore, the term $\sum_{t=1}^T \|z_t\|^2$ given in Equation (3.1) can be bounded by $\sum_{t=1}^T L_t$, where L_i is the Lipschitz constant of f_t.

Regression problem:

Consider the regression problem: $f_t(w) = |w^T x_t - y_t|$ If $\|x_t\| \le L$ then f is L-Lipschitz and we can get the same bound we got in lecture 2: $B \cdot L\sqrt{2T}$

Experts problem:

We would like to use lemma (3.4) in order to show bound to the experts problem. This leads us into a problem because OGD doesn't promise us that w_t would be distribution vectors.

3.2.3 Strong Convexity

Definition $f : S \to \mathbb{R}$ is σ-strong-convex over S with norm $\|\cdot\|$ if:

$$\forall w \in S, \forall\ z \in \partial f_t(w), \forall\ u \in S: \ \ f(u) \ge f(w) + (u-w)^T z + \frac{\sigma}{2}\|u-w\|^2$$

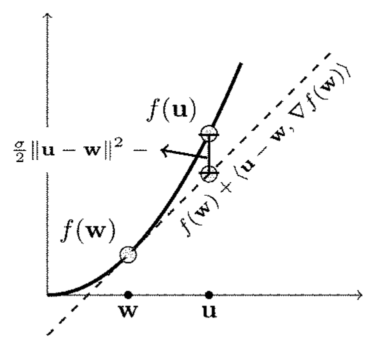

Figure 3.4: A strongly convex function f.

Lemma 3.5 *Let $f : S \to \mathbb{R}$ be a σ-strong-convex function over S with norm $\|\cdot\|$, and $w^* = \arg\min_{v\in S} f(v)$. Then:*

$$\forall u \in S\ \ f(u) - f(w^*) \ge \frac{\sigma}{2}\|u-w^*\|^2$$

Proof:
First case: We will assume that f is differentiable and w^* is inside S.
Then $\nabla f(w^*) = 0$ from the definition of strong convexity. Then:

$$f(u) - f(w^*) \ge (u-w^*)^T \nabla f(w^*) + \frac{\sigma}{2}\|u-w^*\|^2 = \frac{\sigma}{2}\|u-w^*\|^2$$

and we are done.

Second case: w^* is on the boundary of S.
If w^* is on the boundary of S, still $(u-w^*)^T\nabla f(w^*) \ge 0$. (Otherwise, it is possible to advance in the direction of u and improve the minimization of f).

Third case: f is not differentiable. We can extend f's source to $\mathbb{R}^d$ by defining g (a proper function):

$$g(x) = \begin{cases} f(w), & \text{if } w \in S. \\ \infty & \text{if } w \notin S \end{cases} \tag{3.2}$$

It is still guaranteed that

$$w^* = \arg\min g(w) \tag{3.3}$$

Claim 3.6 $0 \in \partial f_t(w^*)$

Proof:

$$\forall w \;\; g(w^*) \leq g(w)$$
$$\forall w : g(w) \leq g(w^*) + 0^T(w - w^*)$$

Therefore, $0 \in \partial g(w)$. □

Using claim (3.6), because g is σ-strong-convex:

$$\forall u \in S \;\; f(u) - f(w) = g(u) - g(w^*) \geq (u-w)^T 0 + \frac{\sigma}{2}\|u - w^*\|^2 = \frac{\sigma}{2}\|u - w^*\|^2$$

□

3.2.4 Hessian function and strong convexity

We will now show a characterization of strong convexity using the Hessian function:

Claim 3.7 *If f is convex, and for all $x \in S$ the Hessian, $H_f(x) = \nabla^2 f(x) = \left[\frac{\partial^2}{\partial x_i \partial x_j} f(x)\right]$, is positive definite, then f is strongly-convex.*

Proof: The Taylor series of f:

$$\forall x \exists z \in [x, y] : \quad f(y) = f(x) + (y-x)^T \nabla f(x) + \frac{1}{2}(y-x)^T H_f(z)(y-x) \tag{3.4}$$

Since H_f is positive-definite, $\forall x : x^T H_f x > 0$. In particular, $\frac{1}{2}(y-x)^t H_f(z)(y-x) > 0$.
This means that by definition, f is strongly-convex (assuming $(y-x)^T H_f(z)(y-x) \geq ||y-x||^2$). □

Corollary 3.8 *Assuming for the regularization function R:*

$$\forall w : \; x^T(\nabla^2 R(w))x \geq \sigma ||x||^2 \tag{3.5}$$

Then R is σ-strongly-convex.

Examples:

1. Euclidean Regularization: $R(w) = \frac{1}{2}||w||_2^2$.
   Then $\nabla^2 R(w) = I$. So R is 1-strongly-convex.

2. Entropic Regularization: $R(w) = \sum_{i=1}^d w_i \log w_i$. (assuming $\sum w_i = B$)
   Then: $\nabla^2 R(w) = \begin{pmatrix} \frac{1}{w_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{w_d} \end{pmatrix}$. So:

$$x^T(\nabla^2 R(w))x = \sum_{i=1}^d \frac{x_i^2}{w_i} = \frac{1}{||w||_1}\left(\sum_{i=1}^d w_i\right)\left(\sum_{i=1}^d \frac{x_i^2}{w_i}\right) \geq \frac{1}{||w||_1}\left(\sum_{i=1}^d w_i \frac{x_i}{w_i}\right)^2 = \frac{||x||_1^2}{||w||_1} = \frac{1}{B}||x||_1^2$$

   Where the inequality is received using the Cauchy-Schwarz inequality.

3.2.5 Follow the Regularized Leader with strongly-convex regularization

Last time we saw that:

$$\text{Regret}_T(u) = \sum_{i=1}^{T} f_t(w_t) - f_t(u) \leq R(u) - R(w_1) + \sum_{t=1}^{T} f_t(w_t) - f_t(w_{t+1})$$

Assuming f_t is L_t-Lipschitz we get:

$$\text{Regret}_T(u) \leq R(u) - R(w_1) + \sum_{i=1}^{T} L_t ||w_t - w_{t+1}||$$

Now we'd like to show that $||w_t - w_{t+1}||$ is small enough.

Lemma 3.9 *Let R be a σ-strongly-convex regularization, $w_1, \ldots, w_T$ the vectors FoReL chooses. If for all t, f_t is L_t-Lipschitz, then:*

$$f_t(w_t) - f_t(w_{t+1}) \leq L_t \|w_t - w_{t+1}\| \leq \frac{L_t^2}{\sigma}$$

Proof: In stage t, the algorithm calculates $F_t(w) = \sum_{i=1}^{t-1} f_i(w) + R(w)$ and chooses $w_t = \arg\min_w F_t(w)$. R is σ-strongly-convex, which means F_t are also σ-strongly-convex. So we have:

$$F_t(w_{t+1}) \geq F_t(w_t) + \frac{\sigma}{2} \|w_t - w_{t+1}\|^2$$
$$F_{t+1}(w_t) \geq F_{t+1}(w_{t+1}) + \frac{\sigma}{2} \|w_t - w_{t+1}\|^2$$

We add the two and get (using the fact that $F_{t+1}(w) = F_t(w) + f_t(w)$:

$$f_t(w_t) \geq f_t(w_{t+1}) + \sigma \|w_t - w_{t+1}\|^2 \tag{3.6}$$

And so:

$$L_t\|w_t - w_{t+1}\| \underset{(1)}{\geq} f_t(w_t) - f_t(w_{t+1}) \underset{(2)}{\geq} \sigma\|w_t - w_{t+1}\|^2$$

Where (1) is by the Lipschitz property, and (2) is by (3.6). This implies,

$$\|w_t - w_{t+1}\| \leq \frac{L_t}{\sigma}$$

□

We will now see how this affects the previous bounds we found:

OQO:

Here $R(w) = \frac{1}{2\eta}\|w\|_2^2$. So $\sigma = \frac{1}{\eta}$. We assume $\exists L$ such that $\forall t : L_t \leq L$. Then we get:

$$\text{Regret} \leq \frac{1}{2\eta}\|u\|_2^2 + \eta L^2 T$$

Which is the same bound we already saw.

The Experts Problem:

We want to always choose a distribution, so $S = \{w : \ w > 0, \ \|w\|_1 = 1\}$. We will make sure that for $w \notin S$, $R(w) = \infty$, and define:

$$R(w) = \begin{cases} \frac{1}{2\eta}\|w\|_2^2 & w \in S \\ \infty & w \notin S \end{cases}$$

So the regret:

$$\text{Regret} \leq \frac{q}{2\eta}\|u\|_2^2 + \eta T L^2$$

For $\|u\| \leq B$, $\eta = \frac{B}{L\sqrt{2T}}$ we get the bound:

$$\text{Regret} \leq BL\sqrt{2T}$$

We got this bound because we used a non-continuous regularization, and used a sub-gradient. In the experts problem, $B = 1$, $L = \sqrt{d}$ (u are unit vectors), so we have the bound Regret $\leq \sqrt{2dT}$.

We would like to reduce this bound, so we can use an entropic regularization, which will guarantee Regret $\leq \sqrt{2\log(d)T}$, using the following corollary.

Corollary 3.10 *If $f_1, \ldots, f_T$ are convex, L-Lipschitz (in relation to norm $\|\cdot\|_2$), then for $R(w) = \frac{1}{\eta}\sum_{i=1}^d w_i \log w_i$ and $S = \{w : \ \|w\|_1 \leq B, w > 0\}$:*

$$\textit{Regret} \leq \frac{B \log d}{\eta} + \eta B T L^2 \leq BL\sqrt{2 \log dT}$$

Notice that $R(u) \leq \frac{1}{\eta}\log d$, so for the experts problem:

$$|f_t(w) - f_t(u)| = |(w-u)^T x_t| \leq \|w-u\|_1 \cdot \|x_t\|_\infty \leq \|w-u\|_1$$

By Holder's inequality.

So $L_t = 1$ with respect to $\| \cdot \|_1$. In addition, $B = 1$ (since we need a distribution), so:

$$\text{Regret} \leq \sqrt{2 \log dT}$$