Homework 2 - Solutions

1. (a) Given, $F(z) = e^{-\eta l(z,y)}$ is concave in z. Then for $z = \lambda z_1 + (1-\lambda)z_2$, we have

$$e^{-\eta l(z,y)} \geq \lambda e^{-\eta l(z_1,y)} + (1-\lambda)e^{-\eta l(z_2,y)}$$

Taking log on both sides, we have

$$\begin{aligned} l(z,y) &\leq -\frac{1}{\eta} \log\left(\lambda e^{-\eta l(z_1,y)} + (1-\lambda)e^{-\eta l(z_2,y)}\right) \\ &\leq -\frac{1}{\eta}(\lambda \log(e^{-\eta l(z_1,y)}) + (1-\lambda)\log(e^{-\eta l(z_2,y)})) = \lambda l(z_1,y) + (1-\lambda)l(z_2,y) \end{aligned}$$

where the second inequality follows by Jensen's inequality applied on $\log x$ which is concave in x. Hence the result.

(b) We need to show that $F(x,y) = e^{-l(x,y)} = \left(\frac{x}{y}\right)^y \left(\frac{1-x}{1-y}\right)^{1-y}$ is concave in x for every $y \in [0,1]$. We will first show that $F''(x) \leq 0, \forall y \in (0,1)$.

$$\begin{aligned} F''(x,y) &= \left(\frac{yx^{y-1}(1-x)^{1-y} - (1-y)x^y(1-x)^{-y}}{y^y(1-y)^{(1-y)}}\right)' \\ &= \frac{-y(1-y)}{y^y(1-y)^{(1-y)}}(x^{y-2}(1-x)^{1-y} + 2x^{y-1}(1-x)^{-y} + x^y(1-x)^{-y-1}) \end{aligned}$$

which clearly is non-positive for every $y \in (0,1)$. We have $F(x,0) = 1-x$, and $F(x,1) = x$, which clearly are concave in x. Hence the result.

(c) We need to show that $F(x,y) = e^{-\frac{(x-y)^2}{2}}$ is concave in x for every $y \in [0,1]$.

$$F''(x,y) = \left(-(x-y)e^{-\frac{(x-y)^2}{2}}\right)' = \left((-1+(x-y)^2)e^{-\frac{(x-y)^2}{2}}\right)$$

which is non-positive for any $x, y \in [0,1]$.

(d) Fix $y > 0$. Then, $F(x,y) = e^{-\eta|x-y|} = e^{\eta(x-y)}$ for any $x \in [0,y]$. But this function is strictly convex in x. Also, $F(x,0) = e^{-\eta x}$, which is strictly convex in x. Thus for any $y \in [0,1]$, the function $\mid x - y \mid$ can never be η-exp-concave for any $\eta > 0$.

2. (a) With advance information, the problem is equivalent to choosing the right expert from a total of N^T experts, since choosing one out of N experts at each time instant means choosing one out of N^T experts in all. The minimax regret, as proved in class, is equal to $\sqrt{\frac{T}{2}\log N}$ when we have N experts. In the presence of N^T experts, the minimax regret happens to be $\sqrt{\frac{T}{2}\log N^T}$ which is equal to $T\sqrt{\frac{\log N}{2}}$, which means that the regret is linear in T.

(b) The number of experts in the set $\mathcal{I}_m$ is given by

$$\begin{aligned} \mid \mathcal{I}_m \mid &= (\text{\# ways to choose the switch time}) \times (\text{\# ways to choose the order of experts}) \\ &\leq \binom{T}{m} \times N^{m+1} \leq \left(\frac{Te}{m}\right)^m N^{m+1} \end{aligned}$$

The minimax regret in the presence of $\mid \mathcal{I}_m \mid$ experts is given by,

$$\sqrt{\frac{T}{2}\log \mid \mathcal{I}_m \mid} = \sqrt{\frac{T}{2}\left((m+1)\log N + m\left(1 + \log\frac{T}{m}\right)\right)}$$

which can be written as $O\left(\sqrt{\frac{T}{2}\left((m+1)\log N + m\log\frac{T}{m}\right)}\right)$. But the minimax regret can be achieved through the exponential weights algorithm. Thus the EXP-WTS algorithm performed on all the experts in $\mathcal{I}_m$ achieves the given regret in the problem.

3. Let $\mathcal{A} = \{w \in \mathbb{R}_+^N : \|w\|_1 \leq B\}$. We need to show that $R(w) = \sum_{i=1}^N w_i \log w_i$ is $\frac{1}{B}$-strongly convex in $\mathcal{A}$, i.e., $x^T(\nabla^2 R(w))x \geq \frac{1}{B}\|x\|_1^2$ for every $x, w \in \mathcal{A}$. But $\nabla^2 R(w)$ is a diagonal matrix with its ith diagonal entry equal to $\frac{1}{w_i}$. Thus it remains to prove $\sum_{i=1}^N \frac{x_i^2}{w_i} \geq \frac{1}{B}\|x\|_1^2$ for every $x, w \in \mathcal{A}$.

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

where the first equation just involves multiplying and dividing by $\|w\|_1$, the first inequality follows from Cauchy-Schwarz, and the last inequality follows since $w \in \mathcal{A}$. Hence the result.

4. Let us consider $\mathcal{Y} = \{1, 2, \ldots, m\}$, and in the sequence y^T, let us consider that the alphabet 1 appears n_1 times, 2 appears n_2 times, $\ldots$ and m appears n_m times. Considering log loss, the best expert $f \in \mathcal{F}$ minimizes $-\log \Pi_{i=1}^m f(i)^{n_i}$, or in other words, maximizes $\Pi_{i=1}^m f(i)^{n_i}$. To find the best expert, we need to solve the following optimization problem:

$$\max_{f\in\mathcal{F}} \Pi_{i=1}^m f(i)^{n_i} \text{ s.t. } \sum_{i=1}^m f(i) = 1, f(i) \geq 0, i = 1, 2, \ldots, m.$$

We solve this problem using the Lagrangian method. Let $\mathcal{L}(f, \lambda, \mu) = -\Pi_{i=1}^m f(i)^{n_i} + \lambda\sum_{i=1}^m f(i) - \sum_{i=1}^m \mu_i f(i)$. Finding $\frac{\partial \mathcal{L}}{\partial f(i)}$ and equating it to zero, we have

$$-\frac{n_i}{f(i)}\Pi_{i=1}^m f(i)^{n_i} + \lambda - \mu_i = 0, \forall i = 1, 2, \ldots, m \tag{1}$$

By KKT conditions, we must also have $\mu_i f(i) = 0, \mu_i \geq 0$ for every i. Let us define $X := \Pi_{i=1}^m f(i)^{n_i}$. Forcing $\mu_i = 0$ for every i, we have $\sum_{i=1}^m f(i) = 1$ implying $\frac{X}{\lambda}\sum_{i=1}^m n_i = 1$. But $\sum_{i=1}^m n_i = T$. Thus $\lambda = XT$. By (1), we must have $\frac{n_i}{f(i)}X = XT$, or $f(i) = \frac{n_i}{T}$. So the best expert happens to be the empirical distribution of the received sequence. The cumulative loss of the best expert, $L(f)$, is given by

$$L(f) = -\log\left(\Pi_{i=1}^m f(i)^{n_i}\right) = -\sum_{i=1}^m n_i \log\left(\frac{n_i}{T}\right) = TH(f).$$

5. (a) The quantity Δ_m represents an $(m-1)$-dimensional unit simplex with vertices $e_1, e_2, \ldots, e_m \in \mathbb{R}^m$, where e_i is an m-dimensional vector with ith element being 1, and all the other elements being 0. We shall prove that $\text{Vol}(\Delta_m) = \frac{\sqrt{m}}{(m-1)!}$ through induction. For $m = 2$, the length between $(1, 0)$ and $(0, 1)$ clearly is $\sqrt{2}$. Let $\text{Vol}(\Delta_k) = \frac{\sqrt{k}}{(k-1)!}$. Now, Δ_{k+1} represents a k-dimensional unit simplex similar to a triangle with Δ_k as its base, and e_{k+1} as the remaining vertex. Then we have

$$\text{Vol}(\Delta_{k+1}) = \frac{1}{k}\text{Vol}(\Delta_k) \times \text{Perpendicular height} = \frac{1}{k}\frac{\sqrt{k}}{(k-1)!}\sqrt{1 + \frac{k}{k^2}} = \frac{\sqrt{k+1}}{k!}$$

where the perpendicular height is the distance between e_{k+1} and the centroid of Δ_k, which is $(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}, 0) \in \mathbb{R}^{k+1}$.

The set $\text{Ball}_\epsilon(b^*)$ represents an $(m-1)$-dimensional simplex with vertices $(1-\epsilon)b^* + \epsilon e_1, (1-\epsilon)b^* + \epsilon e_2, \ldots, (1-\epsilon)b^* + \epsilon e_m \in \mathbb{R}^m$. Using the same argument as above, we will prove that $\text{Vol}(\text{Ball}_\epsilon(b^*)) = \epsilon^{m-1}\frac{\sqrt{m}}{(m-1)!}$. For $m = 2$, the length between $((1-\epsilon)b_1^* + \epsilon, (1-\epsilon)b_2^*)$ and $((1-\epsilon)b_1^*, (1-\epsilon)b_2^* + \epsilon)$ clearly is $\epsilon\sqrt{2}$, for any $b^* \in \Delta_2$. Let $\text{Vol}(\text{Ball}_\epsilon(b^*)) = \epsilon^{k-1}\frac{\sqrt{k}}{(k-1)!}$ when $m = k$. For $m = k+1$, considering $\hat{b}^* \in \Delta_{k+1}$ to have first k of its co-ordinates equal to $b^* \in \Delta_k$, we have

$$\text{Vol}(\text{Ball}_\epsilon(\hat{b}^*)) = \frac{1}{k}\text{Vol}(\text{Ball}_\epsilon(b^*)) \times \text{Perpendicular ht} = \frac{\epsilon^{k-1}}{k}\frac{\sqrt{k}}{(k-1)!}\sqrt{\epsilon + \frac{\epsilon k}{k^2}} = \epsilon^k\frac{\sqrt{k+1}}{k!}$$

where the perpendicular height is the distance between $(1-\epsilon)\hat{b}^* + \epsilon e_{k+1}$ and the centroid of $\text{Vol}(\text{Ball}_\epsilon(b^*))$, which is $((1-\epsilon)\hat{b}_1^* + \frac{\epsilon}{k}, (1-\epsilon)\hat{b}_2^* + \frac{\epsilon}{k}, \ldots, (1-\epsilon)\hat{b}_k^* + \frac{\epsilon}{k}, (1-\epsilon)\hat{b}_{k+1}^*) \in \mathbb{R}^{k+1}$. Thus we have

$$\frac{\text{Vol}(\text{Ball}_\epsilon(b^*))}{\text{Vol}(\Delta_m)} = \frac{\epsilon^{m-1}\frac{\sqrt{m}}{(m-1)!}}{\frac{\sqrt{m}}{(m-1)!}} = \epsilon^{m-1}.$$

(b) We know that $S_T(b^*, X^T) = \Pi_{t=1}^T \left(\sum_{i=1}^n b_i^* x_{i,t}\right)$. So for any $b \in \text{Ball}_\epsilon(b^*)$, considering $b = (1-\epsilon)b^* + \epsilon b'$ for some $b' \in \Delta_m$, we have

$$\begin{aligned} S_T(b, X^T) = \Pi_{t=1}^T \left(\sum_{i=1}^n [(1-\epsilon)b_i^* + \epsilon b_i']x_{i,t}\right) &\geq \Pi_{t=1}^T \left(\sum_{i=1}^n (1-\epsilon)b_i^* x_{i,t}\right) \\ &= (1-\epsilon)^T \Pi_{t=1}^T \left(\sum_{i=1}^n b_i^* x_{i,t}\right) = (1-\epsilon)^T S_T(b^*, X^T). \end{aligned}$$

6. For the universal portfolio problem, one plays $p_t \in \Delta_m$ and receives a loss of $-\log < p_t, x_t >$, at each time t. It is given that $x_t \in [\epsilon, 1]^m$ at each time t. Using the regret expression for P-OGD as derived in the class, we have $R_T \leq DG\sqrt{T}$, if we choose $\eta = \frac{D}{G\sqrt{T}}$. The value of D and G for this problem has to be found. $D = \max_{p,q\in\Delta_m} \|p-q\|_2 = \sqrt{2}$, and $G = \sup_{t\leq T, p\in\Delta_m} \|\nabla f_t(p)\|_2 = \sup_{t\leq T, p\in\Delta_m} \frac{1}{<p_t, x_t>}\|x_t\|_2 \leq \frac{\sqrt{m}}{\epsilon}$. Thus we have

$$R_T \leq \frac{\sqrt{2mT}}{\epsilon}.$$