Final exam (May, 23rd)

Problem 1 (Basics)**.** Give a concise (but justified) answer to the following questions. When using a theorem from the course, state it clearly.

1. Let $X_1, X_2 \in \{0,1\}$ be independent and uniform bits. Show that X_1 and $X_1 \oplus X_2$ are independent random variables, here $\oplus$ is the exclusive-or.

2. Let X_1 and X_2 be two independent exponential random variables with parameters θ_1 and θ_2. Compute $\mathbf{P}\{\min(X_1, X_2) = X_1\}$. Recall that an exponential random variable with parameter θ has a density given by $f(x) = \theta e^{-\theta x}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$.

3. Let U_1 and U_2 be independent random variables that are uniformly distributed on $[0,1]$. Compute $\mathbf{P}\{U_1 \geq 2U_2\}$.

4. Let Y_n be a uniformly distributed random variable on $\{1, \ldots, n\}$. Show that $\frac{Y_n}{n}$ converges in distribution to X where X is uniformly distributed on $[0,1]$.

5. Consider a Markov chain $\{X_n\}_{n\in\mathbb{N}}$ on the state space $\mathbb{Z}$. The transition probabilities are given by $\mathbf{P}\{X_1 = k | X_0 = k'\} = \frac{1}{2}$ if $|k - k'| = 1$ and 0 otherwise. Is the state 0 recurrent or transient?

6. Compute the stationary distribution for the random walk on a path of length n, i.e., the vertices are $V = \{1, \ldots, n\}$ and the edges are $E = \{\{i, i+1\} \text{ for } i \in \{1, \ldots, n-1\}\}$.

7. Take a random graph G according to the model $\mathcal{G}_{n,\frac{1}{2}}$. Find a good upper bound on the probability that the number of edges in G is at least $\frac{3}{4} \cdot \binom{n}{2}$.

8. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with $\mathbf{E}\{X_1\} = \mu$ and $\mathbf{Var}\{X_1\} = v$ (both finite). For which value of a_n and b_n is the following statement true: $\frac{(\sum_{i=1}^n X_i) - a_n}{b_n}$ converges in distribution to $\mathcal{N}(0,1)$.

9. Find a good upper bound on $\mathbf{P}\left\{|\sum_{i=1}^{1000} X_i - 500| \geq 50\right\}$ where X_i are independent and uniform bits. Given 1000 tosses of a coin, you find 550 heads. Is it reasonable to say that the coin is fair? Give a hypothesis testing framework to answer this question and give an upper bound for the p-value for this data.

Problem 2. Consider a Markov chain with state space $S = \{0, \ldots, N\}$ and transition probabilities $P_{i,i+1} = p$ for $0 \leq i \leq N-1$, $P_{i,i-1} = 1-p$ for $1 \leq i \leq N$ and $P_{0,0} = 1-p$, $P_{N,N} = p$, with $0 < p < 1$.

1. Give a graphical representation of the transition probabilities of this Markov chain.

2. Is this Markov chain irreducible? aperiodic?

3. Compute the stationary distribution of this Markov chain as a function of p. *Hint:* It is sufficient to satisfy $\pi_i(1-p) = \pi_{i-1}p$

4. Starting from the state 0, what is the expected time to return to zero as a function of p?

Problem 3 (Isolation lemma)**.** Let $n, N \in \mathbb{N}$ and $\mathcal{F}$ be a family of subsets of $\{1, \dots, n\}$. Our objective is to find a weight function $w : \{1, \dots, n\} \to \{1, \dots, N\}$ with the following property. Defining $w(S) = \sum_{x \in S} w(x)$ for any $S \subseteq \{1, \dots, n\}$, we want that there exists a unique set $S \in \mathcal{F}$ of minimum weight $w(S)$. This means that there exists $S_0 \in \mathcal{F}$ such that $\min_{S \in \mathcal{F}} w(S) = w(S_0)$ and for all $S \in \mathcal{F}$ with $S \neq S_0$, $w(S) > w(S_0)$. We then say that the function w is isolating for $\mathcal{F}$.

1. We start with a very simple family: $\mathcal{F}$ is the family of all non-empty subsets of $\{1, \dots, n\}$. For $N \geq 2$, construct an isolating function for $\mathcal{F}$.

2. We now choose the function w at random, so that the values $\{w(x)\}_{x \in \{1,\dots,n\}}$ are mutually independent and uniformly distributed on $\{1, \dots, N\}$. We will show that for any choice of $\mathcal{F}$, the function w is isolating for $\mathcal{F}$ with probability at least $1 - \frac{n}{N}$. For this we now fix $\mathcal{F}$ as an arbitrary family and we introduce $\alpha(x) = \min\limits_{S \in \mathcal{F} : x \notin S} w(S) - \min\limits_{S \in \mathcal{F} : x \in S} w(S - \{x\})$ for any $x \in \{1, \dots, n\}$.

   (a) Show that $\mathbf{P}\{\exists x \in \{1, \dots, n\} : \alpha(x) = w(x)\} \leq \frac{n}{N}$.

   (b) Using the previous question show that the random weight function w defined above is isolating for $\mathcal{F}$ with probability at least $1 - \frac{n}{N}$.

Problem 4. In this problem, we will study a model for efficient learning of a concept by observing random examples. Intuitively, the objective is to learn a function $f : \mathcal{X} \to \{0, 1\}$ when given examples of the form $(x_1, f(x_1)), \dots, (x_m, f(x_m))$. For instance, we might have $\mathcal{X} = \mathbb{R}$ and $f(x) = \mathbf{1}_{x \geq a}$, where $\mathbf{1}_{x \geq a} = 1$ if $x \geq a$ and 0 otherwise, for some (unknown) value of $a \in \mathbb{R}$.

To model this situation, a *concept class* $\mathcal{C}$ is a set of functions. For the instance described above, the concept class is given by $\mathcal{C}_{\text{half-line}} = \{f : \mathbb{R} \to \{0, 1\} : f(x) = \mathbf{1}_{x \geq a} \text{ for some } a \in \mathbb{R}\}$. When do we say that a concept class $\mathcal{C}$ is learnable? Intuitively, we want an algorithm $\mathcal{A}$ to be able to determine from a reasonable number of examples $(x_i, f(x_i))$ the correct function f among all the possible functions in $\mathcal{C}$. However, learning the exact function might be impossible so our objective will only be to produce a function that approximates f. In addition, we will only require the output of the algorithm to be correct with high probability. This model is called the Probably Approximately Correct (or PAC) model.

Formally, we say that a concept class $\mathcal{C}$ is PAC-learnable if there is an algorithm $\mathcal{A}$ and a polynomial $p(s, t)$ in two variables such that for any $f \in \mathcal{C}$, any probability measure $\mathcal{D}$ on $\mathcal{X}$, any $\epsilon \in]0, 1/2[$, any $\delta \in]0, 1[$, the algorithm $\mathcal{A}$ takes as input $m = p(\frac{1}{\epsilon}, \frac{1}{\delta})$ independent samples $x_1, \dots, x_m$ according to $\mathcal{D}$, together with the evaluations of f at these points $f(x_1), \dots, f(x_m)$ and it produces $\tilde{f}$ (which is random as it depends on $x_1, \dots, x_m$). The output $\tilde{f}$ should satisfy

$$\mathop{\mathbf{P}}_{x_1, \dots, x_m \sim \mathcal{D}^{\times m}} \left\{ \text{Err}_{\mathcal{D}}(\tilde{f}, f) \leq \epsilon \right\} \geq 1 - \delta \, ,$$

where $\text{Err}_{\mathcal{D}}(\tilde{f}, f) = \mathop{\mathbf{P}}\limits_{x \sim \mathcal{D}} \left\{ f(x) \neq \tilde{f}(x) \right\}$.

1. We now show that the concept class $\mathcal{C}_{\text{half-line}} = \{f : f(x) = \mathbf{1}_{x \geq a} \text{ for some } a \in \mathbb{R}\}$ is PAC-learnable. For this, we fix an $f \in \mathcal{C}_{\text{half-line}}$ as $f(x) = \mathbf{1}_{x \geq a}$, a probability measure $\mathcal{D}$ over $\mathbb{R}$, and ϵ and δ.

   (a) Consider the following simple algorithm which takes m examples of the form $(x_i, f(x_i))$ and outputs the function $\tilde{f}$ defined by $\tilde{f}(x) = \mathbf{1}_{x \geq \hat{a}}$ where $\hat{a} = \min\{x_i : i \in \{1, \dots, m\}, f(x_i) = 1\}$. If the set is empty, then define $\hat{a} = +\infty$. Show that $\text{Err}_{\mathcal{D}}(\tilde{f}, f) = \mathcal{D}([a, \hat{a}[)$.

(b) Assume in this question and the next that a is such that $\mathcal{D}([a, +\infty[) \geq \epsilon$. Define $a_\epsilon = \sup\{a' : \mathcal{D}([a, a'[) \leq \epsilon\}$. Show that $\mathcal{D}([a, a_\epsilon[) \leq \epsilon$ and $\mathcal{D}([a, a_\epsilon]) \geq \epsilon$. *Note:* You will get partial credit for doing the special case where $\mathcal{D}$ is finite for example.

(c) Find an upper bound on the probability $\mathbf{P}\{\hat{a} > a_\epsilon\}$ (here the probability is over the choice of $x_1, \ldots, x_m$).

(d) Conclude on the PAC-learnability of the concept class $\mathcal{C}_{\text{half-line}}$. How large should we take m as a function of ϵ and δ?

2. We now consider another concept class. Let $\mathcal{X} = \{0,1\}^n$ and the functions are the disjunctions of a subset of the bits, i.e., $\mathcal{C}_{\text{1-disj}} = \{f : \{0,1\}^n \to \{0,1\} : f(z_1, \ldots, z_n) = \vee_{j \in S} z_j \text{ for some } S \subseteq \{1, \ldots, n\}\}$.

(a) Propose a natural algorithm for learning a function $f \in \mathcal{C}_{\text{1-disj}}$ efficiently. For example, you could start with $S = \{1, \ldots, n\}$ and for each example z such that $f(z) = 0$, remove all elements $j \in S$ for which $z_j = 1$.

(b) Show that this algorithm PAC-learns the class $\mathcal{C}_{\text{1-disj}}$ using a number of examples that is polynomial in $1/\epsilon$, $1/\delta$ and also in n.

French version

Problem 1 (Basiques)**.** Donner une réponse concise (mais justifiée) aux questions suivantes. Si vous utilisez un théorème du cours, énoncez-le clairement.

1. Soit $X_1, X_2 \in \{0,1\}$ des bits indépendants et uniformes. Montrer que X_1 et $X_1 \oplus X_2$ sont des variables aléatoires indépendantes, où $\oplus$ est le "ou exclusif".

2. Soit X_1 et X_2 deux variables aléatoires indépendantes suivant des lois exponentielles de paramètres θ_1 et θ_2. Calculer $\mathbf{P}\{\min(X_1, X_2) = X_1\}$. Rappel: la fonction de densité d'une loi exponentielle de paramètre θ est donnée par $f(x) = \theta e^{-\theta x}$ pour $x \geq 0$ et $f(x) = 0$ pour $x < 0$.

3. Soit U_1 et U_2 des variables aléatoires indépendantes uniformément distribuées sur $[0,1]$. Calculer $\mathbf{P}\{X_1 \geq 2X_2\}$.

4. Soit Y_n une variable aléatoire uniformément distribuée sur $\{1,\ldots,n\}$. Montrer que $\frac{Y_n}{n}$ converge en distribution vers X où X est uniformément distribuée sur $[0,1]$.

5. Soit une chaîne de Markov $\{X_n\}_{n\in\mathbb{N}}$ sur l'espace d'état $\mathbb{Z}$. Les probabilités de transition sont données par $\mathbf{P}\{X_1 = k | X_0 = k'\} = \frac{1}{2}$ si $|k - k'| = 1$ et 0 sinon. Est-ce que l'état 0 est récurrent ou transient ?

6. Calculer la distribution stationnaire de la marche aléatoire sur un chemin de longueur n, i.e., les sommets sont $V = \{1,\ldots,n\}$ et les arêtes sont $E = \{\{i, i+1\} \text{ for } i \in \{1,\ldots,n-1\}\}$.

7. Soit un graphe aléatoire G tiré selon le modèle $\mathcal{G}_{n,\frac{1}{2}}$. Trouver une borne supérieure sur la probabilité que le nombre d'arêtes de G est au mois $\frac{3}{4} \cdot \binom{n}{2}$.

8. Soit $X_1,\ldots,X_n$ des variables aléatoires indépendantes et identiquement distribuées avec $\mathbf{E}\{X_1\} = \mu$ et $\mathbf{Var}\{X_1\} = v$ (μ et v sont finis). Pour quelles valeurs de a_n et b_n avons-nous : $\frac{(\sum_{i=1}^n X_i) - a_n}{b_n}$ converge en distribution à $\mathcal{N}(0,1)$.

9. Trouver une bonne borne supérieure pour $\mathbf{P}\left\{|\sum_{i=1}^{1000} X_i - 500| \geq 50\right\}$ où X_i sont des bits uniformes et indépendants. Je lance une pièce 1000 fois et j'obtiens 550 piles. Est-ce raisonnable de supposer la pièce non-biaisée ? Formuler le problème en tant que test d'hypothèse et donner une borne supérieure sur la p-valeur de ces données.

Problem 2. Soit une chaîne de Markov sur l'espace d'états $S = \{0,\ldots,N\}$ et de probabilités de transition $P_{i,i+1} = p$ pour $0 \leq i \leq N-1$, $P_{i,i-1} = 1-p$ pour $1 \leq i \leq N$ et $P_{0,0} = 1-p$, $P_{N,N} = p$, avec $0 < p < 1$.

1. Donner une représentation graphique des probabilités de transition de cette chaîne de Markov.

2. Cette chaîne de Markov est-elle irréductible? Apériodique?

3. Calculer la distribution stationnaire π de cette chaîne de Markov en fonction de p. *Aide :* Une condition suffisante pour π est de satisfaire $\pi_i(1-p) = \pi_{i-1}p$.

4. En démarrant à l'état 0, quelle est l'espérance du temps de retour en 0 en fonction de p?

Problem 3. Soit $n, N \in \mathbb{N}$, et $\mathcal{F}$ une famille de sous-ensembles de $\{1, \ldots, n\}$. Notre objectif est de trouver une fonction de poids $w : \{1, \ldots, n\} \to \{1, \ldots, N\}$ satisfaisant la propriété suivante. En notant $w(S) = \sum_{x \in S} w(x)$ défini pour tout $S \subseteq \{1, \ldots, n\}$, on souhaite qu'il existe un unique $S \in \mathcal{F}$ de poids minimum $w(S)$. En d'autres termes, on souhaite qu'il existe $S_0 \in \mathcal{F}$ tel que $\min_{S \in \mathcal{F}} w(S) = w(S_0)$ et pour tout $S \in \mathcal{F}$ avec $S \neq S_0$, $w(S) > w(S_0)$. On dit alors que w est une fonction *isolante* pour $\mathcal{F}$.

1. On commence par une famille très simple: $\mathcal{F}$ est la famille de tous les sous-ensembles non-vides de $\{1, \ldots, n\}$. Pour $N \geq 2$, construire une fonction isolante pour $\mathcal{F}$.

2. On choisit maintenant la fonction w aléatoirement de la manière suivante: les valeurs $\{w(x)\}_{x \in \{1,\ldots,n\}}$ sont mutuellement indépendantes et uniformément distribuées sur $\{1, \ldots, N\}$. On va montrer que pour tout choix de $\mathcal{F}$, la fonction w est isolante pour $\mathcal{F}$ avec probabilité au moins $1 - \frac{n}{N}$. Pour ceci, on fixe désormais $\mathcal{F}$ arbitrairement et on définit $\alpha(x) = \min_{S \in \mathcal{F} : x \notin S} w(S) - \min_{S \in \mathcal{F} : x \in S} w(S - \{x\})$ pour tout $x \in \{1, \ldots, n\}$.

   (a) Montrer que $\mathbf{P}\{\exists x \in \{1, \ldots, n\} : \alpha(x) = w(x)\} \leq \frac{n}{N}$.

   (b) En utilisant la question précédente, montrer que la fonction aléatoire w définie ci-dessus est isolante pour $\mathcal{F}$ avec probabilité au moins $1 - \frac{n}{N}$.

Problem 4. Dans ce problème, nous étudierons un modèle pour l'apprentissage efficace d'un concept en observant des exemples aléatoires. Intuitivement, l'objectif est d'apprendre une fonction $f : \mathcal{X} \to \{0, 1\}$ étant donné des exemples de la forme $(x_1, f(x_1)), \ldots, (x_m, f(x_m))$. Nous pouvons avoir par exemple $\mathcal{X} = \mathbb{R}$ et $f(x) = \mathbf{1}_{x \geq a}$, où $\mathbf{1}_{x \geq a} = 1$ si $x \geq a$ et 0 sinon, pour une valeur (inconnue) de $a \in \mathbb{R}$.

Pour modéliser cette situation, une *classe de concepts* $\mathcal{C}$ est un ensemble de fonctions. Pour l'instance décrite ci-dessus, la classe de concept est donnée par $\mathcal{C}_{\text{half-line}} = \{f : \mathbb{R} \to \{0, 1\} : f(x) = \mathbf{1}_{x \geq a} \text{ pour un } a \in \mathbb{R}\}$. Quand disons-nous qu'une classe de concepts $\mathcal{C}$ est apprenable ? Intuitivement, nous voulons qu'un algorithme $\mathcal{A}$ puisse déterminer à partir d'un nombre raisonnable d'exemples $(x_i, f(x_i))$ la bonne fonction f parmi toutes les fonctions possibles dans $\mathcal{C}$. Cependant, apprendre la fonction exacte pourrait être impossible, donc notre objectif sera seulement de produire une fonction qui se rapproche de f. De plus, nous exigerons uniquement que la sortie de l'algorithme soit correcte avec grande probabilité. Ce modèle s'appelle le modèle Probablement Approximativement Correct (ou PAC).

Plus precisement, nous disons qu'une classe de concepts $\mathcal{C}$ est PAC-apprenable s'il existe un algorithme $\mathcal{A}$ et un polynome $p(s, t)$ en deux variables tels que pour $f \in \mathcal{C}$, toute mesure de probabilité $\mathcal{D}$ sur $\mathcal{X}$, tout $\epsilon \in]0, 1/2[$, tout $\delta \in]0, 1[$, l'algorithme $\mathcal{A}$ prend en entrée $m = p(\frac{1}{\epsilon}, \frac{1}{\delta})$ échantillons indépendants $x_1, \ldots, x_m$ selon $\mathcal{D}$, ainsi que les évaluations de f à ces points $f(x_1), \ldots, f(x_m)$ et il produit $\tilde{f}$ (qui est aléatoire car elle dépend de $x_1, \ldots, x_m$). La sortie $\tilde{f}$ doit satisfaire

$$\mathop{\mathbf{P}}_{x_1, \ldots, x_m \sim \mathcal{D}^{\times m}} \left\{ \text{Err}_{\mathcal{D}}(\tilde{f}, f) \leq \epsilon \right\} \geq 1 - \delta$$

où $\text{Err}_{\mathcal{D}}(\tilde{f}, f) = \mathop{\mathbf{P}}_{x \sim \mathcal{D}} \left\{ f(x) \neq \tilde{f}(x) \right\}$.

1. Nous montrons maintenant que la classe de concept $\mathcal{C}_{\text{half-line}} = \{f : f(x) = \mathbf{1}_{x \geq a} \text{ pour un } a \in \mathbb{R}\}$ est PAC-apprenable. Pour cela, nous fixons un $f \in \mathcal{C}_{\text{half-line}}$ comme $f(x) = \mathbf{1}_{x \geq a}$, une mesure de probabilité $\mathcal{D}$ sur $\mathbb{R}$, et ϵ et δ.

   (a) Considérez l'algorithme simple suivant qui prend m examples de la form $(x_i, f(x_i))$ et émet la fonction $\tilde{f}$ définie par $\tilde{f}(x) = \mathbf{1}_{x \geq \hat{a}}$ où $\hat{a} = \min\{x_i : i \in \{1, \ldots, m\}, f(x_i) = 1\}$. Prouver que $\text{Err}_{\mathcal{D}}(\tilde{f}, f) = \mathcal{D}([a, \hat{a}[)$.

(b) Définir $a_\epsilon = \sup\{a' : \mathcal{D}([a, a'[) \leq \epsilon\}$. Montrez que $\mathcal{D}([a, a_\epsilon[) \leq \epsilon$ et $\mathcal{D}([a, a_\epsilon]) \geq \epsilon$. *Note:* Des points seront accordés si la question est traitée dans un cas spécial, par exemple, $\mathcal{D}$ fini.

(c) Trouver une borne supérieure sur la probabilité $\mathbf{P}\{\hat{a} > a_\epsilon\}$ (ici la probabilité est sur le choix de $x_1, \ldots, x_m$).

(d) Conclure sur le PAC-apprenabilité de la classe de concepts $\mathcal{C}_{\text{half-line}}$. Quelle valeur choisir pour m en fonction de ϵ et δ ?

2. Nous considérons maintenant une autre classe de concepts. Soit $\mathcal{X} = \{0,1\}^n$ et les fonctions sont les disjonctions d'un sous-ensemble des bits, c'est-à-dire $\mathcal{C}_{\text{1-disj}} = \{f : \{0,1\}^n \to \{0,1\} : f(z_1, \ldots, z_n) = \vee_{j \in S} z_j$pour un certain$S \subseteq \{1, \ldots, N\}\}$.

(a) Proposer un algorithme naturel pour apprendre une fonction $f \in \mathcal{C}_{\text{1-disj}}$ efficacement. Par exemple, vous pouvez commencer par $S = \{1, \ldots, n\}$ et pour chaque exemple z tel que $f(z) = 0$, supprimez tous les éléments $j \in S$ pour lesquels $z_j = 1$.

(b) Montrer que cet algorithme PAC-apprend la classe $\mathcal{C}_{\text{1-disj}}$ en utilisant un nombre d'exemples polynomial en $1/\epsilon$, $1/\delta$ et aussi en n .