

Epreuve - Matière : 102 - 9423

Session : 2024

CONSIGNES

- Remplir soigneusement, sur CHAQUE feuillet officiel, la zone d'identification en MAJUSCULES.
- Remplir soigneusement le cadre relatif au concours OU à l'examen qui vous concerne.
- Ne pas signer la composition et ne pas y apporter de signe distinctif pouvant indiquer sa provenance.
- Rédiger avec un stylo à encre foncée (bleue ou noire) et ne pas utiliser de stylo plume à encre claire.
- N'effectuer aucun collage ou découpage de sujets ou de feuillet officiel.
- Numéroté chaque PAGE (cadre en bas à droite de la page) sur le nombre total de pages que comporte la copie (y compris les pages vierges).
- Placer les feuilles dans le bon sens et dans l'ordre de numérotation des pages.

Partie I. Définitions et propriétés élémentaires

Question 1

a) 13444 Les transitions $1 \rightarrow 3$, $3 \rightarrow 4$ et $4 \rightarrow 4$ sont bien de probabilités non nulles, ainsi que $P(1)$, la probabilité de commencer en 1. Donc cette marche est possible dans H_1 , avec une probabilité de : $1 \times 0.5 \times 0.3 \times 1^2 = 0.15$.

b) 239 Cette marche n'est pas possible dans H_1 car $P(2) = 0$.

c) 132323232 Cette marche est possible, avec probabilité (dans H_1) :
 $1 \times 0.5 \times 0.2 \times (0.8 \cdot 0.2)^3 = 0.1 \times 0.004096$
 $= 0.0004096 = 4,096 \cdot 10^{-4}$.

d) 1234 Cette marche n'est pas possible dans H_1 car $T(1,2) = 0$.

e) 13432 Cette marche n'est pas possible dans H_1 car $T(4,3) = 0$.

Question 2

a) AATATA Cette séquence peut être produite par la marche 134444 dans H_1 .

b) CATGTG Cette séquence ne peut pas être produite dans H_1 , car le premier état est forcément 1 et n'émet que A.

c) AACGAA Cette séquence peut être produite par 131111 dans H_1 .

d) AAGGGG Cette séquence peut être produite par 134444 dans H_1 .

Question 3

a) AGCGAT peut être produit par 134444 dans H_1 , avec probabilité

$$\text{Prob}_{H_1}(134444) \times \text{Prob}_{134444, H_1}(AGCGAT) = 0.15 \times (1 \times \frac{1}{2} \times \frac{1 \times 2 \times 3 \times 4}{10^4})$$

$$= 1.8 \cdot 10^{-4}$$

b) AGAAA peut être produit par 13444 dans H_1 , avec probabilité

$$\text{Prob}_{H_1}(131111) \times \text{Prob}_{13444, H_1}(AGAAA) = 0.15 \times \frac{1}{2} \times (0.2)^3$$

$$= 6 \cdot 10^{-4}$$

Question 4

- D'après ma réponse à (Q2), la marche 134444 peut produire plusieurs séquences dans H_1 , comme par exemple AACGAA et AAGGGG.
- On peut montrer qu'un modèle H_1 a une séquence pour chaque marche possible si et seulement si $\forall i \in [1, n], \forall c \in \Sigma, E(i, c) \in \{0, 1\}$.

Sens direct : On considère une marche quelconque, elle n'a par hypothèse qu'une seule séquence produite possible.

Donc cette marche m , et sa séquence c vérifient $\prod_{i=1}^k E(m_i, c_i) = 1$

et donc $\forall i \in [1, k], E(m_i, c_i) = 1$.

Ainsi $\forall i \in [1, k],$ si $a \neq c_i$ alors $E(m_i, a) = 0$

Comme tous les états sont accessibles, on peut étendre ceci à tous les états du modèle.

Sens indirect: On considère une marche quelconque m , une séquence qu'elle peut produire c .

On a $\prod_{i=1}^k E(m_i, c_i) > 0$ Or $\forall i \in [1, k] E(m_i, c_i) \in \{0, 1\}$

Donc $\prod_{i=1}^k E(m_i, c_i) = 1$, alors elle ne peut produire que cette séquence.

Question 5

La séquence AA peut être produite par 11 et 13 dans H_1 , avec probabilités respectives (sachant la marche) 1 et 0.5.

Question 6

AAG peut être produite dans H_1 par les marches :

- 113 avec probabilité $\text{Prob}_{H_1}(113, \text{AAG}) = 1/8$
- 133 ... $\text{Prob}_{H_1}(133, \text{AAG}) = 1/16$
- 132 $\text{Prob}_{H_1}(132, \text{AAG}) = 1/40$
- 134 $\text{Prob}_{H_1}(134, \text{AAG}) = 3/100$

Donc 113 maximise cette probabilité.

Question 7

Je ne peux que l'interpréter via le contexte, sans démonstration.

Pour une séquence w , $\sum \text{Prob}_H(m, w)$

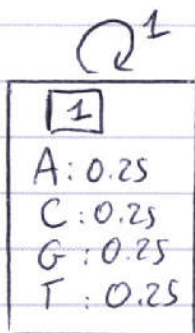
on imagine le modèle H émettre des lettres à l'infini sans s'arrêter, produisant un mot infini W , ~~alors w est un préfixe de W .~~ est la probabilité que W ait pour préfixe w .

Nous sommes obligés d'envisager des mots infinis car le modèle ne prend pas en compte l'arrêt du mot.

Ainsi $\text{Prob}_H(m)$ ne désigne pas la probabilité de "suivre une marche" mais de "commencer par cette marche".

Exemple : les marches 1, 11 ont des probabilités qui somment à 1.5, ce qui est plus grand que 1.

Question 8



est la représentation sous forme de graphe de H_0 .

$$\begin{aligned} \text{On a} \quad & \bullet \sum_{i=1}^4 P(i) = P(1) = 1 \\ & \bullet \sum_{j=1}^4 T(1, j) = T(1, 1) = 1 \\ & \bullet \sum_{c \in \Sigma} E(1, c) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = 1 \end{aligned}$$

Donc (1), (2) et (3) sont bien vérifiées.

Question 9

Pour tout $k \in \mathbb{N}^*$, on considère la séquence $u = A^k$ et la marche $m = 1^k$.

$$\begin{aligned} \text{On a } \text{Prob}_{H_0}(m, u) &= \text{Prob}_{H_0}(m) \cdot \text{Prob}_{m, H_0}(u) = P(1) \times \prod_{i=1}^{k-1} T(m_i, m_{i+1}) \cdot \prod_{i=1}^k E(m_i, u_i) \\ &= 1 \times \left(\prod_{i=1}^{k-1} 1 \right) \cdot \prod_{i=1}^k \frac{1}{4} \\ &= \left(\frac{1}{4} \right)^k > 0 \end{aligned}$$

Epreuve - Matière : 102 - 0923

Session : 2024

CONSIGNES

- Remplir soigneusement, sur CHAQUE feuillet officiel, la zone d'identification en MAJUSCULES.
- Remplir soigneusement le cadre relatif au concours OU à l'examen qui vous concerne.
- Ne pas signer la composition et ne pas y apporter de signe distinctif pouvant indiquer sa provenance.
- Rédiger avec un stylo à encre foncée (bleue ou noire) et ne pas utiliser de stylo plume à encre claire.
- N'effectuer aucun collage ou découpage de sujets ou de feuillet officiel.
- Numéroté chaque PAGE (cadre en bas à droite de la page) sur le nombre total de pages que comporte la copie (y compris les pages vierges).
- Placer les feuilles dans le bon sens et dans l'ordre de numérotation des pages.

On en déduit que H_0 peut produire la séquence A^k et ce pour tout $k \in \mathbb{N}^*$
 ainsi H_0 peut produire des séquences de longueur arbitraire.

Question 10

On considère une séquence non vide, de longueur $k \in \mathbb{N}^*$, notée w .
 Ainsi que la marche $m = 1^n$

On a $\text{Prob}_{H_0}(w)$

$$\text{On a } \text{Prob}_{H_0}(m, w) = \text{Prob}_H(m) \times \text{Prob}_{w, H}(w)$$

$$= P(1) \times \prod_{i=1}^{k-1} T(m_i, m_{i+1}) \times \prod_{i=1}^k E(m_i, w_i)$$

$$= P(1) \times \prod_{i=1}^{k-1} T(1, 1) \times \prod_{i=1}^k E(1, w_i)$$

$$= P(1) \times (1)^{k-1} \times \left(\frac{1}{4}\right)^k$$

$$= \left(\frac{1}{4}\right)^k > 0$$

Donc H_0 peut produire n'importe quelle séquence non vide de Σ .

Question 11

(a) Par définition, on a $|\Sigma_o^k| = |\Sigma_o|^k = 4^k$
 Il y a donc 4^k séquences possibles dans Σ_o^k

(b) On a montré en question 10 que :

$$\forall w \in \Sigma_o^k, \text{Prob}_H(1^k, w) = \left(\frac{1}{4}\right)^k$$

Or une marche de longueur k ne produit que des séquences de longueur k

Donc pour tout $w \in \Sigma_o^k$, seule une marche de longueur k peut produire w . Or il n'y a qu'une seule marche de longueur k dans H_o qui est 1^k .

Finalement la probabilité d'une séquence de Σ_o^k dans H_o est $\frac{1}{4^k}$

(c) On a $\sum_{w \in \Sigma_o^k} \text{Prob}_H(w) = \sum_{w \in \Sigma_o^k} \left(\frac{1}{4}\right)^k = 4^k \cdot \left(\frac{1}{4}\right)^k = 1$
on s'autorise cette notation, bien qu'elle n'est définie que pour des marches.

La somme des probabilités des séquences de Σ_o^k dans H_o est donc 1

Question 12

H_1 ne peut produire qu'une seule séquence de taille 1, qui est A.
 En effet, H_1 ne peut y a qu'une seule marche de longueur 1 dans H_1 qui est 1, qui ne peut produire que A avec probabilité 1

Séquences: A

Probabilités: 1

Somme des probabilités: 1

Question 13

~~les séquences de taille 2 possibles~~

les mots de taille 2 possibles dans H_1 sont 11 et 13
de probabilités respectives $\frac{1}{2}$ et $\frac{1}{2}$

11 produit donc AA avec probabilité $\frac{1}{2}$

13 produit AA avec probabilité $\frac{1}{4}$ et AG avec la même probabilité

Finalement,

Séquences: AA AG

Probabilités: $\frac{3}{4}$ $\frac{1}{4}$

Somme des probabilités: 1

Question 14

On considère la séquence $w = A^{k+3}$ et la marche $m = 13(4^{k+1})$ pour $k \in \mathbb{N}$

$$\begin{aligned} \text{On a } \text{Prob}_H(m, w) &= P(1) \prod_{i=1}^{k+3} T(m_i, m_{i+1}) \prod_{i=1}^{k+2} E(m_i, A) \\ &= 1 \times T(1, 3) \times T(3, 4) \times (T(4, 4))^k \times E(1, A) E(3, A) (E(4, A))^{k+1} \\ &= 0.9 \times 0.3 \times 1^k \times 1 \times 0.5 \times (0.2)^{k+1} > 0 \end{aligned}$$

On en déduit que H_1 peut produire des séquences de longueur arbitraire

Question 15

- Pour qu'un modèle de Markov caché puisse produire des séquences arbitrairement longues, il suffit que le graphe qui le représente ait un cycle
- Pour montrer que le graphe représentant H_1 a un cycle, on remarque d'abord que l'équation (2) impose que tous les sommets de ce graphe possèdent un arc sortant

On montre alors par récurrence sur $k \in \mathbb{N}^*$ que pour tout k : "il existe une marche de longueur k dans un chemin de longueur k dans le graphe de H "

- Pour $k=1$, on utilise le fait que d'après (1), $\sum_{i=1}^n P(i) = 1$, donc il existe i_0 tel que $P(i_0) > 0$
- Pour l'hérédité, on considère un chemin de taille k (hypothèse de récurrence) $s_1 \cdot s_2 \dots s_k$. d'après (3), le sommet s_k a un arc sortant vers un sommet a , que l'on note s_{k+1} .
 $s_1 \dots s_{k+1}$ est un chemin de longueur $k+1$ dans le graphe de H .

Question 1b

On veut montrer par récurrence sur k que :

$$\sum_{w \in \Sigma^k} \sum_{m \in S^k} \text{Prob}_H(m, w) = 1$$

- Pour $k=1$ cette valeur devient
$$\begin{aligned} & \sum_{w \in \Sigma^1} \sum_{m \in S^1} \text{Prob}_H(m, w) \\ &= \sum_{c \in S} \sum_{i \in S} P(c) E(i, c) \\ &= \sum_{c \in S} P(c) \left(\sum_{i \in S} E(i, c) \right) \\ &= \sum_{c \in S} P(c) \quad (\text{d'après (3)}) \\ &= 1 \quad (\text{d'après (1)}) \end{aligned}$$

• Supposons cette propriété vérifiée au rang $k \in \mathbb{N}^*$, on veut montrer que

$$\sum_{w \in \Sigma^{k+1}} \sum_{m \in S^{k+1}} \text{Prob}_H(m, w) = 1$$

L'idée qui nous est de calculer cette somme en distinguant selon la dernière lettre de w et le dernier état de m .

Epreuve - Matière : 102 - 9423

Session : 2024

CONSIGNES

- Remplir soigneusement, sur CHAQUE feuillet officiel, la zone d'identification en MAJUSCULES.
- Remplir soigneusement le cadre relatif au concours OU à l'examen qui vous concerne.
- Ne pas signer la composition et ne pas y apporter de signe distinctif pouvant indiquer sa provenance.
- Rédiger avec un stylo à encre foncée (bleue ou noire) et ne pas utiliser de stylo plume à encre claire.
- N'effectuer aucun collage ou découpage de sujets ou de feuillet officiel.
- Numéroté chaque PAGE (cadre en bas à droite de la page) sur le nombre total de pages que comporte la copie (y compris les pages vierges).
- Placer les feuilles dans le bon sens et dans l'ordre de numérotation des pages.

$$\begin{aligned}
 \text{On a } \sum_{w \in \Sigma^{k+1}} \sum_{m \in S^{k+1}} \text{Prob}_H(m, w) &= \sum_{w \in \Sigma^k} \sum_{m \in S^k} \sum_{a \in \Sigma} \sum_{q \in S} \text{Prob}_H(mq, wa) \\
 &= \sum_w \sum_m \sum_a \sum_q P(m_0) \left(\prod_{i=1}^{k-1} T(m_i, m_{i+1}) \right) T(m_k, q) \left(\prod_{i=1}^k E(m_i, w_i) \right) E(q, a) \\
 &= \sum_w \sum_m \left[P(m_0) \prod_{i=1}^{k-1} T(m_i, m_{i+1}) \prod_{i=1}^k E(m_i, w_i) \right] \sum_{a \in \Sigma} \sum_{q \in S} T(m_k, q) E(q, a) \\
 &= \sum_w \sum_m \text{Prob}_H(m[1 \dots k], w[1 \dots k]) \sum_{a \in \Sigma} \sum_{q \in S} T(m_k, q) E(q, a) \\
 &= \sum_w \sum_m \left(\text{Prob}_H(m[1 \dots k], w[1 \dots k]) \underbrace{\sum_{q \in S} T(m_k, q)}_1 \right) \\
 &= \sum_w \sum_m \text{Prob}_H(m[1 \dots k], w[1 \dots k]) = 1 \quad (\text{hypothèse de récurrence})
 \end{aligned}$$

Partie II Représentation des modèles de Markov cachésQuestion 17

La structure d'un objet HMM pourrait être la suivante :

Comme attributs :

- les valeurs m et n

- un tableau alp de longueur m représentant Σ (pourrait être remplacé par un ensemble)
- un tableau dep de longueur n représentant P
- une matrice trans de taille $n \times n$ représentant T
- une liste de m dictionnaires appelé émission telle que $\text{émission}[i][a] = E(i, a)$

Comme méthodes :

- une méthode pr-marche prenant en entrée une liste d'entiers de 1 à n et renvoyant la probabilité $\text{Prob}_H(m)$

- une méthode pr-marche-seq prenant en entrée une liste d'entiers de 1 à n représentant une marche m , ainsi qu'une liste de lettres de Σ représentant une séquence w , renvoyant $\text{Prob}_H(m, w)$

Question 18 :

class HMM:

```
def __init__(self, m, n, alp, dep, trans, emission):
    m, n = m, n
    alp = alp.copy()
    dep = dep.copy()
    trans = [trans[i].copy() for i in range(n)]
    emission = [emission[i].copy() for i in range(n)]
```



```

def print_param(self):
    print("nombre d'états: {}".format(m))
    print("taille de l'alphabet: {}".format(n))
    print("alphabet: {}".format(alp))
    print("probabilités de départ: {}".format(dep))
    print("probabilités de transitions: {}".format(trans))
    print("probabilités d'émissions: {}".format(emission))

```

```

def pr_marche(self, mch):
    pr = dep[mch[0]]
    for i in range(len(mch) - 1):
        pr *= trans[mch[i]][mch[i+1]]
    return pr

```

```

def pr_marche_seq(self, mch, seq):
    pr = 1
    for i in range(len(mch)):
        pr *= emission[mch[i]][seq[i]]
    return pr * self.pr_marche(mch)

```

Question 19

Il suffit de remplacer les arguments par

```

def __init__(self, m=4, n=1, alp=['A','C','G','T'], dep=[1], trans=[[]],
            emission=[{'A':0.25,'C':0.25,'G':0.25,'T':0.25}]):
    ...

```

En donnant des arguments par défaut.

Question 20:

```

m, n = 4, 4
alp = ['A', 'C', 'G', 'T']
dep = [1, 0, 0, 0]
trans = [[0.5, 0, 0.5, 0], [0, 0.1, 0.8, 0.1], [0, 0.2, 0.5, 0.3], [0, 0, 0, 1]]
emission = [{ 'A': 1, 'C': 0, 'G': 0, 'T': 0 }, { 'A': 0, 'C': 0, 'G': 0.5, 'T': 0.5 },
             { 'A': 0.5, 'C': 0, 'G': 0.5, 'T': 0 }, { 'A': 0.2, 'C': 0.1, 'G': 0.4, 'T': 0.3 } ]
H1 = HMM(m, n, alp, dep, trans, emission)

```

Question 21

```
def __eq__(self, other):  
    if (self.n != other.n) or (self.m != other.m):  
        return False  
    for c in self.alp:  
        if not (c in other.alp):  
            return False  
    for i in range(n):  
        for j in range(m):  
            if self.trans[i][j] != other.trans[i][j]:  
                return False  
    if self.dep[i] != other.dep[i]:  
        return False  
    for c in self.alp:  
        if self.emission[i][c] != other.emission[i][c]:  
            return False  
    return True
```

Annotations in red:

- n, m (next to the first if statement)
- vérifier l'alphabet (next to the first for loop)
- vérifier trans (next to the nested for loops)
- vérifier dep (next to the if statement for dependencies)
- vérifier émission (next to the for loop for emission probabilities)

Question 22

(la Q ne demande pas de vérifier que les probabilités sont dans $[0, 1]$)

```
def est_val estValide(self):  
    if sum(dep) != 1:  
        return False  
    for i in range(n):  
        if (sum(trans[i]) != 1) or (sum(emission[i].values()) != 1):  
            return False  
    return True
```

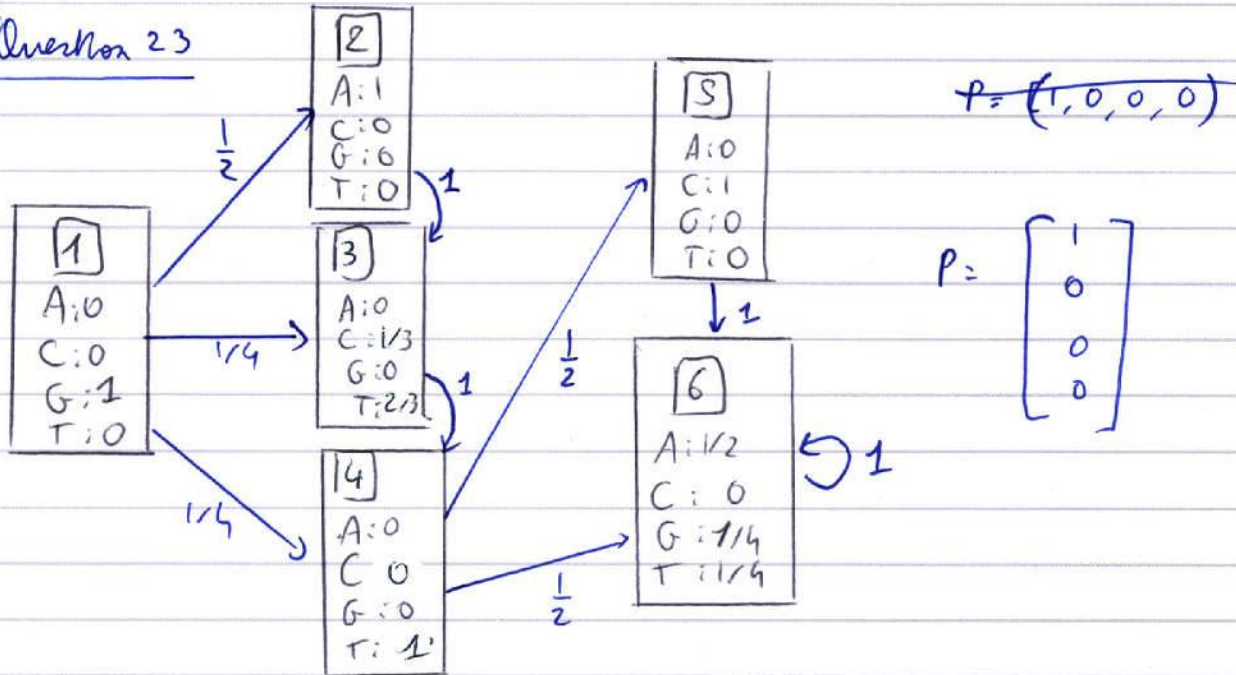

Epreuve - Matière : 102 - 9523

Session : 2024

CONSIGNES

- Remplir soigneusement, sur CHAQUE feuillet officiel, la zone d'identification en MAJUSCULES.
- Remplir soigneusement le cadre relatif au concours OU à l'examen qui vous concerne.
- Ne pas signer la composition et ne pas y apporter de signe distinctif pouvant indiquer sa provenance.
- Rédiger avec un stylo à encre foncée (bleue ou noire) et ne pas utiliser de stylo plume à encre claire.
- N'effectuer aucun collage ou découpage de sujets ou de feuillet officiel.
- Numéroté chaque PAGE (cadre en bas à droite de la page) sur le nombre total de pages que comporte la copie (y compris les pages vierges).
- Placer les feuilles dans le bon sens et dans l'ordre de numérotation des pages.

Question 23



Question 29

Etant donné que les longueurs sont fixées, une liste de liste (qui en Python sont implémentés par des tableaux qui ont une très bonne performance lorsque leur longueur ne varie pas) convient parfaitement.

On considère donc une matrice align telle que $\text{align}[i][j]$ est le jème caractère de la ième séquence.

Question 25

```
def from Alignement (self, align):
```

```
    n = len(align[0])
```

```
    m = 4
```

```
    alp = ['A', 'C', 'G', 'T']
```

```
    dep = [0] * n
```

```
    dep[0] = 1
```

```
    emission = []
```

```
    for i in range(n):
```

```
        d = {'A': 0, 'G': 0, 'C': 0, 'T': 0}
```

```
        occ = 0
```

```
        for j in range(len(align)):
            c = align[j][i]
```

```
            if c != '-':
```

```
                occ += 1
```

```
                d[c] += 1
```

on compte les occurrences de ACGT

```
        for k in dep:
```

```
            d[k] /= occ
```

on divise pour obtenir la fréquence.

```
        emission.append(d)
```

```
    trans = [[0] * n for i in range(n)]
```

```
    for i in range(n):
```

```
        ci = 0
```

```
        for l in range(len(align)):
            c = align[l][ci]
```

```
            if c != '-':
```

```
                ci += 1
```

```
                next = ci + 1
```

```
                while align[l][next] == '-':
```

```
                    next += 1
```

```
                    trans[i][next] += 1
```

```
                    trans[i][next] += 1
```


$\text{for } j \text{ in range}(n):$
 $\quad \text{trans}[i][j] /= c_i$
 $\text{trans}[n-1][n-1] = 1$

Question 26 :

• L'équation (1) est évidemment vérifiée car on a forcément $P = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$

• On veut montrer que $\sum_{j=1}^n T(i, j) = 1$ pour un i quelconque entre 1 et n

-- Si $i = n$ alors on a $T(i, i) = 1$ et $T(i, j) = 0$ pour $j \neq i$. Donc (2) est vérifiée

-- Si $i < n$, alors $T(i, j) = \frac{|\{l: \text{liges de ligne } j, l\}|}{c_i}$ ~~\neq~~

$$\text{Donc } \sum_{j=1}^n T(i, j) = \frac{1}{c_i} \cdot \sum_{j=1}^n |\{l \mid j_{i,l} = j\}|$$

Or pour chaque ligne l , $j_{i,l} \in \{1, \dots, n\}$ soit la lettre en l, i n'est pas un '-'.
 On en déduit qu'il y en a c_i , et donc que $\sum_{j=1}^n |\{l \mid j_{i,l} = j\}| = c_i$
 et finalement que

$$\sum_{j=1}^n T(i, j) = 1$$

Finalement (2) est vérifiée.

• Par définition de l'émission, (3) est vérifiée, car il s'agit de la fréquence d'apparition de caractères dans chaque colonne.

Question 27

On considère un état $i < p$

i peut transiter vers j si et seulement si il existe une ligne telle qu'il y a un gap de i à j (i.e. de longueur $j-i-1$).

On en déduit que si $j \geq i+g+1$, alors la transition (i,j) a une probabilité nulle. Finalement, i ne peut avoir qu'au plus $(g+1)$ successeurs dans le graphe de H .

Donc dans les $(p-1)$ premières lignes de T , il y a au plus $(g+1)$ entrées non nulles.

La dernière ligne n'ayant qu'une entrée non nulle (le dernier état se dirige dans lui-même), il y a en tout,

au plus $1 + (p-1)(g+1)$ entrées non nulles dans T . ($p(g+1)$ est une forme plus simple)

Question 28

Le nombre d'éléments de la matrice est p^2 (c'est le nombre d'états du modèle)

Donc la densité de la matrice de transitions est majorée par $\frac{p(g+1)}{p^2}$

~~Or n (le nombre d'états) est égal à p~~

Finalement, on obtient comme majorant $\frac{g+1}{p}$ pour la densité de la matrice

Question 29

On peut utiliser un tableau de listes d'adjacence, beaucoup moins coûteux en mémoire lorsque la matrice est peu dense.

Si une matrice est de longueur n et est de densité p on passe d'une complexité en mémoire de $O(n^2)$ à $O(n + pn^2)$

Ici $n = p$ et $p \approx \frac{g}{p}$ donc on passe de $O(p^2)$ à $O(gp)$

Concours section : AGRÉGATION EXTERNE INFORMATIQUE

Epreuve matière : Etude d'un problème informatique

N° Anonymat : N240NAT1030219

Nombre de pages : 32

17.56 / 20

Epreuve - Matière : 102 - 9923

Session : 2024

CONSIGNES

- Remplir soigneusement, sur CHAQUE feuillet officiel, la zone d'identification en MAJUSCULES.
- Remplir soigneusement le cadre relatif au concours OU à l'examen qui vous concerne.
- Ne pas signer la composition et ne pas y apporter de signe distinctif pouvant indiquer sa provenance.
- Rédiger avec un stylo à encre foncée (bleue ou noire) et ne pas utiliser de stylo plume à encre claire.
- N'effectuer aucun collage ou découpage de sujets ou de feuillet officiel.
- Numéroté chaque PAGE (cadre en bas à droite de la page) sur le nombre total de pages que comporte la copie (y compris les pages vierges).
- Placer les feuilles dans le bon sens et dans l'ordre de numérotation des pages.

Pour H_2 , on a le tableau de liste d'adjacence suivant :

[
1 → $[(2, \frac{1}{2}), (3, \frac{1}{4}), (4, \frac{1}{4})]$,
2 → $[(3, 1)]$,
3 → $[(4, 1)]$,
4 → $[(5, \frac{1}{2}), (6, \frac{1}{2})]$
5 → $[(6, 1)]$
6 → $[(6, 1)]$
]

Question 30

On peut améliorer cela avec un tableau de dictionnaires, qui donnerait

[
 $\{2: \frac{1}{2}, 3: \frac{1}{4}, 4: \frac{1}{4}\}$,
 $\{3: 1\}$,
 $\{4: 1\}$,
 $\{5: \frac{1}{2}, 6: \frac{1}{2}\}$,
 $\{6: 1\}$,
 $\{6: 1\}$,
]

17 / 32

Question 30

En utilisant un tableau de dictionnaires, il y a énormément d'avantages :

① Le code ne change pas beaucoup

en effet, on accède toujours à $T[i][j]$ via la syntaxe `trans[i][j]`

On peut les remplacer par des `get` de la librairie `dict` de python afin d'utiliser 0 comme valeur par défaut si la clé ne se trouve pas dans le dictionnaire.

Donc on remplace `trans[i][j]` par `trans[i].get(j, 0)`

② La complexité en moyenne reste la même.

Les dictionnaires python sont implémentés par des tables de hachage dont la complexité moyenne d'accès à un élément est en $O(1)$ (le pire cas est linéaire) au coût d'un facteur constant de mémoire gachée.

Partie III Algorithmes

Question 3)

On considère w_1, \dots, w_k , et m_1, \dots, m_k une marche de probabilité maximale produisant w .
On raisonne par l'absurde.

Supposons qu'il existe $k' \leq k$ telle que $m_1, \dots, m_{k'}$ a une plus faible probabilité que $s_1, \dots, s_{k'}$ avec $s_{k'} = m_{k'}$ et que $s_1, \dots, s_{k'}$ produit aussi $w_1, \dots, w_{k'}$.

Alors on considère la marche

$s_1, \dots, s_{k'}, m_{k'+1}, \dots, m_k$

la probabilité que cette marche produise w est

$$P_1 = P(s_1) \left(\prod_{i=1}^{k'-1} T(s_i, s_{i+1}) \right) T(m_{k'}, m_{k'+1}) \left(\prod_{i=k'+1}^{k-1} T(m_i, m_{i+1}) \right) \prod_{i=1}^{k'} E(s_i, w_i)$$

et celle de

m_1, \dots, m_k est $P(m_1) \prod_{i=1}^k E(m_i, w_i)$

$$P_2 = P(m_1) \prod_{i=1}^{k'-1} T(m_i, m_{i+1}) \prod_{i=k'+1}^k E(m_i, w_i)$$

$$\text{On a donc } \frac{P_1}{P_2} = \frac{P(s_1) \prod_{i=1}^{k'-1} T(s_i, s_{i+1}) \prod_{i=1}^{k'} E(s_i, w_i)}{P(m_1) \prod_{i=1}^{k'-1} T(m_i, m_{i+1}) \prod_{i=1}^{k'} E(m_i, w_i)}$$

Donc $\frac{P_1}{P_2} > 1$, car $s_1, \dots, s_{k'}$ est la marche qui a le plus de chance de produire $w_1, \dots, w_{k'}$.

On a donc une marche $s_1, \dots, s_{k'}, m_{k'+1}, \dots, m_k$ qui a une plus grande probabilité que m de produire w .

C'est absurde, on en déduit la propriété voulue.

Question 32

(a) On a $\forall j \in \mathcal{N}, \delta[1, j] = E(j, w_1) \cdot P(j)$ car pour produire w_1 , il faut que la marche soit de longueur 1. Ainsi la probabilité d'une marche de longueur 1 terminant dans l'état j est $P(j)$, et donc celle qu'elle produise w_1 est $P(j) E(j, w_1)$.

(b) On veut déterminer $\delta[i, j]$ pour un certain i et j .

On considère une marche $m_1 \dots m_i$ terminant dans l'état j produisant $w_1 \dots w_i$ (m est aussi de longueur i) avec probabilité maximale.

On sait que, d'après Q31, $m_1 \dots m_{i-1}$ est la marche qui a la probabilité maximale de produire $w_1 \dots w_{i-1}$, parmi celles terminant dans l'état m_{i-1} .

Donc cette marche a pour probabilité $\delta[i-1, m_{i-1}]$ de produire $w_1 \dots w_{i-1}$. La marche $m_1 \dots m_i = m$ a pour probabilité de produire $w_1 \dots w_i$ la valeur

$$P(m_1) \prod_{a=1}^{i-1} T(m_a, m_{a+1}) \prod_{a=1}^i E(m_a, w_a) = \delta[i-1, m_{i-1}] \cdot T[m_{i-1}, m_i] \cdot E[m_i, w_i]$$

Ainsi on obtient la propriété suivante.

Il existe un état j_m tel que

$$\delta[i, j] = \delta[i-1, j_m] \cdot T[j_m, j] \cdot E[j, w_i]$$

Par maximalité de $\delta[i, j]$ on a

$$\delta[i, j] = \max_{j_m \in \mathcal{N}} \delta[i-1, j_m] \cdot T[j_m, j] \cdot E[j, w_i]$$

On peut donc remplir la i ème ligne de S avec la $i-1$ ème et la matrice de transition (et d'émission)

17.56 / 20

Epreuve - Matière : 102 - 4423

Session : 2024

CONSIGNES

- Remplir soigneusement, sur CHAQUE feuillet officiel, la zone d'identification en MAJUSCULES.
- Remplir soigneusement le cadre relatif au concours OU à l'examen qui vous concerne.
- Ne pas signer la composition et ne pas y apporter de signe distinctif pouvant indiquer sa provenance.
- Rédiger avec un stylo à encre foncée (bleue ou noire) et ne pas utiliser de stylo plume à encre claire.
- N'effectuer aucun collage ou découpage de sujets ou de feuillet officiel.
- Numéroté chaque PAGE (cadre en bas à droite de la page) sur le nombre total de pages que comporte la copie (y compris les pages vierges).
- Placer les feuilles dans le bon sens et dans l'ordre de numérotation des pages.

(c) On définit $S[k, j]$ est la probabilité maximale qu'une marche terminant en j produise w .

Donc la probabilité maximale qu'une marche quelconque produise w est

$$\max_{j \in \{1, n\}} S[k, j]$$

(cette valeur est donnée case du tableau de la dernière ligne du tableau, mais on ne sait à priori pas laquelle)

Question 33 $w = AGG$

$i \backslash j$	1	2	3	4
1	1	0	0	0
2	0	0	1/4	0
3	0	1/40	1/16	3/100

tableau S

$i \backslash j$	1	2	3	4
1	0	0	0	0
2	?	?	1	?
3	?	3	3	3

tableau b

? désigne que la case n'a pas de valeur déterminée.

l'algorithme calcule Set b, ensuite

l'algorithme commence par déterminer $z[3]$, qui est ici 3 (encadré en rouge dans (la colonne qui maximise la dernière ligne de S) puis construit z à l'envers en remontant via b.

On obtient alors $z = (1, 3, 3)$

Question 34

Ma réponse à la Q32 prouve la correction du calcul du tableau S .

Il en est de même pour le calcul de b . En effet, si on considère i et j , et qu'on veut calculer $S[i, j]$, l'indice j_m maximisant la probabilité voulue que j'ai énoncé en Q32 est ce qui est calculé dans $b[i, j]$.

Pour rappel, on a montré que si m_1, \dots, m_i est une marche qui a la probabilité maximale de produire w_1, \dots, w_i en terminant dans l'état j , alors $m_{i-1} = j_m = \arg \max_{l \in S} \{S[i-1, l] \cdot T(l, j) \cdot E[j, w_i]\}$.

Or $E(j, w_i)$ ne dépend pas de l et est positif.

$$\text{Donc } \underline{m_{i-1} = \arg \max_{l \in S} \{S[i-1, l] T(l, j)\} = b[i, j] = b[i, m_i]}$$

On a aussi prouvé que $m_k = \arg \max_{l \in S} S[k, l]$.

Finalement, on remarque qu'en ligne 16, l'algorithme calcule bien m_k .

~~Et en ligne~~

Pour la boucle en ligne 17, on a l'invariant suivant:

"Pour tout $j \in [i-1, k]$, il existe une marche terminant par $z[i-1] z[i] \dots z[k]$ qui maximise la probabilité de produire w ."

- Avant le premier tour de boucle, on considère $i = k+1$ (la boucle fait décroître i)
on a donc $z[k] = \arg \max_{l \in S} S[k, l] = m_k$ où m_1, \dots, m_k

Or on a montré qu'il existait un m maximal finissant dans cet état.

- Quand on passe de i à $i-1$, on pose $z[i-1] = b[i, z[i]]$
Il existe une marche m telle maximale telle que m remise
par $z[i]$, $z[i+1]$, ..., $z[k]$ (hypothèse d'invariant)

On a donc (montré plus haut) $m_{i-1} = \arg \max \{S[i-1, l] + (l, j')\} = b[i, m_i]$

Or $m_i = z[i]$ donc $m_{i-1} = z[i-1]$

Finalement, z est bien une marche maximisant la probabilité de
prothine w .

Question 35

Toutes les opérations élémentaires (sauf max et argmax) sont en $O(1)$
Donc :

la boucle entre la ligne 17 et la ligne 19 est en $O(k)$

la ligne 16 est en $O(|S|) = O(n)$

les boucles imbriquées des lignes 10 à 15 font $O(kn)$ nous choisissons
en $O(|S|) = O(n)$ à cause des appels à max et argmax.

le reste est en $O(n)$

Finalement, $Viterbi(k, w)$ est en $O(kn^2)$

Question 36

la probabilité que le modèle H_1 émette la séquence AAG est de

$$\frac{1}{8} + \frac{1}{16} + \frac{1}{40} + \frac{3}{100} = \frac{50 + 25 + 10 + 12}{400} = \frac{97}{400}$$

Question 37

$$\begin{aligned} \text{On a } \alpha(i, j) &= \sum_{\substack{m_1, \dots, m_i \in S^i \\ m_i = j}} \text{Prob}_{m, H}(w_1, \dots, w_i) \text{Prob}_H(m) \\ &= \sum_{\substack{m_1, \dots, m_i \in S^i \\ m_i = j}} \prod_{a=1}^i E(m_a, w_a) \text{Prob}_H(m) \\ &= \sum_{\substack{m_1, \dots, m_i \in S^i \\ m_i = j}} \text{Prob}_{m[i-1], H}(w_1, \dots, w_{i-1}) \cdot E(m_i, w_i) \text{Prob}_H(m) \\ &= E(j, w_i) \sum_{\substack{m_1, \dots, m_{i-1} \in S^{i-1} \\ m_{i-1} = j}} \text{Prob}_{m[i-1], H}(w_1, \dots, w_{i-1}) \text{Prob}_H(m[i-1]) T(m_{i-1}, j) \\ &= E(j, w_i) \sum_{m_1, \dots, m_{i-1} \in S^{i-1}} \text{Prob}_{m, H}(w_1, \dots, w_{i-1}) \text{Prob}_H(m) T(m_{i-1}, j) \\ &= E(j, w_i) \sum_{q \in S} \sum_{\substack{m_1, \dots, m_{i-1} \in S^{i-1} \\ m_{i-1} = q}} \text{Prob}_{m, H}(w_1, \dots, w_{i-1}) \text{Prob}_H(m) T(m_{i-1}, j) \\ &= E(j, w_i) \sum_{q \in S} T(q, j) \sum_{\substack{m_1, \dots, m_{i-1} \in S^{i-1} \\ m_{i-1} = q}} \text{Prob}_{m, H}(w_1, \dots, w_{i-1}) \text{Prob}_H(m) \\ &= E(j, w_i) \sum_{q \in S} T(q, j) \alpha(i-1, q) \end{aligned}$$

Donc $\alpha(i, j) = \sum_{q \in S} \alpha(i-1, q) E(j, w_i) T(q, j)$

Epreuve - Matière : 102-9423

Session : 2024

CONSIGNES

- Remplir soigneusement, sur CHAQUE feuillet officiel, la zone d'identification en MAJUSCULES.
- Remplir soigneusement le cadre relatif au concours OU à l'examen qui vous concerne.
- Ne pas signer la composition et ne pas y apporter de signe distinctif pouvant indiquer sa provenance.
- Rédiger avec un stylo à encre foncée (bleue ou noire) et ne pas utiliser de stylo plume à encre claire.
- N'effectuer aucun collage ou découpage de sujets ou de feuillet officiel.
- Numéroté chaque PAGE (cadre en bas à droite de la page) sur le nombre total de pages que comporte la copie (y compris les pages vierges).
- Placer les feuilles dans le bon sens et dans l'ordre de numérotation des pages.

Nous n'avons plus besoin ni de b ni de z car on veut juste renvoyer la valeur et non une marche.

On remplace S par α , avec $\alpha[1, j] = P(j) \times E(j, w_i)$ (pareil que $V[i, j]$)

et $\alpha[i, j] = E(j, w_i) \times \sum_{l \in S} (T(l, j) \times \alpha[i-1, l])$

1 Début

2 Soit k la taille de w et n la taille de S

3 Soit $S[1 \dots k, 1 \dots n]$ un tableau bidimensionnel

4 Pour $j = 1$ à n

5 $\alpha[1, j] = P(j) \times E(j, w_i)$

6 Fin Pour

7 Pour $i = 2$ à k

8 Pour $j = 1$ à n

9 $S[i, j] = 0$

10 Pour $l = 1$ à n

11 $S[i, j] = S[i, j] + E(j, w_i) \times T(l, j) \times \alpha[i-1, l]$

12 Fin Pour

13 Fin Pour

14 Fin Pour

15

Question 38

On a par une démonstration antérieure à celle de (Q37) (un peu plus simple) que cette probabilité est $\sum_{l=1}^n \alpha[k, l]$

Question 39

Forward calcule les probabilités de produire $w_1 \dots w_i$ en incrémentant i , alors que Backward calcule la probabilité d'émettre $w_i \dots w_k$ en décrémentant i , à partir d'un certain état j .

Notamment, $B(i, j)$ est la probabilité qu'on de produire $w_i \dots w_k$ à partir de l'état j .

On a bien $B(k, j) = E(j, w_k)$ selon cette définition par exemple

Question 40

On a que $B[1, j]$ est la probabilité de produire w en partant de l'état j .

Donc la probabilité de produire w , depuis n'importe quel état est

$$\sum_{j=1}^n P(j) B[1, j]$$

⚠ dans back ward, la première ligne est déjà multipliée par les valeurs $P(j)$
Donc il faut lire

$$\sum_{j=1}^n B[1, j]$$

Question 41

La complexité en temps de Backward et Forward est la même que celle de Viterbi ($O(kn^2)$) car on a juste remplacé $\arg\max$, qui était en $O(n)$ par une boucle faisant des additions/multiplications.

Noter qu'on suppose ici que l'accès aux valeurs de $\tau(l, j)$ etc... est en $O(1)$. On verra que c'est le cas en Python.

Pour la complexité en espace, on n'utilise pas plus qu'un tableau bidimensionnelle de longueur k et largeur n .

Donc la complexité en espace de Forward et Backward est en $O(nk)$, tout comme Viterbi.

Question 42 (On suppose `trons` après est un tableau de dictionnaires)

```
def backward(self, w):  
    k = len(w)  
    Beta = [[0] * n for _ in range(k)]  
    for j in range(n):  
        Beta[k-1][j] = emission[j][w[k-1]]  
  
    for i in range(k-2, -1, -1):  
        for j in range(n):  
            Beta[i][j] = 0  
            for l in range(n):  
                Beta[i][j] += emission[j][w[i]] * Beta[i+1][l]  
                * trons[j].get(l, 0)  ← renvoie 0 si l  
                                   n'est pas une clé  
                                   de trons[j]  
  
    for j in range(n):  
        Beta[0][j] *= dep[j]  
  
    return Beta
```

Par 14 Pour aller plus loin

Question 43 On note E' la nouvelle matrice d'émission.

$$\begin{aligned}\text{Soit } i \in [1, n], \text{ on a } \sum_{c \in \Sigma} E'(i, c) &= \sum_{\substack{c \in \Sigma \\ E(i, c) = 0}} E'(i, c) + \sum_{\substack{c \in \Sigma \\ E(i, c) > 0}} E'(i, c) \\ &= \sum_{\substack{c \in \Sigma \\ E(i, c) = 0}} \frac{1}{n \cdot n_i} + \sum_{\substack{c \in \Sigma \\ E(i, c) > 0}} \left(\frac{n-1}{n}\right) E(i, c)\end{aligned}$$

$$= \frac{1}{n \cdot n_i} |\{c \in \Sigma \mid E(i, c) = 0\}| + \frac{n-1}{n} \sum_{\substack{c \in \Sigma \\ E(i, c) > 0}} E(i, c)$$

$$= \frac{1}{n \cdot n_i} \cdot n_i + \frac{n-1}{n} \sum_{c \in \Sigma} E(i, c) \quad (0)$$

$$= \frac{1}{n} + \frac{n-1}{n} = 1$$

pour (0) on a retenu la condition $E(i, c) > 0$, cela n'ajoute que des termes nuls.

Pour (3) est toujours vérifié.

(1) et (2) sont évidemment toujours vérifiés

Question 44

$$E = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

$$E' = \begin{bmatrix} \frac{1}{12} & \frac{1}{12} & \frac{3}{4} & \frac{1}{12} \\ \frac{3}{4} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & \frac{1}{2} \\ \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{3}{4} \\ \frac{1}{12} & \frac{3}{4} & \frac{1}{12} & \frac{1}{12} \\ \frac{3}{8} & \frac{1}{4} & \frac{3}{16} & \frac{3}{16} \end{bmatrix}$$

17.56 / 20

Epreuve - Matière : 102 - 96.23

Session : 2029

CONSIGNES

- Remplir soigneusement, sur CHAQUE feuillet officiel, la zone d'identification en MAJUSCULES.
- Remplir soigneusement le cadre relatif au concours OU à l'examen qui vous concerne.
- Ne pas signer la composition et ne pas y apporter de signe distinctif pouvant indiquer sa provenance.
- Rédiger avec un stylo à encre foncée (bleue ou noire) et ne pas utiliser de stylo plume à encre claire.
- N'effectuer aucun collage ou découpage de sujets ou de feuillet officiel.
- Numéroté chaque PAGE (cadre en bas à droite de la page) sur le nombre total de pages que comporte la copie (y compris les pages vierges).
- Placer les feuilles dans le bon sens et dans l'ordre de numérotation des pages.

Question 95

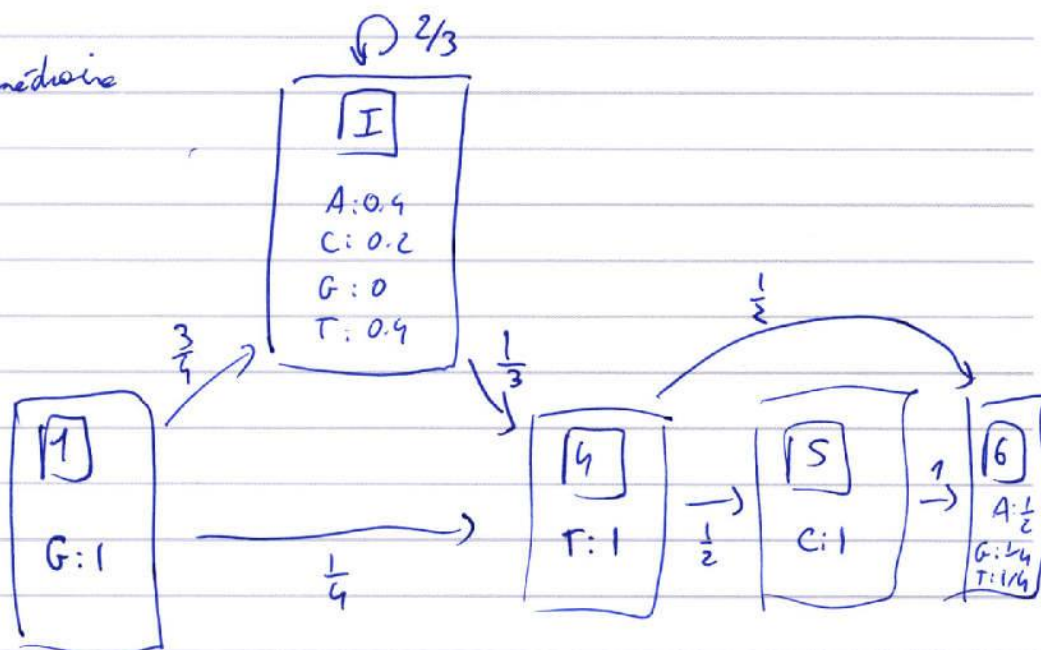
Question 96

l'intervalle mal ordonné est formé des deux colonnes 2 et 3

A a pour fréquence $\frac{2}{5}$, T: $\frac{2}{5}$ et C: $\frac{1}{5}$

on compte 2 transitions internes (de 2 à 3) et 1 externe de 3 à 4

Voici l'état intermédiaire



Concours section : AGRÉGATION EXTERNE INFORMATIQUE

Epreuve matière : Etude d'un problème informatique

N° Anonymat : **N240NAT1030219** Nombre de pages : 32

17.56 / 20

30 / 32

