ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Foundations of Data Science — Assignment date: Friday, February 3rd, 2023, 9:15 am
Fall 2022 — Due date: Friday, Feburary 3rd, 2023, 12:15 noon

Final Exam – SG0211

This exam is open book. No electronic devices of any kind are allowed. There are 4 problems. Choose the ones you find easiest and collect as many points as possible. We do not necessarily expect you to finish all of them. Good luck!

Name: ______________________________

Problem 1	/ 15
Problem 2	/ 15
Problem 3	/ 12
Problem 4	/ 20
Total	/62

Problem 1 (*Fisher Goes Exponential*)**.** [15 pts]

Let $p_\theta(x)$ denote a family of distributions parameterized by θ. Define the Fisher information as

$$I_\theta = \mathbb{E}_\theta[\nabla_\theta \log p_\theta(X)(\nabla_\theta \log p_\theta(X))^T].$$

(1) [5pts] Let $p_\theta(x) = h(x)e^{\langle\theta,\phi(x)\rangle - A(\theta)}$ be an exponential family. What is the Fisher information in terms of the parameters of the family?

(2) [5pts - 1pt per question] Consider distributions of the form $p_\lambda(x) = \lambda e^{-\lambda x}$, where $\lambda \in \mathbb{R}^+$.

1. Write it in the form of an exponential family.
2. What is $\Theta = \{\theta \in \mathbb{R} : A(\theta) < \infty\}$.
3. Is the family regular?
4. Is it minimal?
5. What is the Fisher information?

(3) [5pts - 1pt per question] Consider distributions of the form $p_p(k) = (1-p)^k p$, where $p \in (0,1)$ and $k \in \mathbb{N}$.

1. Write it in the form of an exponential family.
2. What is $\Theta = \{\theta \in \mathbb{R} : A(\theta) < \infty\}$.
3. Is the family regular?
4. Is it minimal?
5. What is the Fisher information?

Solution:

(1) We know from the notes that the Fisher information can also be written as $-\mathbb{E}_\theta[\nabla^2_\theta \log p_\theta(X)]$. This shows that $I_\theta = \nabla^2_\theta A(\theta)$.

Alternatively, full score also given for showing one of the following equivalent statements: $I_\theta = \mathbb{E}[\phi(x)\phi(x)^\top] - \mathbb{E}[\phi(x)]\mathbb{E}[\phi(x)]^\top$, $I_\theta = \mathrm{Cov}(\phi(x))$, $I_\theta = \mathbb{E}[(\phi(x) - \mathbb{E}[\phi(x)])(\phi(x) - \mathbb{E}[\phi(x)])^\top]$. (Note that rewriting $\mathbb{E}[\phi(x)] = \nabla_\theta A(\theta)$ is also possible)

(2)
1. $p_\Theta(x) = e^{\Theta\phi(x) - \log(1/\Theta)}$ with $h(x) = 1$, $\theta = \lambda$, $\phi(x) = -x$, and $A(\theta) = \log(1/\theta)$,
2. $\Theta = \{\theta > 0\}$
3. The family is regular since the region Θ is open.

4. Yes, the family is minimal.
5. The Fisher information is $\frac{\partial^2 A(\theta)}{\partial \theta^2} = \frac{\partial^2 \log(1/\theta)}{\partial \theta^2} = \frac{1}{\theta^2}$.

(3)

1. $p_\theta(k) = e^{\theta\phi(k) - A(\theta)}$ with $h(k) = 1$, $\theta = \log(1-p)$, $\phi(k) = k$, and $A(p) = \log(1/p)$ so that $p = 1 - e^\theta$ and $A(\theta) = \log(1/(1-e^\theta))$,
2. We have $\Theta = \{\theta < 0\}$.
3. The family is regular, since Θ is not open.
4. Yes, the family is minimal.
5. The Fisher information is $\frac{\partial^2 A(\theta)}{\partial \theta^2} = \frac{\theta^2 \log(1/(1-e^\theta))}{\partial \theta^2} = \frac{e^\theta}{(1-e^\theta)^2} = (1-p)/p^2$.

Problem 2 (*Compression*). [15 pts]

Suppose $\mathcal{P} \in \Pi(\mathcal{X}, \mathcal{Y})$ be a probability distribution on $\mathcal{X} \times \mathcal{Y}$ and (X, Y) be a joint random variable with distribution P_{XY} with marginals P_X and P_Y.

In what follows, assume that **all codes are optimal, prefix-free, and binary**. Optimal here means having smallest possible average length. All logs are to the base 2.

(1) [1 pt] Let $c_X : \mathcal{X} \to \{0,1\}^*$ and $c_Y : \mathcal{Y} \to \{0,1\}^*$ be optimal prefix free codes. What are lower and upper bounds for the expected length of these codes c_X and c_Y?

(2) [1 pt] Let $c_{XY} : \mathcal{X} \times \mathcal{Y} \to \{0,1\}^*$ be an optimal prefix free code. What are lower and upper bounds for the expected length of this code?

(3) [10 pts total] In this sub problem, assume that X, Y have a joint distribution according to the following table:

	Y=0	Y=1
X=0	$1/4$	0
X=1	$1/8$	$1/8$
X=2	$1/8$	$1/8$
X=3	0	$1/4$

(a) [4 pts] What are lower and upper bounds for the expected lengths of c_X and c_Y? Are the lower bounds tight?

(b) [3 pts] What are lower and upper bounds for the expected lengths of c_{XY} ? Is the lower bound tight?

(c) [3 pts] For the above joint distribution, is it more efficient to compress separately and concatenate the individual code words (which, as we saw in the lecture, is guaranteed to yield a prefix free code), or to compress (X, Y) jointly (again, in a prefix free manner)?

(4) [3 pts] Assume that (X, Y) has some generic joint distribution. Assume further that $I(X; Y) > 1$. Show that in this case optimal joint prefix free compression is more efficient than compressing individually and concatenating.

Solution 1. *(1)*

$$H(X) \leq \mathbb{E}[\text{length}(c_X(X))] \leq H(X) + 1 \tag{1}$$

$$H(Y) \leq \mathbb{E}[\text{length}(c_Y(Y))] \leq H(Y) + 1 \tag{2}$$

(2)

$$H(X,Y) \le \mathbb{E}[\text{length}(c_{XY}(X,Y))] \le H(X,Y) + 1 \tag{3}$$

(3)

(a) We calculate $H(X) = 2, H(Y) = 1$. Therefore,

$$2 \le \mathbb{E}[\text{length}(c_X(X))] \le 3 \tag{4}$$

$$1 \le \mathbb{E}[\text{length}(c_Y(Y))] \le 2 \tag{5}$$

Considering (for example) the following code $c_X(0) = 00, c_X(1) = 01, c_X(2) = 10, c_X(3) = 11$, we see that $\mathbb{E}[\text{length}(c_X(X))] = 2$.
Similarly, constructing a code with $c_Y(0) = 0, c_Y(1) = 1$, we have $\mathbb{E}[\text{length}(c_Y(Y))] = 1$.

Hence both lower bounds are tight.

Alternatively: tightness follows from the existence of a prefix free code with code word lengths $l_i = \lceil -\log(p_i) \rceil$ (Shannon-Fano coding) + computing $\mathbb{E}[l_i]$.

Alternative 2: tightness follows from the fact that the marginal distributions P_X and P_Y are uniform.

(b) We calculate $H(X,Y) = 2.5$. Therefore,

$$2.5 \le \mathbb{E}[\text{length}(c_{XY}(X,Y))] \le 3.5 \tag{6}$$

We construct (for example) the following code: $c_{XY}(0,0) = 00, c_{XY}(3,1) = 01, c_{XY}(1,0) = 100, c_{XY}(1,1) = 101, c_{XY}(2,0) = 110, c_{XY}(2,1) = 111$, we have $\mathbb{E}[\text{length}(c_{XY}(X,Y))] = 2.5$ Hence, the lower bound is tight.

Alternatively: tightness follows from the existence of a prefix free code with code word lengths $l_i = \lceil -\log(p_i) \rceil$ (Shannon-Fano coding) + computing $\mathbb{E}[l_i]$.

(c) We have that $\mathbb{E}[\text{length}(c_X(X))] + \mathbb{E}[\text{length}(c_Y(Y))] \ge H(X) + H(Y) = 3 > 2.5 = \mathbb{E}[\text{length}(c_{XY}(X,Y))]$. Thus, from the tightness of the bounds in 3 a) and 3 b), it follows that it is better to compress jointly.

(4) When $I(X;Y) > 1$, compressing jointly is guaranteed to be better as

$$\begin{aligned}
\mathbb{E}[\text{length}(c_X(X))] + \mathbb{E}[\text{length}(c_Y(Y))] &\ge H(X) + H(Y) && (7)\\
&= H(X,Y) + I(X;Y) && (8)\\
&> H(X,Y) + 1 && (9)\\
&\ge \mathbb{E}[\text{length}(c_{XY}(X,Y))] && (10)
\end{aligned}$$

Problem 3 (*Stability implies Generalization*)**.** [12 pts]

Let $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ be a training dataset composed of n i.i.d. samples drawn from $\mathcal{D}$. As usual, we denote $L_{\mathcal{D}}(h) = E_{(x,y)\sim\mathcal{D}}[l(h(x), y)]$ and $L_{\mathcal{S}}(h) = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)$ the true and empirical risks of a hypothesis h, respectively. For simplicity, let us denote by h_S the output of a learning algorithm when trained with dataset S.

An important property of learning algorithms is their ability to generalize, i.e., the true and empirical risks of the output hypothesis should be close in expectation. Formally, we say that a learning algorithm $\mathcal{A}$ ϵ-generalizes in expectation if

$$|E_S[L_S(h_S) - L_{\mathcal{D}}(h_S)]| < \epsilon \ . \tag{11}$$

An interesting connection arises when we investigate the *stability* of a learning algorithm. Formally, we call a learning algorithm ϵ*-uniformly stable* if $\forall S, S'$ datasets of size n that differ in at most one sample we have

$$\sup_{(x,y)} l(h_S(x), y) - l(h_{S'}(x), y) < \epsilon \ . \tag{12}$$

Notations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), (\widetilde{x}_1, \widetilde{y}_1), \ldots, (\widetilde{x}_n, \widetilde{y}_n)$ are $2n$ independently sampled training examples. We define $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $\widetilde{S} = \{(\widetilde{x}_1, \widetilde{y}_1), \ldots, (\widetilde{x}_n, \widetilde{y}_n)\}$ and $S^{(i)} = \{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (\widetilde{x}_i, \widetilde{y}_i), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}$.

(1) [2 pts] Prove that $L_{\mathcal{D}}(h_S) = E_{\widetilde{S}}[\frac{1}{n}\sum_{i=1}^{n} l(h_S(\widetilde{x}_i), \widetilde{y}_i)]$.

(2) [3 pts] Prove that $E_{S,\widetilde{S}}[l(h_S(\widetilde{x}_i), \widetilde{y}_i)] = E_{S,S^{(i)}}[l(h_{S^{(i)}}(x_i), y_i)]$.

(3) [7 pts] Prove that an ϵ-uniformly stable learning algorithm ϵ-generalizes in expectation, by justifying each step in the following sequence.

$$
\begin{aligned}
|E_S[L_S(h_S) - L_{\mathcal{D}}(h_S)]| &\overset{(a)}{=} |E_S\left[L_S(h_S) - E_{\tilde{S}}\left[\frac{1}{n}\sum_{i=1}^{n} l(h_S(\tilde{x_i}), \tilde{y_i})\right]\right]| \\
&\overset{(b)}{=} |E_S\left[L_S(h_S)\right] - E_{S,\tilde{S}}\left[\frac{1}{n}\sum_{i=1}^{n} l(h_S(\tilde{x_i}), \tilde{y_i})\right]| \\
&\overset{(c)}{=} |E_S\left[L_S(h_S)\right] - \frac{1}{n}\sum_{i=1}^{n} E_{S,\tilde{S}}\left[l(h_S(\tilde{x_i}), \tilde{y_i})\right]| \\
&\overset{(d)}{=} |E_S\left[L_S(h_S)\right] - \frac{1}{n}\sum_{i=1}^{n} E_{S^{(i)},(x_i,y_i)}\left[l(h_{S^{(i)}}(x_i), y_i)\right]| \\
&\overset{(e)}{=} |E_S\left[\frac{1}{n}\sum_{i=1}^{n} l(h_S(x_i), y_i))\right] - \frac{1}{n}\sum_{i=1}^{n} E_{S,S^{(i)}}\left[l(h_{S^{(i)}}(x_i), y_i)\right]| \\
&\overset{(f)}{=} |\frac{1}{n}\sum_{i=1}^{n} E_{S,S^{(i)}}\left[l(h_S(x_i), y_i)) - l(h_{S^{(i)}}(x_i), y_i)\right]| \\
&\overset{(g)}{\leq} \frac{1}{n}\sum_{i=1}^{n} \epsilon = \epsilon
\end{aligned}
$$

Solution:

1. Note that since $\tilde{S}$ is composed of n i.i.d. samples $L_{\mathcal{D}}(h_S) = E_{(\tilde{x_i},\tilde{y_i})\sim\mathcal{D}}[l(h_S(\tilde{x_i}), \tilde{y_i})]$ for all i. Thus, by linearity of expectation $L_{\mathcal{D}}(h_S) = E_{\tilde{S}}[\frac{1}{n}\sum_{i=1}^{n} l(h_S(\tilde{x_i}), \tilde{y_i})]$.

2. $E_{S,\tilde{S}}[l(h_S(\tilde{x_i}), \tilde{y_i})] = E_{S,(\tilde{x_i},\tilde{y_i})}[l(h_S(\tilde{x_i}), \tilde{y_i})] =$
   (since $(x_1, y_1), \ldots, (x_n, y_n), (\tilde{x_i}, \tilde{y_i})$ are i.i.d. we can interchange (x_i, y_i) with $(\tilde{x_i}, \tilde{y_i})$)
   $= E_{S^{(i)},(x_i,y_i)}[l(h_{S^{(i)}}(x_i), y_i)]$

3.
$$
\begin{aligned}
&|E_S[L_S(h_S) - L_{\mathcal{D}}(h_S)]| \overset{(1)}{=} |E_S\left[L_S(h_S) - E_{\tilde{S}}\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n} l(h_S(\tilde{x_i}), \tilde{y_i})\right]\right]| = \\
&= |E_S\left[L_S(h_S)\right] - E_{S,\tilde{S}}\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n} l(h_S(\tilde{x_i}), \tilde{y_i})\right]| = \\
&= |E_S\left[L_S(h_S)\right] - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} E_{S,\tilde{S}}\left[l(h_S(\tilde{x_i}), \tilde{y_i})\right]| \overset{(2)}{=} \\
&= |E_S\left[L_S(h_S)\right] - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} E_{S^{(i)},(x_i,y_i)}\left[l(h_{S^{(i)}}(x_i), y_i)\right]| = \\
&= |E_S\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n} l(h_S(x_i), y_i))\right] - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} E_{S,S^{(i)}}\left[l(h_{S^{(i)}}(x_i), y_i)\right]| = \\
&= |\tfrac{1}{n}\textstyle\sum_{i=1}^{n} E_{S,S^{(i)}}\left[l(h_S(x_i), y_i)) - l(h_{S^{(i)}}(x_i), y_i)\right]| \overset{(\text{ ϵ-uniform stability})}{\leq} \\
&\leq \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \epsilon = \epsilon
\end{aligned}
$$

Problem 4 (*Multi-arm Bandits*)**.** [20 pts]

We consider the following game where in each round t we can choose between $[N] = \{1, 2, \ldots, N\}$ different actions. After we choose an action $a_t \in [N]$ an adversary reveals the loss of each action in this round, call it $l_i^t \in [0,1]$, $i \in [N]$. Note that this is an adversarial setting, where the losses do not come from a probability distribution. This setting differs from what we had discussed in class where only the loss for the chosen action was revealed.

Our goal is to design a randomized algorithm $\mathcal{A}$ which maintains a probability distribution p^t over actions, and achieves a sub-linear regret, i.e., $\mathcal{R}(T) = \max_i\{\sum_{t=1}^T E_{A_t\sim p^t}\left[l_{A_t}^t - l_i^t\right]\} \leq o(T)$. We also note that the adversary may know the probability distribution p^t, but does not know the realizations A_t. We will analyze the following algorithm:

Algorithm 1: Multiplicative Weights Update

Input: learning parameter ϵ
Initialization: $p_i^1 = 1/N, w_i^1 = 1, \forall i \in [N], \Phi^1 = N$
for $t = 1$ to T **do**
  $A_t \sim p^t$
  Adversary reveals the loss vector l^t and we suffer $l_{A_t}^t$
  Update weights $w_i^{t+1} = w_i^t \cdot \exp(-\epsilon \cdot l_i^t), \forall i \in [N]$ and let $\Phi^{t+1} = \sum_i w_i^{t+1}$
  Update the probability distribution: $p_i^{t+1} = w_i^{t+1}/\Phi^{t+1}, \forall i$
end for

---

(1) [2 pts] Prove that $w_i^{T+1} = \exp(-\epsilon \cdot \sum_{t=1}^T l_i^t), \forall i \in [N]$

(2) [8 pts] Prove that $\Phi^{t+1} \leq \Phi^t \cdot \exp(\epsilon^2 - \epsilon\langle p^t,\ l^t\rangle)$

Hint: Note that $w_i^{t+1} = p_i^{t+1} \cdot \Phi^{t+1}$ and use the inequalities: (a) $e^x \leq 1+x+x^2, \forall x \in [0,1]$ and (b) $e^x \geq x+1, \forall x$.

(3) [2 pts] Prove that $\Phi^{T+1} \leq \Phi^1 \cdot \exp(\epsilon^2 \cdot T - \epsilon \sum_{t=1}^T \langle p^t,\ l^t\rangle)$

(4) [8 pts] By noting that $\Phi^1 \cdot \exp(\epsilon^2 \cdot T - \epsilon \sum_{t=1}^T \langle p^t,\ l^t\rangle) \geq \Phi^{T+1} \geq w_i^{T+1}, \forall i \in [N]$ set the learning parameter ϵ so that $\mathcal{R}(T) \leq 2\sqrt{\log(N) \cdot T}$.

Solution:

1. Using induction we will prove that $w_i^{t'+1} = \exp(-\epsilon \cdot \sum_{t=1}^{t'} l_i^t), \forall i \in [N]$. Note that for $t' = 1$, we get that $w_i^2 = w_i^1 \cdot \exp(-\epsilon \cdot l_i^1) = \exp(-\epsilon \cdot l_i^1)$. Assume that the hypothesis is true for $t'-1$ then we get that $w_i^{t'+1} = w_i^{t'} \cdot \exp(-\epsilon \cdot l_i^{t'}) \overset{\text{(induction hypothesis)}}{=} \exp(-\epsilon \cdot \sum_{t=1}^{t'-1} l_i^t) \cdot \exp(-\epsilon \cdot l_i^{t'}) = \exp(-\epsilon \cdot \sum_{t=1}^{t'} l_i^t)$

2.

$$\begin{aligned}
\Phi^{t+1} &= \textstyle\sum_i w_i^{t+1} = \sum_i w_i^t \cdot \exp(-\epsilon \cdot l_i^t) \overset{(a)}{\leq} \\
&\textstyle\sum_i w_i^t \cdot (1 - \epsilon \cdot l_i^t + \epsilon^2 \cdot (l_i^t)^2) \overset{l_i^t \in [0,1]}{\leq} \\
&\textstyle\sum_i w_i^t \cdot (1 - \epsilon \cdot l_i^t + \epsilon^2) = \\
&\textstyle\sum_i w_i^t \cdot (1 + \epsilon^2) - \sum_i w_i^t \cdot \epsilon \cdot l_i^t = \\
&\textstyle\sum_i w_i^t \cdot (1 + \epsilon^2) - \sum_i p_i^t \cdot \Phi^t \cdot \epsilon \cdot l_i^t = \\
&\textstyle(1 + \epsilon^2) \cdot \Phi^t - \Phi^t \cdot \sum_i p_i^t \cdot \epsilon \cdot l_i^t = \\
&\textstyle(1 + \epsilon^2) \cdot \Phi^t - \Phi^t \cdot \epsilon \cdot \sum_i p_i^t \cdot l_i^t = \\
&(1 + \epsilon^2) \cdot \Phi^t - \Phi^t \cdot \epsilon \cdot \langle p^t,\ l^t \rangle = \\
&\Phi^t \cdot (1 + (\epsilon^2 - \epsilon \cdot \langle p^t,\ l^t \rangle)) \overset{(b)}{\leq} \\
&\Phi^t \cdot \exp\left(\epsilon^2 - \epsilon \cdot \langle p^t,\ l^t \rangle\right)
\end{aligned}$$

3. It it sufficient to reapply the inequality proven in sub-question (2) for $t = T,\ t = T-1,\ t = T-2,\ \ldots,\ t = 2$.

4. From sub-questions (1) and (3) we get that for all i:

$$\exp(-\epsilon \cdot \textstyle\sum_{t=1}^T l_i^t) \leq \Phi^1 \cdot \exp(\epsilon^2 \cdot T - \epsilon \sum_{t=1}^T \langle p^t,\ l^t \rangle) = N \cdot \exp(\epsilon^2 \cdot T - \epsilon \sum_{t=1}^T \langle p^t,\ l^t \rangle) \Longrightarrow$$

$$\Longrightarrow -\epsilon \cdot \textstyle\sum_{t=1}^T l_i^t \leq \log(N) \cdot + \epsilon^2 \cdot T - \epsilon \sum_{t=1}^T \langle p^t,\ l^t \rangle \overset{\text{divide by } \epsilon}{\Longrightarrow}$$

$$\Longrightarrow \textstyle\sum_{t=1}^T \left(\langle p^t,\ l^t \rangle - l_i^t\right) \leq \frac{\log(N)}{\epsilon} \cdot + \epsilon T \Longrightarrow$$

$$\Longrightarrow \textstyle\sum_{t=1}^T E_{A_t \sim p^t}\left[l_{A_t}^t - l_i^t\right] \leq \frac{\log(N)}{\epsilon} \cdot + \epsilon T \overset{\epsilon = \sqrt{\frac{\log(N)}{T}}}{=} 2\sqrt{\log(N) \cdot T}$$