Homework 3 - Solutions

1. Given, there exists an algorithm for OCO, which for a sequence of loss functions $f_t \in \mathcal{L}$ gives a regret of $R(T) = o(T)$. Let us feed the function f continuously into this algorithm at every time t. Then, $R(T) = \sum_{t=1}^{T}(f(x_t) - \inf_{x\in\mathcal{K}} f(x))$. As $R(T) = o(T)$, we have $\lim_{n\to\infty} \frac{R(T)}{T} = 0$, i.e., for every $\epsilon > 0$, $\exists N_\epsilon$ s.t. $\frac{R(T)}{T} \leq \epsilon$ for every $T \geq N_\epsilon$. So for any $T \geq N_\epsilon$, we have,

$$\frac{R(T)}{T} = \frac{1}{T}\left(\sum_{t=1}^{T}(f(x_t) - \inf_{x\in\mathcal{K}} f(x))\right) \geq f\left(\sum_{t=1}^{T}(x_t/T)\right) - \inf_{x\in\mathcal{K}} f(x)$$

where the inequality follows since f is convex within $\mathcal{K}$. Thus choosing $x^* = \sum_{t=1}^{T}(x_t/T)$ at $T = N_\epsilon$, gives

$$f(x^*) \leq \inf_{x\in\mathcal{K}} f(x) + \frac{R(T)}{T} \leq \inf_{x\in\mathcal{K}} f(x) + \epsilon.$$

2. Thanks to the hint, we know that executing the FTRL algorithm, and executing (unconstrained minimization + Bregman Projection), are one and the same. So to minimize $\sum_{i=1}^{t-1} f_i(x) + R_\eta(x)$ over $x \in \mathbb{R}_n$, let us differentiate it and equate it to zero.

$$\frac{\partial}{\partial x_i}(< z_{1:t-1}, x > + \frac{1}{\eta}\sum_{i=1}^{N} x_i \log x_i) = 0 \Rightarrow (z_i)_{1:t-1} + \frac{1}{\eta}(1+\log x_i) = 0 \Rightarrow x_i = \exp\left(-\eta(z_i)_{1:t-1} - 1\right).$$

Now thanks again to the hint, we know that the Bregman projection of x_i's onto Δ_N is equivalent to scaling the x_i's by its L_1 norm. Thus we have

$$x_i^*(t) = \frac{\exp\left(-\eta(z_i)_{1:t-1} - 1\right)}{\sum_{j=1}^{N}\exp\left(-\eta(z_j)_{1:t-1} - 1\right)} = \frac{\exp\left(-\eta(z_i)_{1:t-1}\right)}{\sum_{j=1}^{N}\exp\left(-\eta(z_j)_{1:t-1}\right)}.$$

Running the exponential weights algorithm with N experts with losses z_t, we obtain $x_{i,t} = \exp\left(-\eta\sum_{s=1}^{t-1}(z_i)_s\right)$. These weights can be normalized, since proportional weights give rise to the same prediction. So when we normalize $x_{i,t}$ by its L_1 norm, we have

$$x_{i,t} = \frac{\exp\left(-\eta(z_i)_{1:t-1}\right)}{\sum_{j=1}^{N}\exp\left(-\eta(z_j)_{1:t-1}\right)}.$$

Note that $x_i^*(t)$ and $x_{i,t}$ are the same. Hence the result.

Aside: To prove that the Bregman projection of $y \geq 0$ w.r.t. to R onto Δ_N is just scaling it by its L_1 norm, let us consider the optimization problem of minimizing $D_R(x, y) = R(x) - R(y) - \nabla R(y)^T(x - y)$ subject to $\sum_{i=1}^{N} x_i = 1$. The Lagrangian $\mathcal{L}$ is given by

$$\mathcal{L} = \sum_{i=1}^{N}(x_i \log x_i - y_i \log y_i - (1 + \log y_i)(x_i - y_i) + \lambda x_i) - \lambda = \sum_{i=1}^{N}(x_i \log\frac{x_i}{y_i} + (\lambda - 1)x_i + y_i) - \lambda.$$

Differentiating it partially w.r.t. x_i and equating it to 0, we get $\lambda = \log\frac{y_i}{x_i}$ for every i, implying $x = ay$, $a > 0$ being a constant. But we know that $\sum_{i=1}^{N} x_i = 1$. Thus $x_i = \frac{y_i}{\sum_{i=1}^{N} y_i}$.

3. We know that $D_R(x, y) = R(x) - R(y) - \nabla R(y)^T(x - y)$.

(a)

$$\begin{aligned}D_R(u,v)+D_R(v,w)-D_R(u,w) &= [R(u)-R(v)-\nabla R(v)^T(u-v)] \\ &+ [R(v)-R(w)-\nabla R(w)^T(v-w)]-[R(u)-R(w)-\nabla R(w)^T(u-w)] \\ &= -\nabla R(v)^T(u-v)-\nabla R(w)^T(v-u) = (\nabla R(w)-\nabla R(v))^T(u-v).\end{aligned}$$

(b)

$$\begin{aligned}\nabla_x D_R(x,y) &= \nabla_x(R(x)-R(y)-\nabla R(y)^T(x-y)) \\ &= \nabla R(x) - 0 - \nabla_x\left(\sum_{i=1}^d (x_i-y_i)\frac{\partial}{\partial y_i}R(y)\right) = \nabla R(x)-\nabla R(y).\end{aligned}$$

4. Fenchel dual $h(\theta)$ of $F(x)$ is defined to be $h(\theta) = \sup_{x\in\mathbb{R}^d}(< x,\theta > -F(x))$. Let us find the stationary point that maximizes $< x,\theta > -F(x)$ by differentiating it and equating the differential to zero. Note that as the given functions are convex in $\mathbb{R}^d$, the Hessian $-\nabla^2 F(x)$ is always negative semidefinite, and thus no more checking is required.

   (a) $\theta - \nabla F(x) = 0$ implies $x_i = \log\theta_i$ for any $\theta_i \geq 0$. Thus $h(\theta) = \sum_{i=1}^d \theta_i(\log\theta_i - 1)$, for every $\theta \geq 0$. The case $\theta_i = 0$ works since we consider $0\log 0 = 0$. If there exists an i s.t. $\theta_i < 0$, then the corresponding x_i can be set to $-\infty$, thus making $h(\theta) = \infty$. Thus we have

$$h(\theta) = \begin{cases} \sum_{i=1}^n \theta_i(\log\theta_i - 1) & \text{if } \theta \geq 0; \\ \infty & \text{otherwise.}\end{cases}$$

   (b) $\theta - \nabla F(x) = 0$ implies $\theta_i = \frac{e^{x_i}}{\sum_{j=1}^d e^{x_j}}$. This can be satisfied for every i only when $\sum_{i=1}^d \theta_i = 1, \theta_i \geq 0$. In such a case, we have $x_i = \log\theta_i$, and thus $h(\theta) = \sum_{i=1}^d \theta_i\log\theta_i = -H(\theta), \forall\theta\in\Delta_d$, where H refers to the entropy function. If there exists an i s.t. $\theta_i < 0$, then the corresponding x_i can be set to $-\infty$, thus making $h(\theta) = \infty$. If $\theta \geq 0$ but $\|\theta\|_1 \neq 1$, then let $x_i = \lambda$ for every i. We have $< x,\theta > -F(x) = \lambda(\|\theta\|_1 - 1) - \log d$. This means that we can drive λ either to ∞ or $-\infty$ appropriately for any $\theta \geq 0$ s.t.$\|\theta\| \neq 1$, thus making $h(\theta) = \infty$. Now we have

$$h(\theta) = \begin{cases} -H(\theta) & \text{if } \theta\in\Delta_d; \\ \infty & \text{otherwise.}\end{cases}$$

   (c) $\theta - \nabla F(x) = 0$ implies $\theta_i = \mathrm{sgn}(x_i) \mid x_i \mid^{p-1}\left(\sum_{i=1}^d \mid x_i \mid^p\right)^{(2/p)-1}, p\in(1,\infty)$. This means that

$$\begin{aligned}\|\theta\|_{\frac{p}{p-1}} = \left(\sum_{i=1}^d \mid\theta_i\mid^{\frac{p}{p-1}}\right)^{(p-1)/p} &= \left[\left(\sum_{i=1}^d \mid x_i\mid^p\right)^{(2-p)/(p-1)}\cdot\left(\sum_{i=1}^d \mid x_i\mid^p\right)\right]^{(p-1)/p} \\ &= \left(\sum_{i=1}^d \mid x_i\mid^p\right)^{1/p} = \|x\|_p.\end{aligned}$$

   Thus to maximize $< x,\theta > -F(x)$, x is chosen s.t. $\|x\|_p = \|\theta\|_q$, where $q = \frac{p}{p-1}$, or $\frac{1}{p}+\frac{1}{q} = 1$. More explicitly, $x_i = \frac{\mathrm{sgn}(\theta_i)|\theta_i|^{\frac{1}{p-1}}}{\left(\sum_{i=1}^d|\theta_i|^q\right)^{\frac{2-p}{q(p-1)}}}$. So we have

$$h(\theta) =< x,\theta > -F(x) = \frac{\left(\sum_{i=1}^d \mid\theta_i\mid^{\frac{p}{p-1}}\right)}{\left(\sum_{i=1}^d \mid\theta_i\mid^q\right)^{\frac{2-p}{q(p-1)}}} - \frac{1}{2}\|\theta\|_q^2 = \left(\sum_{i=1}^d \mid\theta_i\mid^q\right)^{\frac{qp-q-2+p}{q(p-1)}} - \frac{1}{2}\|\theta\|_q^2 = \frac{1}{2}\|\theta\|_q^2$$

   where pq was substituted by $p+q$ in the last equality. Hence, $h(\theta) = \frac{1}{2}\|\theta\|_q^2, \forall\theta\in\mathbb{R}^d$, if $p\in(1,\infty)$.

For $p = 1$, $\theta - \nabla F(x) = 0$ implies $\theta_i = \mathrm{sgn}(x_i)\|x\|_1$, but this can hold only if $\mid \theta_i \mid$ is same for every i. If not, we would get infeasible conditions. In that case, we look at the boundary conditions, where $x_i = 0$ for a few components. If we choose $x_i = 0$ for any i having $\theta_i < \max_i \mid \theta_i \mid$, then we get feasible conditions to satisfy. We can then derive $x_i = \mathrm{sgn}(\theta_i)\frac{|\theta_i|\mathbf{1}(|\theta_i|=\max_i|\theta_i|)}{\#|\theta_i|=\max_i|\theta_i|}$, and thus have $h(\theta) = \max_i \theta_i^2 = \|\theta\|_\infty^2$.
For $p = \infty$, $\theta - \nabla F(x) = 0$ implies $\theta_i = \mathrm{sgn}(x_i)\max_i \mid x_i \mid$, but this can hold only if $\mid \theta_i \mid$ is same for every i. If not, we would get infeasible conditions. In that case, we look at the boundary conditions, where $\mid x_i \mid$ are equal for a few components. If we choose all $\mid x_i \mid$'s to be equal, then we get feasible conditions to satisfy. We can then derive $x_i = \mathrm{sgn}(\theta_i)\|\theta\|_1$, and thus have $h(\theta) = \frac{1}{2}\|\theta\|_1^2$.
Hence, $h(\theta) = \frac{1}{2}\|\theta\|_q^2, \forall\theta \in \mathbb{R}^d$, $q = \frac{p}{p-1}$ for any $p \in [1, \infty]$.