TD n°03

09 février 2016

Homework exercises

Exercice 1 -

In class, we covered a randomized algorithm to check the product of matrices over $\mathbb{F}_2$. In an analogous way, design and analyze a randomized algorithm for checking the product of matrices over $\mathbb{F}_p$ where $p$ is a prime number. The error probability should be at most $1/100$.

Réponse :

Consider the following algorithm for a probabilistic check $AB \equiv C \pmod{p}$ for $A, B, C \in \mathbb{F}_p^{n \times n}$. Repeat $k$ times:

1. Take $\vec{r} \leftarrow \mathbb{F}_p^n$ uniformly at random

2. Check if $AB\vec{r} = C\vec{r} \pmod{p}$. Continue, if true; return "not equal", otherwise.

The running time of the algorithm is $O(k \cdot n^2 \log^2 p)$.

The algorithm has one-sided error: for a uniformly random $\vec{r}$, if $AB \neq C \pmod{p}$, the check on the second step returns true with probability $\leq 1/p$. Repeating $k$-times, the error probability gets decreased to $\leq (1/p)^k$. We set $k \geq \lceil \frac{\log 100}{\log p} \rceil$ to satisfy the error probability bound.

Exercice 2 - Show that the events $\{A_i\}_{1 \leq i \leq n}$ are mutually independent if and only if

$$P\{\cap_{i=1}^n B_i\} = \prod_{i=1}^n P\{B_i\}$$

where for every $i$, either $B_i = A_i$ or $B_i = A_i^c$. We use the notation $A^c$ for the complement of $A$ in $\Omega$.

Réponse :

$\Rightarrow$ Use induction on $n$ and the fact that $P\{\cap_j B_j \cap A_i^c\} = P\{\cap_j B_j\} - P\{A_i^c\} = \prod_j P\{B_j\} - \prod_j P\{B_j\} \cdot A_i$, where we used (1) the induction on $n$ (to obtain the first summand) and the induction on the number of complements (to obtain the second summand).

$\Leftarrow$ Need to show that $P\{\cap_{i \in \{1,...,n\}} A_i\} = \prod_{i \in I} A_i$ for all subsets $I \subset \{1...n\}$. Use induction on $|I|$. For $|I| = n$, we have the result setting $B_i = A_i$. Let us show that we also have $P\{\cap_{i \in \{1,...,n\}} A_i\} = \prod_{i \in I} A_i$ for $|I| = n-1$. We can express $\cap_{i=1}^{n-1} B_i = (\cap_{i=1}^{n-1} B_i) \cup (\cap_{i=1}^{n-1} (B_i \setminus B_n))$, where the union is taken over two disjoint sets. Therefore, $P\{\cap_{i=1}^{n-1} B_i\} = \prod_{i=1}^n P\{B_i\} + P\{\cap_{i=1}^{n-1} B_i \setminus B_n\}$. For the last summand, we use the rule $A \setminus B = A \cap B^c$, obtain $P\{B_n\} \cdot \prod_{i=1}^{n-1} P\{B_i\} + P\{B_n^c\} = \prod_{i=1}^n P\{B_i\}$.

Exercice 3 -

Write an algorithm that takes as input $n$ and generates a uniformly random permutation of $\{1,\ldots,n\}$ (represented in an array in the natural way). You can use the function $\text{RandInt}(m)$ which returns a uniform number between $\{1,\ldots,m\}$. You should justify why each permutation has probability $1/n!$. You should aim for a running time of $O(n)$, where you may assume that a
call to RandInt takes constant time and that accessing a given index of an array takes constant time. Less efficient (but correct) algorithms will be given partial credit.

Réponse: Consider the following algorithm (aka Knuth’s shuffle originally considered by Fisher and Yates):

1. Initialize an array $A$ of size $n$ with any permutation from $S_n$.
2. For $i$ from 1 to $n$ do
   (a) $i \leftarrow \text{RandInt}[i \ldots n]$

Clearly, the runtime is $O(n)$. The correctness can be easily shown by induction on $n$.

Exercice 4 - Pourquoi a-t-on besoin des $\sigma$-algèbres?

The purpose of this exercise is to show that one cannot make a “reasonable” definition of probability that is defined for all possible subsets of $\Omega$. On other words, one needs $\sigma$-algebras.

To define a uniform probability measure on $[0,1]$, the following properties should hold:

Propriété 1. For $0 \leq a \leq b \leq 1$, $P([a,b]) = P([a,b)) = P(]a,b]) = P(]a,b[) = b - a$.

Propriété 2 (Stability under shifts). For $A \subseteq [0,1]$ et $0 \leq r \leq 1$, we define the $r$-shift of $A$ by

$$A \oplus r = \{a + r \mid a \in A, a + r \leq 1\} \cup \{a + r - 1 \mid a \in A, a + r > 1\}.$$ 
Then $P$ must satisfy $P(A \oplus r) = P(A)$.

Propriété 3 (Countable additivity). If $(A_n)_{n \in \mathbb{N}}$ are disjoint subsets of $[0,1]$, then

$$P \left( \bigcup_{n \in \mathbb{N}} A_n \right) = \sum_{n \in \mathbb{N}} P(A_n).$$

We shall show that there does not exist such a function $P$ defined for all subsets of $[0,1]$.

4.1 Define a relation on $[0,1]$ as follows: $x \sim y$ if and only if the difference $y - x$ is rational. Show that this is an equivalence relation (i.e., it is reflexive, symmetric, and transitive).

Réponse: One can easily verify that
   — $\sim$ is reflexive since $x - x = 0$ is rational;
   — $\sim$ is symmetric since if $y - x$ is rational, then $x - y$ is also;
   — $\sim$ is transitive since if $x \sim y$ and $y \sim z$, then $z - x = (z - y) + (y - x)$ is rational.

4.2 Construct a subset $H$ of $[0,1]$ that contains exactly one element of each equivalence class. Such $H$ must exist by the Axiom of Choice.

Axiom of Choice. Given a collection $(A_i)_{i \in I}$ of non-empty sets (i.e., $A_i \neq \emptyset$ for all $i \in I$), their Cartesian product $\prod_{i \in I} A_i$ is also non-empty. In other words, one can choose one element simultaneously from each of the non-empty sets $A_i$.

Réponse: $H = [0,1]/\mathbb{Q}$, i.e., for each real $r$ there exist only one $h \in H$, s.t. $h - r$ is rational.
4.3 Obtenir une contradiction pour $P$ en utilisant le ensemble $H$.

Réponse : Soient $(A_i)$ des ensembles d'équivalences équidistantes. Par définition, chaque classe d'équivalence $A_i$ est non- vide. En vertu de l'axiome de choix, il est possible de choisir $H$ qui contient exactement une représentative de chaque classe d'équivalence.

Fonctionnalité de $H$, $H \oplus r_1 \cap H \oplus r_2 = \emptyset$ pour tout deux rationnels $r_1 \neq r_2$ de $[0,1]$. Si le contraire est le cas, $x \in H \oplus r_1 \cap H \oplus r_2$, ce qui implique que $h_1, h_2 \in H$ tels que $x = h_1 + r_1$ et $x = h_2 + r_2$. Or $h_1 = h_2 + (r_2 - r_1)$, ce qui est en contradiction avec le fait que $H$ contient exactement une représentative de chaque classe.

Exercice 5 - Le singe savant

Un singe tape sur un clavier de 26 lettres, chaque lettre est choisie indépendamment et uniformément aléatoirement. Si le singe tape 1,000,000 lettres, qu'est-ce que le nombre d'occurrences attendu de la phrase "preuve" ?

Réponse : Si $X_i$ est une variable aléatoire égale à 1 si la phrase "preuve" apparaît sur la position $i$, et 0 sinon. Alors $P(X_i = 1) = (1/26)^6$, $P(X_i = 0) = 1 - (1/26)^6$. Le nombre attendu d'apparitions de la phrase "preuve" est alors $\sum_{i=1}^{10^6-5} \mathbb{E}(X_i) = (1/26)^6 \cdot (10^6 - 5)$.

Exercice 6 - Test de dépistage

Nous voulons effectuer un test médical sur un grand nombre $n$ de spécimens. Cela peut être fait de deux façons :

1. Le test est effectué sur chaque spécimen séparément. Dans ce cas, $n$ tests sont effectués.
2. Les spécimens sont groupés en (disjoints) ensembles de taille $k$ et les tests sont effectués sur les groupes. Si un groupe a un résultat positif, chaque spécimen dans ce groupe est alors testé individuellement.
Let $p$ be the probability that a sample is positive. We assume that samples are independent and identically distributed and that $k$ divides $n$.

6.1 What is the probability $p_k$ that a set of $k$ samples produces positive result?

Réponse:
For a set of $k$ samples, the probability that all these samples are negative is $(1-p)^k$. Hence, the probability that at least 1 sample is positive is $p_k = 1 - (1-p)^k$.

6.2 Consider a random variable $X$ that represents the total number of tests performed when we follow the second method. Which values can $X$ take and what are the corresponding probabilities?

Réponse:
We start with performing $\frac{n}{k}$ tests for sets of samples of size $k$. For each positive set (the number of positive tests ranges from 0 to $\frac{n}{k}$), we perform $k$ more additional tests.

Hence, $X$ takes values from the set $\{\frac{n}{k} + tk, 0 \leq t \leq \frac{n}{k}\}$ and

$$P(X = \frac{n}{k} + tk) = P("t sets are positive") = \binom{n/k}{t} p_k^t (1-p_k)^{n/k-t}.$$ 

6.3 What is the expected value of the number of tests performed when the second method is used, i.e., what is $E[X]$?

Réponse: Méthode 1, dite de la brute:

$$E(X) = \sum_{t=0}^{n/k} (\frac{n}{k} + tk)P(X = \frac{n}{k} + tk)$$

$$= \sum_{t=0}^{n/k} (\frac{n}{k} + tk) \binom{n/k}{t} p_k^t (1-p_k)^{n/k-t} + \sum_{t=1}^{n/k} tk \binom{n/k}{t} p_k^t (1-p_k)^{n/k-t}$$

$$= \frac{n}{k} (p_k + (1-p_k))^{n/k} + np_k \sum_{t=1}^{n/k} \binom{n/k-1}{t-1} p_k^{t-1} (1-p_k)^{n/k-t}$$

$$= \frac{n}{k} + np_k \sum_{t=0}^{n/k-1} \binom{n/k-1}{t} p_k (1-p_k)^{(n/k-1)-t} = \frac{n}{k} + np_k$$

Méthode 2, dite "tiens, il y a « indépendance » dans le titre du TD":

We introduce a random variable $Y$ that represents the number of tests performed on one set. Since the samples are independent and identically distributed, this random variable is the same for each set (i.e., $Y$ is identical to $Y_t$).

$Y$ takes two values:

— $k + 1$ ($k$ tests after the first showed positive result) with probability $p_k = 1 - (1-p)^k$
— 1 (the first test was negative) with probability $(1-p)^k$. 

4
Hence, $\mathbb{E}(Y) = (k+1)p_k + 1(1-p_k) = 1 + kp_k$. Using linearity of expectation,

$$
\mathbb{E}(X) = \mathbb{E}\left( \sum_{t=1}^{n/k} Y \right) = \frac{n}{k} \mathbb{E}(Y) = \frac{n}{k}(1 + kp_k) = \frac{n}{k} + np_k.
$$

6.4 For small $p$, which value of $k$ minimizes the number of tests? What is $\mathbb{E}(X)$ for such $k$?

**Réponse:** For small $p$, it holds the $p_k = 1 - (1-p)^k \simeq 1 - (1-pk) = pk$. Hence, $\mathbb{E}(X) \simeq \frac{n}{k} + npk$. Taking derivative, we determine $k$ as

$$
-n\frac{k^2}{k^2} + np = 0 \iff p = \frac{1}{k^2} \iff k = \sqrt{\frac{1}{p}}.
$$

We obtain $\mathbb{E}(X) \simeq 2n\sqrt{p}$.

**Exercice 7 - 66**

We throw a fair die until we obtain two 6s in a row.

7.1 What is the expected number of throws until we stop?

**Réponse:** Attention, la réponse n’est pas 36!

Let $a$ be the expected number of throws until we stop, which is the quantity we want to obtain.

Now, consider a second game where we stop if we get a 6 in the first try or when we get two 6s in a row. Let $b$ be the expected number of throws in this game.

We can relate $a$ and $b$ by

$$
a = 1 + \frac{1}{6}b + \frac{5}{6}a
$$

$$
b = 1 + \frac{5}{6}a.
$$

Indeed, in the original game, after one throw, we play the second game if we get a 6 and we replay the first game otherwise. In the second game, after the first throw, we stop if we get a 6 and play the first game otherwise. Hence, $a = 42$.

**Exercice 8 - Stagiaire L3**

Bob wants to recruit a probationer (fr. stagiaire) L3 among $n$ candidates and naturally, he wants to recruit the best one. The candidates present themselves on an interview one by one in a random order. During the interview, Bob grades a candidate (all scores are different). The rule of the game is the following : after the interview, Bob either had to hire the candidate or had to refuse him.

Skillful Bob uses the following strategy : first, interview $m$ candidates and refuse them all ; next, after the $m$-th candidate, hire the first candidate that has a better score than all previously interviewed.

8.1 Show that the probability that Bob chooses the best candidate is

$$
P(n,m) = \frac{m}{n} \sum_{j=m+1}^{n} \frac{1}{j-1}
$$
Réponse : Let $E_j$ denote the event that "the $j$-th candidate is the best and is being recruited". First, we argue that $P(E_j) = \frac{1}{n} \frac{m}{j-1}$ for $j > m$ (and clearly 0, otherwise). Indeed, the $j$-th candidate is the best with probability $\frac{1}{n}$. Also, for $E_j$ to happen, it requires that the best candidate among $j-1$ first candidates appears in the group of first $m$ rejected candidates (s.t. we do not choose him). This happens with probability $\frac{m}{j-1}$. The result follows from the fact that $P(m, n) = P(\cup_{j=m+1}^{n} E_j) = \sum_{j=m+1}^{n} P(E_j)$.

8.2 Conclude that $\lim_{n \to \infty} \max_{m} P(n, m) \geq \frac{1}{e}$.

Réponse : Computing the above sum by the integral, we obtain for the limit

$$\frac{m}{n} (\ln(n) - \ln(m)) \leq P(n, m) \leq \frac{m}{n} (\ln(n-1) - \ln(m-1)).$$

The function $\frac{\ln(x)}{x}$ attains its maximum at $x = e$. For $m = \frac{n}{e}$, we obtain $\lim_{n \to \infty} \max_{m} P(n, m) \geq \frac{1}{e}$.

Exercice 9 - Variante de la distribution géométrique

The geometric distribution arises as the distribution of the number of times we flip a coin until it comes up heads. Consider now the distribution of the number of flips $X$ until the $k$-th head appears, where each coin flip comes up heads independently with probability $p$. This is known as the negative binomial distribution.

9.1 Provide a formula for the distribution of this random variable.

9.2 Find the expectation of this distribution.