Advanced Topics in ML and AGT **Fall Semester, 2018/19**

Lecture 3: October 29th

Lecturer: Yishay Mansour *Scribe: Ben Bogin and Tal Dim*

3.1 Convex functions

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

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

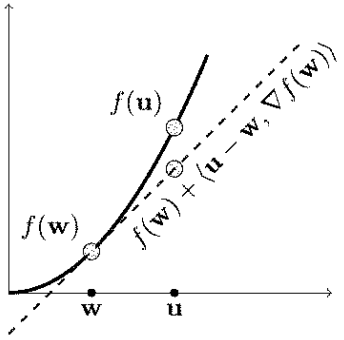

Figure 3.1: A convex function.

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

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

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

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

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

[0]Based on the scribes by Dana Cohen, Ofir Epstein, Roey Rozi (2017/18)

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 for both x and y:

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

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 one-dimensional 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: The function 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, therefore 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: $\forall x,y\in S$: if $f(y)\geq f(x)+(\nabla f(x))^T(y-x)$, then f is convex
For any 2 points x and y, define:

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

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 a one-dimensional function we can use the one-dimensional theorem we have already proved, showing that g is convex.

The function g is defined and convex on all lines between x and y, therefore f is convex (f is convex iff it is convex on each of the lines between x and y).

□

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 corollary 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 $\partial f(w) = \{\nabla f(w)\}$

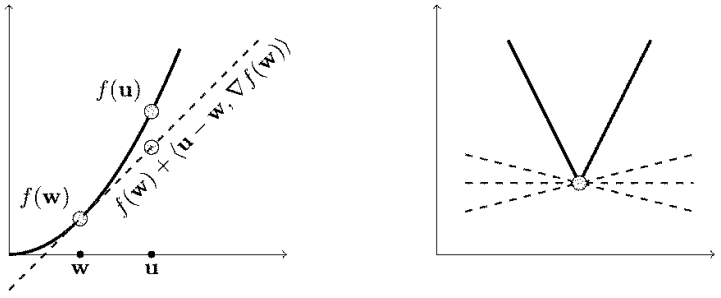

Figure 3.2: Sub-gradient of a differentiable function vs. non-differentiable function.

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

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

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 (OCO) problem:

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

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

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

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.*

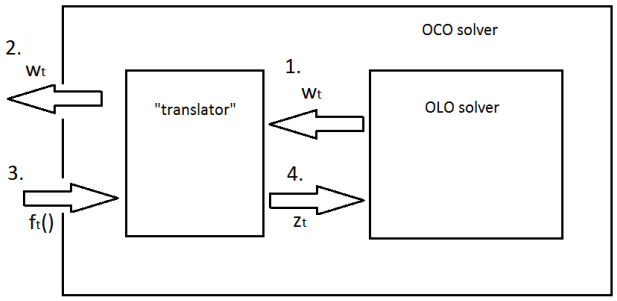

Figure 3.3: A scheme showing how to solve OCO using OLO as described in corollary 3.3.

3.2.1 Online Gradient Descent (OGD)

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:

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

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:

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

Examples:

We will show some examples of common dual norms:

1. L_2 :

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

2. L_1 :

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

3. L_∞ :

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

Lemma 3.4 *Let f be a convex function. Then, $f(x)$ 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 $u - w = \arg\max_{v,\|v\|=1} v^T z$.
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 subgradient, the second is from the Lipschitz property and the third is from the definition of u, namely $\|v\| = \|u - w\| = 1$.

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

□

Corollary 3.5 *If we run OGD on $f_1, ..., f_T$ such that f_t is L_t-Lipshitz with respect to $\|\cdot\|_2$, then*

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

Since $L_t \geq \|z_t\|_{*,2}^2 = \|z_t\|_2^2$.

Note: Lemma 3.4 implies that 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_t is the Lipschitz constant of f_t.

Regression problem (with absolute loss):

Consider the regression problem: $f_t(w) = |w^T x_t - y_t|$ If $\|x_t\|_2 \leq L$ then f is L-Lipschitz with respect to $\|\cdot\|_2$, 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 a bound for the experts problem. The problem is that 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:

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

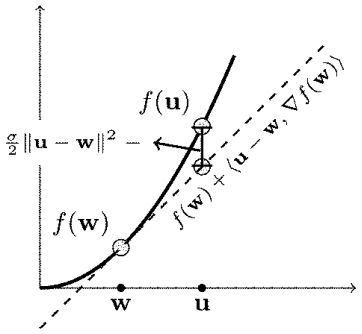

Figure 3.4: A strongly convex function f.

Lemma 3.6 *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:*

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

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:

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

and we are done.

Second case: w^* is on the boundary of S.

If w^* is on the boundary of S, still $(u - w^*)^T \nabla f(w^*) \geq 0$. (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):

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

It is still guaranteed that

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

Claim 3.7 $0 \in \partial f(w^*)$

Proof:

$$\forall w \;\; g(w) \geq g(w^*)$$

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

Therefore, $0 \in \partial g(w)$. □
Using Claim 3.7, 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 3.8 *If f is convex, and $\forall x \in S$ 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, and $\forall z$: $z^T(\nabla^2 f(x))z \geq \frac{\sigma}{2}\|z\|$ then f is σ-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 and $\forall x : x^T H_f x > \frac{\sigma}{2}\|x\|^2$. In particular, $\frac{1}{2}(y-x)^T H_f(z)(y-x) > \frac{\sigma}{2}\|x-y\|^2$. This means that by definition, f is σ-strongly-convex. □

Corollary 3.9 *Assuming for the regularization function R:*

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

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:

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

   Where the inequality is received using the Cauchy-Schwarz inequality. Here $\sigma = \frac{1}{B}$.

3.2.5 Follow the Regularized Leader with strongly-convex regularization

In Lecture 2 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 would like to bound $||w_t - w_{t+1}||$, as a function of f, namely, L_t and σ.

Lemma 3.10 *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:*

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

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 and all f_i are convex, which means F_t are also σ-strongly-convex. So we have:

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

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

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

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 inequality (3.6). This implies,

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

□

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

OQO:

Here $R(w) = \frac{1}{2\eta}\|w\|_2^2$. So $\sigma = \frac{1}{\eta}$.
Using $L^2 = \frac{1}{T}\sum_{t=1}^{T} L_t^2$ the lower bound using the previous part will be:

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

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:

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

If $\max_{w \in S} ||u||_2 \leq B$, $\eta = \frac{B}{L\sqrt{2T}}$ 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. Now we can use that in the experts problem. Since for $u \in S$ we have $|u\|_1 = 1$, then $\|u\|_2 \leq 1$, which implies that $B = 1$. For $x_t \in [0,1]^d$ we have $|x\|_2 \leq \sqrt{d}$. Using $f_t(w) = w^T x_t$ we get:

$$\forall u, w \in S, \ |f_t(u) - f_t(w)| = |(u-w)^T x_T| \leq \|(u-x)\|_{2,*}\|x_t\|_2 \leq \|u-w\|_2\sqrt{d}$$

Thus we infer that $L = \sqrt{d}$. Since $B = 1$ (u are unit vectors), we get the bound: Regret $\leq \sqrt{2dT}$. We would like to reduce this bound and get to the bound: Regret $\leq \sqrt{2\log(d)T}$. To do so we can use an entropic regularization, which is: $R(w) = \frac{1}{\eta}\sum_{i=1}^{d} w_i \log w_i$. using the following corollary.

Corollary 3.11 *If $f_1, \ldots, f_T$ are convex, L-Lipschitz (in relation to norm $\|\cdot\|_1$), then for $R(w) = \frac{1}{\eta}\sum_{i=1}^{d} w_i \log w_i$ and $S = \{w : \ \|w\|_1 \leq B, w > 0\}$:*

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

Notice that $R(u) \leq \frac{1}{2\eta}\log d/B$, so for the experts problem, By Holder's inequality, we get:

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

Since $x_d \in [0,d]^d$ implies that $\|x\|_\infty \leq 1$.
So $L_t = 1$ with respect to $\|\cdot\|_1$. In addition, $B = 1$ with respect to $\|\cdot\|_1$ Since $u \in S$ implies that $\|u\|_1 = 1$.
Now we can look at the regret:

$$\text{Regret} \leq R(u) - R(w1) + \frac{1}{\sigma}\sum_{t=1}^{T} L_t^2$$

Since $B = 1$, $R(u) \leq \frac{1}{\eta}Log(d)$, $L_t = 1$, and $\sigma = \frac{1}{2\eta}$, We have:

$$\text{Regret} \leq \frac{Log(d)}{2\eta} + \eta T, \text{ (For } \eta = \sqrt{\frac{Log(d)}{2T}})$$
$$\text{Regret} \leq \sqrt{2\log dT}$$

Even though the two bounds seem similar, the meaning of L and B are different.
In OGD, B is a ball over u (the norm L_2), and the Lipschitz constant is with respect to the norm L_2.
In the other one, the B is in L_1 and also is the Lipschitz constant.

In the Experts problem, since each expert is a vector (0,...,0,1,0,...0), It's norm is 1 in both L_1 and in L_2, but the Lipschitz constant of norm L_2 is $\sqrt{d}$ and of norm L_1 is 1.