**E1 245 - Online Prediction and Learning, Aug-Dec 2014**
**Final Exam**
**December 9, 2014**

---

Instructions:

- There are a total of 5 questions with a maximum score of 50 points. The total time allotted is 3 hours.
- No electronic devices or aids are allowed. You may use notes made on one A4 size sheet of paper for reference.
- Academic dishonesty will not be tolerated.

---

1. *(6 points)* Suppose I give you a finite-state, finite-action Markov Decision Process $(\mathcal{S}, \mathcal{A}, T, R)$ where $\mathcal{S}$ and $\mathcal{A}$ are the set of states and actions respectively, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$ is the transition probability function and $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ denotes the reward function. I present you with a policy $\pi : \mathcal{S} \to \mathcal{A}$ for this MDP, claiming that $\pi$ is optimal under the infinite-horizon discounted reward criterion with a discount factor $\gamma \in (0, 1)$.

   How will you check if my claim is true or false (i.e., give an algorithm that decides if $\pi$ is optimal or not)?

   **Solution.** We can apply one step of policy iteration to decide whether $\pi$ is optimal or not. Given the MDP and the policy $\pi$, find the value function $V_\gamma^\pi : \mathcal{S} \to \mathbb{R}$ by solving the system of linear equations

$$V_\gamma^\pi(s) = \sum_{s' \in \mathcal{S}} T(s, \pi(s), s') \left[ R(s, \pi(s), s') + \gamma V_\gamma^\pi(s') \right], \ s \in \mathcal{S}.$$

   Next, if $\pi(s) \in \arg\max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s, a, s') \left[ R(s, a, s') + \gamma V_\gamma^\pi(s') \right]$ for every state $s \in \mathcal{S}$, then $\pi$ is an optimal discounted-cost policy. If not, $\pi$ can always be strictly improved and thus cannot be optimal.

2. *(6 points)* Consider the problem of sequentially predicting a (fixed and unknown) sequence $y_1, y_2, \ldots$ in $\mathbb{R}^d$. At each round $t = 1, 2, \ldots, T$, the prediction algorithm picks a point $p_t \in \mathbb{R}^d$ (knowing only $y_1, \ldots, y_{t-1}$), the current element $y_t$ of the sequence is revealed, and the algorithm suffers a loss $l(p_t, y_t) = ||p_t - y_t||_2^2$.

   If the sequence $\{y_n\}$ comes from the unit ball $B := \{x \in \mathbb{R}^d : ||x||_2 \leq 1\}$ and we use the Follow-The-Leader (FTL) algorithm[1] to predict from $B$, then we have seen that the regret with respect to points in $B$ is $O(\log T)$ over $T$ rounds. Show that this upper bound is order-wise tight, i.e., for any $T \geq 1$, construct a sequence $y_1, y_2, \ldots, y_T$ from $B$ such that FTL suffers regret $\Omega(\log T)$ with respect to $B$.

   [Hint: Think about $\pm v$ for any unit vector $v$.]

   **Solution.** Let $y_t := (-1)^t v$, $t = 1, 2, \ldots$ for some fixed $v \in B$, $||v||_2 = 1$. FTL picks $p_1 := 0$ and $p_t := \frac{1}{t-1} \sum_{s=1}^{t-1} y_s$ for $t \geq 2$, i.e., the sequence $0, -v, 0, -v, \ldots$. Hence, FTL's cumulative loss is

$$1 + (1+1)^2 + 1 + \left(1 + \frac{1}{3}\right)^2 + 1 + \left(1 + \frac{1}{5}\right)^2 + \ldots$$

[1] Assume that the initial prediction is $p_1 := 0$.

In comparison, the loss of the single point $0 \in B$ over the input sequence $\{y_t\}$ is $1+1+1+\ldots = T$. Hence, the regret of FTL wrt all of $B$ is at least

$$\begin{aligned}
&1 + (1+1)^2 + 1 + \left(1+\frac{1}{3}\right)^2 + 1 + \left(1+\frac{1}{5}\right)^2 + \ldots - T \\
&= \sum_{k=1}^{\lfloor T/2 \rfloor} \left[ \left(1+\frac{1}{2k-1}\right)^2 - 1 \right] \\
&= \sum_{k=1}^{\lfloor T/2 \rfloor} \left[ \frac{2}{2k-1} + \frac{1}{(2k-1)^2} \right] \\
&\geq \sum_{k=1}^{\lfloor T/2 \rfloor} \left[ \frac{2}{2k} + 0 \right] = \sum_{k=1}^{\lfloor T/2 \rfloor} \frac{1}{k} \geq \log \lfloor T/2 \rfloor = \Omega(\log T).
\end{aligned}$$

3. *(8 points)* Consider an online convex optimization problem over $\mathbb{R}^d$ with convex, differentiable losses $l_t : \mathbb{R}^d \to \mathbb{R}$, $t = 1, 2, \ldots$ and a strictly convex, differentiable regularizer $R : \mathbb{R}^d \to \mathbb{R}$. Consider the Follow-The-Regularized-Leader (FTRL) algorithm: Choose $w_1 \in \mathbb{R}^d$ such[2] that $\nabla R(w_1) = 0$. For $t \geq 1$, pick $w_{t+1} := \arg\min_{w \in \mathbb{R}^d} \left[ \sum_{s=1}^{t} \eta l_s(w) + R(w) \right]$. Define $\Phi_0(w) := R(w)$ and $\Phi_t(w) := \Phi_{t-1}(w) + \eta l_t$, $w \in \mathbb{R}^d$, $t \geq 1$. Prove the equivalence

$$w_{t+1} = \arg\min_{w \in \mathbb{R}^d} \left[ \eta l_t(w) + D_{\Phi_{t-1}}(w, w_t) \right], \ t \geq 1, \tag{1}$$

where $D_F$ stands for the Bregman divergence[3] corresponding to $F : \mathbb{R} \to \mathbb{R}$. In other words, FTRL searches for a point that "balances" between minimizing the loss on the most recently observed loss function and staying close (in terms of Bregman "distance") to the previous decision. [Hint: The minimizer of the right hand side in (1) must make its gradient vanish. You may use the fact that a strictly convex, differentiable function $F$ is minimized at $x$ if and only if $\nabla F(x) = 0$.]

**Solution.** Let $v := \arg\min_{w \in \mathbb{R}^d} \left[ \eta l_t(w) + D_{\Phi_{t-1}}(w, w_t) \right]$. We will show that $v = w_{t+1}$ as defined in FTRL. Since $v$ is a minimizer by definition, we must have

$$\eta \nabla l_t(v) + \nabla_x D_{\Phi_{t-1}}(x, w_t)\big|_{x=v} = 0.$$

Since $\nabla_x D_{\Phi_{t-1}}(x, w_t) = \nabla \Phi_{t-1}(x) - \nabla \Phi_{t-1}(w_t)$, this means that

$$\eta \nabla l_t(v) + \nabla \Phi_{t-1}(v) - \nabla \Phi_{t-1}(w_t) = 0 \ \Rightarrow \ \nabla \Phi_t(v) = \nabla \Phi_{t-1}(w_t).$$

By the definition of FTRL, $w_t$ minimizes $\Phi_{t-1}$ over $\mathbb{R}^d$, and thus $\nabla \Phi_{t-1}(w_t) = 0$ which gives $\nabla \Phi_t(v) = 0$. Since $\Phi_t$ is strictly convex and differentiable, $v = w_{t+1}$ and we are done.

4. *(10 points)* **Inventory Control.** The manager of a warehouse for a product factory faces the following (idealized) problem. The warehouse has infinite capacity in terms of units of products. At the beginning of each month, the warehouse is restocked with an additional (integer) number of units as desired by the manager. The additional number of units added can at most be $M$. The total market demand $D$ for the entire month (again in integer units) is random with a (fixed) probability distribution given by $p_j := \mathbb{P}[D = j]$, $j = 0, 1, 2, \ldots$

[2] Assume that this is indeed possible.

[3] $D_F(x, y) := F(x) - F(y) - \langle \nabla F(y), x - y \rangle$.

At the end of the month, up to $D$ units are sold from the warehouse subject to the restriction that the warehouse cannot sell more units than what it holds. Thus, the inventory level (no. of units in the warehouse) for the next month $Y$, the inventory level for the current month $X$, the no. of restocked units in the current month $W$ and the demand $D$ in the current month are related by the equation $Y = \max(0, X + W - D)$. The cost of ordering $u$ fresh units of the product per month into the warehouse is a function $c(u)$, $u = 0, 1, 2, \ldots$ The cost of storing $u$ units in the warehouse at the beginning of each month is $h(u)$, $u = 0, 1, 2, \ldots$ The revenue obtained when $u$ units are sold at the end of a month is captured by the function $f(u)$, $u = 0, 1, 2, \ldots$ The warehouse manager wishes to maximize net expected profit over a whole year.

Describe how you would model the above problem as a (possibly infinite) Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, T, R)$, where $\mathcal{S}$ and $\mathcal{A}$ are the set of states and actions respectively, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$ is the transition probability function and $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ denotes the reward function. Write down clearly each component of the MDP.

**Solution.** $\mathcal{S} := \{0, 1, 2, \ldots\}$, $\mathcal{A} := \{0, 1, 2, \ldots, M\}$,

$$T(s, a, s') := \begin{cases} p_{s+a-s'}, & 0 < s' \leq s + a \\ \sum_{k=s+a}^{\infty} p_k, & s' = 0 \\ 0, & s' > s + a, \end{cases}$$

$$R(s, a, s') := \begin{cases} f(s + a - s') - c(a) - h(s), & 0 \leq s' \leq s + a \\ 0, & s' > s + a. \end{cases}$$

5. *(20 points)* **Bandits with side information.** Consider a stochastic 2-armed bandit where each arm $i$'s reward sequence is generated independently from a Bernoulli distribution with parameter $\mu_i$, $i = 1, 2$. Further, it is known that $\mu_1 \neq \mu_2$ and $\mu_1, \mu_2 \in \{a, b\}$ where $0 < a < b < 1$ are known constants, i.e., the only uncertainty is in the order. Denote $\Delta := b - a$. The aim is to obtain low regret $R_T := Tb - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right]$ where $I_t \in \{1, 2\}$ is the arm played at time $t$.

   We know that running the UCB algorithm gives $O\left(\frac{\log T}{\Delta}\right)$ regret. But this completely ignores the (potentially huge) side information about the arms' rewards which is known beforehand. The problem asks you to analyze a bandit algorithm whose regret does not scale with $T$!

   Consider the following (rather simple) algorithm. In the beginning, play each arm once, i.e., $I_1 = 1$, $I_2 = 2$. At every subsequent time $t \geq 3$, if there exists an arm whose observed empirical mean so far exceeds $(a+b)/2$, then play the arm with the highest empirical mean. Else, play both arms one after another, i.e., $I_t = 1$ followed by $I_{t+1} = 2$.

   (a) *(3 points)* Without loss of generality, let arm 1 be the optimal arm when running the algorithm. Split the set of times when arm 2 is played according to whether its observed empirical mean so far is (i) greater than or (ii) at most $(a + b)/2$.

   (b) *(5 points)* Bound (from above) the sum of probabilities of playing arm 2 at all times when event (i) occurs, using Hoeffding's inequality[4]. [Note: You may find the inequality $e^x \geq 1 + x$ useful.]

[4]Hoeffding's inequality: For iid random variables $X_1, X_2, X_3, \ldots$ bounded in $[0, 1]$ with $\mu := \mathbb{E}[X_1]$, $\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq n(\mu + \epsilon)\right] \leq \exp(-2n\epsilon^2)$. Likewise for the left tail.

(c) *(5 points)* Bound (from above) the sum of probabilities of playing arm 2 at all times when event (ii) occurs by using the definition of the algorithm and relating event (ii) to an event involving the empirical mean of arm 1. Obtain a bound by applying Hoeffding as before.

(d) *(7 points)* Put together the conclusions of the previous parts to derive a regret bound independent of $T$ and depending only on $\Delta$.

**Solution.**

(a) Let $c := (a+b)/2$. We have, for any time $t \geq 1$,

$$\{I_t = 2\} = \{t = 2\} \cup \{\hat{\mu}_{2,T_2(t)} > c, I_t = 2, t \geq 3\} \cup \{\hat{\mu}_{2,T_2(t)} \leq c, I_t = 2, t \geq 3\},$$

where $T_i(t)$ is the number of times arm $i$ has been played up to time $t$, and $\hat{\mu}_{i,k}$ is the observed empirical mean of arm $i$'s rewards when it has been played $k$ times.

(b) We have the bound

$$\sum_{t=1}^{T} \mathbb{P}\left[\hat{\mu}_{2,T_2(t)} > c, I_t = 2, t \geq 3\right] \leq \sum_{k=1}^{T} \mathbb{P}\left[\hat{\mu}_{2,k} > c\right] \overset{(*)}{\leq} \sum_{k=1}^{T} \exp\left(-2k\Delta^2/4\right)$$
$$\leq \sum_{k=1}^{\infty} \exp\left(-k\Delta^2/2\right) = \frac{1}{\exp\left(\Delta^2/2\right) - 1} \leq \frac{2}{\Delta^2}$$

by applying Hoeffding's inequality in $(*)$.

(c) We also have, using the property of the algorithm,

$$\sum_{t=1}^{T} \mathbb{P}\left[\hat{\mu}_{2,T_2(t)} \leq c, I_t = 2, t \geq 3\right] \leq \sum_{t=3}^{T} \mathbb{P}\left[\hat{\mu}_{1,T_1(t-1)} \leq c, I_{t-1} = 1\right]$$
$$\leq \sum_{k=1}^{T} \mathbb{P}\left[\hat{\mu}_{1,k} \leq c\right] \leq \frac{2}{\Delta^2}.$$

(d) Using the conclusions of the previous parts and the definition of regret, the regret is bounded by

$$\sum_{t=1}^{T} \Delta \cdot \mathbb{P}[I_t = 2] = \Delta \cdot 1 + \sum_{t=3}^{T} \Delta \cdot \mathbb{P}[I_t = 2] \leq \Delta + \frac{4}{\Delta},$$

uniformly over all $T$.