Computational Learning Theory **Fall Semester, 2017/18**

Lecture 4: November 13

Lecturer: Yishay Mansour *Scribe: Guy Dolinsky, Yogev Bar-On, Yuval Lewi*

4.1 Fenchel-Conjugate

4.1.1 Motivation

Until this lecture we saw our problem in the primal space $(x, f(x))$. In this lecture we will look at the dual space representation of our problem, meaning, looking at $(f(x), \nabla f(x))$. For convex functions, this representation contains all the data we have in the regular problem while giving us a new geometric view of our problem.

Let us define the dual function:$f^*(y) = \max_{w \in S} y^T u - f(w)$

Theorem 4.1 *Assume that* $x = \arg\max_{w \in S} y^T w - f(w)$. *Then* $y \in \partial f(x)$.

Proof: The definition of f^* derives that:

$$\forall u : f^*(y) \geq y^T u - f(u)$$

Hence,

$$\forall u : f(u) \geq y^T u - f^*(y)$$

From our assumption that: $f^*(y) = y^T x - f(x)$, it implies that,

$$\forall u : f(u) \geq y^T u - y^T x + f(x) = f(x) + y^T (u - x),$$

which is the definition of $y \in \partial f(x)$. □

4.1.2 Examples

Example from economics

Assume that a manufacturer produces d products with quantities of $q \in \mathbb{R}_+^d$. Let us also assume the the cost function per quantity is defined by a convex function $C(q)$. Then the revenue is defined by:

$$Rev(p, q) = p^T q - C(q),$$

where $p \in \mathbb{R}^d$ is the price per unit of product.
The dual problem in that case is:

$$C^*(p) = \max_q p^T q - C(q) = \max_q Rev(p, q).$$

Namely, the dual problem, given prices outputs quantities that maximize revenue. In addition, the marginal cost per per product is $\nabla C(q)$ meaning at optimum we have $p = \nabla C(q)$.

A single dimension example

Define $f(w) = w \log w$ where $f : \mathbb{R} \to \mathbb{R}$. Then:

$$f^*(y) = \max_w y^T w - w \log w$$

Hence, by taking the derivative and comparing to 0 we get:

$$y - 1 - logw^* = 0 \Rightarrow w^* = e^{y-1}.$$

Therefore

$$f^*(y) = ye^{y-1} - (y-1)e^{y-1} = e^{y-1}.$$

L-2 distance example

Define $f(w) = \frac{1}{2}\|w\|_2^2$ where $f : \mathbb{R}^d \to \mathbb{R}$. As before:

$$f^*(y) = \max_{w \in S} y^T w - \frac{1}{2}\|w\|_2^2$$

Taking the derivative and zeroing we get $w^* = y$. This implies,

$$f^*(y) = y^T y - \frac{1}{2}\|y\|_2^2 = \frac{1}{2}\|y\|_2^2.$$

4.1.3 Fenchel-young inequality

Theorem 4.2 *Fenchel-young inequality:* $f^*(y) + f(u) \geq y^T u$

Proof: By definition:

$$f^*(y) = \max_{w \in S} y^T w - f(w)$$

Hence:

$$\forall u : f^*(y) \geq y^T u - f(u)$$

Rearranging,

$$\forall u : f^*(y) + f(u) \geq y^T u.$$

□

Theorem 4.3 $f(w) \geq f^{**}(w)$

Proof: Since $f^{**}(w) = (f^*(w))^*$,

$$f^{**}(w) = \max_y y^T w - f^*(y) = \max_y(y^T w - (\max_z z^T y - f(z))).$$

We can set $z = w$ and get,

$$f^{**}(w) \leq max_y(y^T w - w^T y + f(w)) = f(w)$$

□

It is possible to prove that $f^{**} = f$ if the epigraph of $f$is a close and convex group. As the graph of f is $\{(x, f(x))|x \in S\}$ and the epigraph is $\{(x, t)|x \in S, f(x) \leq t\}$.

Lemma 4.4 *If the epigraph of f is close and convex then:*

$$y \in \partial f(x) \Leftrightarrow x = arg\max_z(y^T z - f(z)) \Leftrightarrow x \in \partial f^*(y)$$

Proof:

1. Let us assume $x = arg\max_z(y^T z - f(z))$. Hence:

   $$f^*(y) = y^T x - f(x)$$

   From that we conclude:

   $$f^*(w) - f^*(y) \geq ((w^T x - f(x)) - (y^T x - f(x))) = x^T(w - y)$$

   Thus $x \in \partial f^*(y)$.

2. Let $x \in \partial f^*(y)$, Then

   $$\begin{gathered} x \in \partial f^*(y) \Leftrightarrow \\ \forall w : f^*(w) - f^*(y) \geq x^T(w - y) \Leftrightarrow \\ \forall w : x^T y - f^*(y) \geq f^*(w) - x^T w \end{gathered}$$

   As a result,

   $$y = arg\max_w(x^T w - f^*(w))$$

Combining this with the definition of f^{**}, this leads us to

$$f^{**}(x) = x^T y - f^*(y)$$

Recall $f = f^{**}$ (since the epigraph of f is closed and convex)

$$f^*(y) = x^T y - f(x) = \max_z(z^T y - f(z))$$

Hence

$$x = arg\max_z(z^T y - f(z)).$$

3. We would like to show $y \in \partial f(x) \Leftrightarrow x = arg\max_z(y^T z - f(z))$

$$y \in \partial f(x) \Leftrightarrow \forall z \in S : f(z) - f(x) \geq y^T(z - x) \Leftrightarrow$$
$$\forall z \in S : y^T x - f(x) \geq y^T z - f(z) \Leftrightarrow x = arg\max_z(y^T z - f(z)).$$

□

Lemma 4.5 *If f is differentiable then $y = \nabla f^*(\nabla f(y))$.*

Proof: Set z such that $f(y) + f^*(z) = z^T y$. Then

$$\nabla_y f(y) = \nabla_y(z^t y - f^*(z)) = z$$
$$\nabla_z f^*(z) = \nabla_z(z^t y - f(y)) = y$$

Therefore $\nabla f^*(\nabla f(y)) = \nabla f^*(z) = y$. □

Theorem 4.6 *If $f = f^{**}$ Then $x = \nabla f^*(\nabla f(x))$.*

Proof: Recall the definition of f^* that $f^*(y) = max_z z^T y - f(z)$.
If x maximizes the function then $y = \nabla f(x)$. Hence,

$$f^*(y) = x^T y - f(x)$$

Since $f^{**} = f$

$$f^*(y) + f^{**}(x) = x^T y$$

Therefore

$$f^{**}(x) = max_z z^T y - f^*(z)$$

Meaning $\nabla f^*(y) = x$. Combining $y = \nabla f(x)$ and $x = \nabla f^*(y)$ lead us to $x = \nabla f^*(y) = \nabla f^*(\nabla f(x))$. □

4.2 Bergman-Divergence

4.2.1 Bergman-Divergence of a Convex Function

Let R be some convex function on a set S. We would like to use its dual space for an algorithm we will present later this lecture (the Mirror-Decent Algorithm). We can go from a point in S to a point in the dual space by using ∇R, but we cannot always use ∇R^* to go back - the resulting point is not necessarily in S. To fix this we will use the Bergman-Divergence:

Definition *Bergman-Divergence* of a convex function R is defined as:

$$B_R(x||y) = R(x) - R(y) - [\nabla R(y)]^T (x - y)$$

We can now use $\text{argmin}_{x \in S} B_R(x||y)$ as the *projection* of point y on S.

4.2.2 Examples

L^2-Norm

Let $R(w) = \frac{1}{2}||w||_2^2$. So $\nabla R(w) = w$ and we obtain:

$$B_R(x||y) = \tfrac{1}{2}||x||_2^2 - \tfrac{1}{2}||y||_2^2 - y^T(x - y) = \tfrac{1}{2}||x||_2^2 - \tfrac{1}{2}||y||_2^2 - y^T x + ||y||_2^2 = \tfrac{1}{2}||x||_2^2 + \tfrac{1}{2}||y||_2^2 - y^T x = \tfrac{1}{2}||x - y||_2^2$$

Negative Entropy

Let $R(w) = \sum_i w_i \log w_i$. We obtain $\nabla R(w) = (..., \log(w_i) + 1, ...)^T$, so:

$$B_R(x||y) = \sum_i x_i \log x_i - \sum_i y_i \log y_i - \sum_i ((\log(y_i) + 1)(x_i - y_i)) = \sum_i (x_i (\log(x_i) - \log(y_i))) - \sum_i x_i + \sum_i y_i = \sum_i x_i \log \tfrac{x_i}{y_i} - \sum_i x_i + \sum_i y_i$$

If $S = \{w | w_i \geq 0, ||w||_1 = 1\}$ is the simplex, i.e all distributions, we obtain that $B_R(x||y)$ is the KL-divergence of $x, y \in S$. Also, in this case we obtain that the projection on S is:

$$\text{argmin}_{x \in S} B_R(x||y) = \text{argmin}_{x \in S} \left(\sum_i x_i \log \frac{x_i}{y_i} + \sum_i y_i - 1 \right)$$

We will solve using Lagrange-multipliers:

$$F(x, \lambda) = \sum_i x_i \log \frac{x_i}{y_i} + \sum_i y_i - 1 - \lambda \left(\sum_i x_i - 1 \right)$$

$$\forall i \left(\frac{\partial F}{\partial x_i} = 1 + \log \frac{x_i}{y_i} - \lambda = 0 \right)$$

$$\forall i \left(x_i = y_i e^{(\lambda - 1)} \right)$$

And since $\sum_i x_i = e^{(\lambda-1)} \sum_i y_i = 1$, we obtain that $x_i = \frac{y_i}{||y_i||_1}$. Thus, the projection of y on S is the normalization of y.

4.3 Online Mirror Decent

4.3.1 The Online Mirror Decent Algorithm

The Online Mirror Decent algorithm is an online learning algorithm, similar to the ones we have already seen. The big difference is that OMD uses the dual space to update the current point, instead of the primal space, and projects the update on the primal space with the Bergman-Divergence function. We will present the algorithm with linear loss functions:

ONLINE MIRROR DECENT
begin
Set y_1 s.t. $\nabla R(y_1) = 0$
Set $w_1 = \text{argmin}_{w \in S} B_R(w||y_1)$
for $t \in [1, T]$ **do**
Play w_t and get $f_t(x) = z_t^T x$
Set y_{t+1} s.t $\nabla R(y_{t+1}) = \nabla R(y_t) - \eta \nabla f_t(w_t) = \nabla R(y_t) - \eta z_t$
Namely, $y_{t+1} = \nabla R^{-1}(\nabla R(y_t) - \eta z_t) = \nabla R^*(\nabla R(y_t) - \eta z_t)$
Set $w_{t+1} = \text{argmin}_{w \in S} B_R(w||y_{t+1})$
end for
end ONLINE MIRROR DECENT

4.3.2 Regret Analysis

Theorem 4.7 *Let R be some σ-strong-convex function. The OMD with linear loss functions outputs the same predictions as FoReL.*

Proof: We will denote by w_t^F and w_t^O the predictions in time t of FoReL and OMD, respectively. First, we notice that in OMD:

$$\nabla R(y_{t+1}) = \nabla R(y_t) - \eta z_t = \ldots = -\eta \sum_{i=1}^{t} z_i$$

In FoReL, as we seen in Lecture 2, the update rule is:

$$w_{t+1}^F = \text{argmin}_{w \in S} \left(\eta \sum_{i=1}^{t} z_i^T w + R(w) \right)$$

Hence,

$$\nabla \left(\eta \sum_{i=1}^{t} z_i^T w + R(w) \right) (w_{t+1}^F) = 0$$

and we obtain:

$$\eta \sum_{i=1}^{t} z_i^T + \nabla R(w_{t+1}^F) = 0$$

Therefore,

$$\nabla R(w_{t+1}^F) = -\eta \sum_{i=1}^{t} z_i^T = \nabla R(y_{t+1})$$

Since R is a σ-strong-convex function, we obtain that $w_{t+1}^F = y_{t+1}$. If $y_{t+1} \in S$, we also have $y_{t+1} = w_{t+1}^O$ and we are done. Otherwise, we obtain that:

$$w_{t+1}^O = \text{argmin}_{w \in S} B_R(w||y_{t+1}) = \text{argmin}_{w \in S} \Big(R(w) - R(y_{t+1}) - \left[\nabla R(y_{t+1}) \right]^T (w - y_{t+1}) \Big)$$

$$= \text{argmin}_{w \in S} \Big(R(w) - \left[\nabla R(y_{t+1}) \right]^T w \Big) = \text{argmin}_{w \in S} \left(R(w) + \eta \sum_{i=1}^{t} z_i w \right)$$

Which is again the same as w_{t+1}^F. □

4.4 Exponentiated Gradient Algorithm

4.4.1 The Exponentiated Gradient Algorithm

We can now proceed to retrieve the Randomized Weighted Majority algorithm (the Exponentiated Gradient Algorithm in this context) from the Online Mirror Descent Algorithm.

Regularization Analysis

Setting the regularization function to $R(w) = \sum_{i=1}^{d} w_i \log(w_i)$, we have that $\nabla R(w) = (\ldots, \log(w_i) + 1, \ldots)^T$. Now solving for $\max_w R(w)$ s.t. $\sum_i w_i = 1$ using the Lagrangian $F(w, \lambda) = R(w) - \lambda(\sum_i w_i - 1)$ yields that: $\frac{\partial}{\partial w_i} F = 1 + \log w_i - \lambda \Rightarrow \log w_i = \log w_j = \lambda - 1 \Rightarrow w_i = \frac{1}{d}$. We therefore conclude that $R(w) \leq \log d$.

An Online Mirror Descent Step

The OMD step is defined as $\nabla R(y_{t+1}) = \nabla R(y_t) - \eta \ell_t$, meaning that: $1 + \log y_{t+1}^{(i)} = 1 + \log y_t^{(i)} - \eta \ell_t^{(i)} \Rightarrow y_{t+1}^{(i)} = y_t^{(i)} e^{-\eta \ell_t^{(i)}}$, which is both the definition of RWM and coincides with FoReL.

Exponentiated Gradient Algorithm
begin
Set $y_1 = 1$
Set $w_1 = \frac{y_1}{|y_1|_1} = \frac{1}{d}$
for $t \in [1, T]$ **do**
Play w_t and get ℓ_t
Set $y_{t+1}^{(i)} = y_t^{(i)} e^{-\eta \ell_t^{(i)}}$
Set $w_{t+1}^{(i)} = \frac{y_t^{(i)}}{\sum_{j=1}^{d} y_t^{(j)}}$
end for
end Exponentiated Gradient Algorithm

4.4.2 Regret Analysis

Lemma 4.8 *The regret of this algorithm is bounded a follows:*

$$\forall_u \sum_{t=1}^{T} (w_t - u)^T z_i \leq R(u) - R(w_1) + \sum_{t=1}^{T} B_{R^*}(-Z_{1;t} || - Z_{1;t-1}),$$

Where $Z_{1;t} = \eta \sum_{i=1}^{t} z_i$ *and equality holds for* $u = arg\min_u(R(u) - \sum_i u^T z_i)$

Proof: By the Fenchel-Young inequality,

$$R(u) + R^*(-Z_{1;T}) \geq u^T(-Z_{1;T}) \Rightarrow R(u) + u^T(Z_{1;T}) \geq -R^*(-Z_{1;T}),$$

and from the definition of OGD:

$$w_t = \nabla R^*(-Z_{1;t-1}).$$

Now using a telescopic series, we get that:

$$-R^*(-Z_{1;T}) = -R^*(0) - \left(\sum_{t=1}^{T} \left(R^*(Z_{1;t}) - R^*(Z_{1;t-1})\right)\right)$$

Which by adding and subtracting $\nabla R^*(-Z_{1;t-1})^T z_t$ to the sum and then splitting it, equals:

$$-R^*(0) + \sum_{t=1}^{T} \nabla R^*(-Z_{1;t-1})^T z_t - \sum_{t=1}^{T} \left(R^*(Z_{1;t}) - R^*(Z_{1;t-1}) - \nabla R^*(-Z_{1;t-1})^T(-z_t)\right) =$$
$$-R^*(0) + \sum_{t=1}^{T} w_t^T z_t - \sum_{t=1}^{T} B_{R^*}(-Z_{1;t} || - Z_{1;t-1})$$

Combining everything, yields the following inequalities:

$$R(u) + u^T(Z_{1;T}) \geq -R^*(0) + \sum_{t=1}^T w_t^T z_t - \sum_{t=1}^T B_{R^*}(-Z_{1;t}|| - Z_{1;t-1}) \Rightarrow$$
$$R(u) + R^*(0) + \sum_{t=1}^T B_{R^*}(-Z_{1;t}|| - Z_{1;t-1}) \geq \sum_{t=1}^T (w_t - u)^T z_t$$

Finally, we evaluate $R^*(0)$:

$$R^*(0) = \max_w 0^T w - R(w) = \max_w -R(w) = -\min_w R(w) = -R(w_1)$$

□

Theorem 4.9 *For the Normalized Exponentiated Gradient Algorithm, the regret is:*

$$regret = \sum_{t=1}^T (w_t^T - u)^T z_t \leq \frac{\log d}{\eta} + \eta \sum_{t=1}^T \sum_i w_t^{(i)} {z_t^{(i)}}^2$$

Proof: From the previous lemma, and since $R^*(\theta) = \frac{1}{\eta}\log(\sum_{i=1}^d e^{-\eta\theta_i})$, it suffices to show that $B_{R^*}(-Z_{1;t}|| - Z_{1;t-1}) \leq \eta \sum_i w_t^{(i)} z_t^{2(i)}$. We evaluate $B_{R^*}(-Z_{1;t}|| - Z_{1;t-1})$ to get:

$$B_{R^*}(-Z_{1;t}|| - Z_{1;t-1}) = R^*(-Z_{1;t}) - R^*(-Z_{1;t-1}) + w_t^T z_t = \frac{1}{\eta}\log\frac{\sum_i e^{-\eta Z_{1;t}^{(i)}}}{\sum_j e^{-\eta Z_{1;t-1}^{(j)}}} + w_t^T z_t$$

However, $e^{-\eta Z_{1;t}} = e^{-\eta Z_{1;t-1}} e^{-\eta z_t}$, and therefore:

$$B_{R^*}(-Z_{1;t}|| - Z_{1;t-1}) = \frac{1}{\eta}\log(\sum_i w_t^{(i)} e^{-\eta z_t^{(i)}}) + w_t^T z_t.$$

Since $e^{-a} \leq 1 - a + a^2$ for $-1 \leq a$, that value is bounded from above by:

$$\frac{1}{\eta}\log(\sum_i w_t^{(i)} \left(1 - \eta z_t^{(i)} + \eta^2 {z_t^{(i)}}^2\right)) + w_t^T z_t = \frac{1}{\eta}\log\left(1 - \eta w_t^T z_t + \eta^2 \sum_i w_t^{(i)} {z_t^{(i)}}^2\right) + w_t^T z_t,$$

Since $e^{-a} \geq 1 - a$ we have $\log(1 - a) \leq -a$, and bound from above by:

$$B_{R^*}(-Z_{1;t}|| - Z_{1;t-1}) \leq \frac{1}{\eta}(-\eta w_t^T z_t + \eta^2 \sum_i w_t^{(i)} {z_t^{(i)}}^2) + w_t^T z_t = \eta \sum_i w_t^{(i)} {z_t^{(i)}}^2.$$

□