MAPSI – Examen Réparti 2 – 45 pts

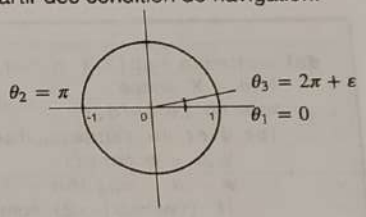
Durée : 1h30

Seuls documents autorisés : Calculatrice, antisèche recto-verso, tables de lois de probabilité
Le barème sur 45 pts (1 pt = 2 minutes)
– Barème indicatif –

Exercice 1 (8 pts) – **Tenir le cap**

Dans le cadre du développement d'un pilote automatique de bateau, on souhaite prédire le cap de navigation d'un voilier, exprimé en radian, à partir d'une description des conditions de navigation dans $\mathbb{R}^d$ (état du bateau, données météorologiques, etc...). Ce cap prédit sera comparé à celui mesuré par le GPS (la vérité terrain). Sans surprise, notre but est d'apprendre une fonction qui soit une bonne prédiction de la vérité terrain à partir des condition de navigation.

Définir une fonction de coût pour ce problème est difficile, dans la mesure où géométriquement $\theta_1 = 0$ est très éloigné de $\theta_2 = \pi$ mais très proche de $\theta_3 = 2\pi + epsilon$.
La figure ci-contre rappelle le fonctionnement du cercle trigonométrique, la projection sur l'axe des abscisses permettant de lire la valeur du cosinus, entre -1 et 1.



Q 1.1 (0.5pt) A quelle classe de problème a-t-on à faire ici ?

Régression

Q 1.2 (0.5pt) Est-il possible d'utiliser les moindres carrés ici ? Expliquer très succinctement pourquoi.

Du fait de la périodicité, on peut pas utiliser la mse qui donnerait des écarts alors que les angles sont les mêmes. $\mathcal{L}$ vaut bien 0 si les angles sont les mêmes, et maximale (2) si les angles sont opposées.
Il faudrait mettre un modulo 2π... Ce qui ne va pas être agréable dans l'optimisation.

Q 1.3 On propose d'utiliser la fonction de coût suivante : $\mathcal{L}(y, \hat{y}) = 1 - \cos(y - \hat{y})$ et un **modèle linéaire** f pour prédire le cap réel y en fonction des conditions de navigation $\mathbf{x} \in \mathbb{R}^d$.

Q 1.3.1 (0.5pt) Donner l'expression de f et le nombre de paramètres à optimiser.

$$f(\mathbf{x}_i) = \mathbf{x}_i \mathbf{w} = \sum_{j=1}^{d} w_j x_{ij}$$

Il y a d paramètres à optimiser

Q 1.3.2 (1.5pt) Soit un jeu de données $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, exprimer la fonction coût en développant l'expression de f, calculer les dérivées partielles puis en déduire l'expression du gradient.
Note pour le calcul de la dérivée : $\cos'(u) = -u' \sin(u)$

$$\mathcal{L} = \sum_{i=1}^{N} 1 - \cos(y_i - \hat{y}_i) = \sum_{i=1} 1 - \cos(y_i - \sum_{j=1}^{d} w_j x_{ij})$$

$$\frac{\partial \mathcal{L}}{\partial w_k} = \sum_{i=1}^{N} -x_{ik} \sin(y_i - \sum_{j=1}^{d} w_j x_{ij})$$

$$\nabla_{\mathbf{w}} \mathcal{L} = \sum_{i=1}^{N} -\mathbf{x}_i \sin(y_i - \mathbf{x}_i \mathbf{w})$$

Q 1.3.3 (2pts) Donner le code python respectant la signature suivante :
`optimise_cap(X,Y,n_iter_max=100, epsilon=1e-3, gamma=1e-2)` qui permet de retrouver les paramètres optimaux.
Classiquement, `epsilon` et `gamma` désignent respectivement le pas de gradient et le seuil de convergence.
Vous porterez une attention particulière à la compatibilité entre les dimensions des matrices dans les calculs.

```python
def optimise_cap(X,Y,n_iter_max=100, epsilon=1e-3, gamma=1e-2):
    N,d = X.shape
    w = np.zeros(d,1)
    for iter in range(n_iter_max):
        w_o = w.copy()
        w = w + epsilon * X.T @ np.sin(Y - X@w)
        if ((w-w_o)**2).sum() < gamma:
            break
    return w
```

Q 1.3.4 (1pt) En reprenant l'exemple donné en introduction avec $\theta_1, \theta_2, \theta_3$, la fonction coût proposée vous semble-t-elle robuste ou pas ? Justifier rapidement.

Dans le cas présent, la fonction coût est justement prise en défaut : entre θ_1 et θ_3, l'angle vaut ε et la mise à jour sera faible... mais entre θ_1 et θ_2, l'angle vaut π et le sinus sera également très petit malgré le besoin de mise à jour.

Q 1.4 (2pts) Question exploratoire : Ce problème ne permet pas directement de construire un pilote automatique... Pourquoi ?
Comment un industriel pourrait utiliser le résultat de nos calculs pour construire un prototype opérationnel ?

L'industriel a besoin de jouer sur les paramètres modifiable du bateau pour garder le cap fourni en consigne par le navigateur...
Un usage possible serait d'utiliser le modèle en inférence sur des paramètres flexibles du bateau échantillonnés (e.g. gouvernail) pour trouver les paramètres qui permettent d'obtenir le bon cap.

Exercice 2 (3pts) – **COVID (évidemment)**

Un journaliste propose l'analyse suivante sur les vagues Beta, Delta et Omicron en Afrique du Sud :
Au cours de la deuxième, troisième et quatrième vague, les enfants représentaient respectivement 3,9%, 3,5% et 17,7% de l'ensemble des admissions à l'hôpital. Cependant, lors des trois vagues Beta, Delta, Omicron, respectivement 7,1%, 3,8% et 6,1% des enfants et adolescents de moins de vingt ans [touchés] ont été admis à l'hôpital. Il apparaît donc que la proportion des cas [graves] était globalement similaire parmi les moins de vingt ans lors des trois vagues successives.
Le Monde, 4 janvier 2022

Q 2.1 (1pt) Les deux séries de chiffres différents sont-elles possibles ou le journaliste écrit-il n'importe quoi ?

Les chiffres sont compatibles (même si la phrase est ambigue au premier abord)
Nous verrons dans la question suivante qu'il s'agit de question de probas conditionnelles ou jointes

Q 2.2 (2pts) Si l'article est plausible, donner les variables aléatoires (et leurs modalités) en présence et les probabilités calculées ; sinon, expliquer en quoi c'est impossible.

Les variables aléatoires sont les suivantes :
Malade $M = \{m, \bar{m}\}$ [Optionnel, on ne considère de toutes façons que les malades m]
Hospitalisation $H = \{h, \bar{h}\}$
Age $A = \{e, \bar{e}\}$ (enfant -20 ans / pas enfant)
Variant/Vague $V = \{\beta, \delta, o\}$
Première série de chiffre : $p(A = e|H = h, V)$ pour les trois valeur de V
Seconde série de chiffre : $p(A = e, H = h|V)$ pour les trois valeur de V

Exercice 3 (11 points) – **Régression tri-logistique**

On s'attaque à un problème de classification à 3 classes en utilisant un modèle adapté de la régression logistique. On considérera un ensemble de données supervisé $\{(\mathbf{x}^i, y^i) \in \mathbb{R}^d \times \{1, 2, 3\}\}_{i=1}^N$.
Cette formulation s'appuie sur les fonctions f_k *suivantes :*

$$f_k(\mathbf{x}) = \frac{e^{\mathbf{x}\mathbf{w}_k}}{\sum_{j=1}^{3} e^{\mathbf{x}\mathbf{w}_j}} \text{ pour } k = 1, 2, 3$$

Q 3.1 (0.5pt) *Identifier les paramètres et leur nombre.*

$\mathbf{w}_k \in \mathbb{R}^d$
Il y a $3 \times d$ paramètres

Q 3.2 (2pts) *Montrer que la fonction* $f_k(\mathbf{x})$ *peut être un estimateur de* $P(Y = k|X = \mathbf{x})$ *car :*

1. *elle respecte bien la contrainte de normalisation des probabilités conditionnelles ;*
2. *elle évolue bien dans* $[0, 1]$ *(calculer les valeurs de* $f_k(\mathbf{x})$ *pour des valeurs de* $\mathbf{x}\mathbf{w}_k \to +\infty$ *et* $\mathbf{x}\mathbf{w}_k \to -\infty$*) ;*
3. *en déduire l'équation (brute) de la frontière de décision de la classe* k*.*

Ca tend respectivement vers 1 et 0
L'affectation se fait dans la classe k lorsque $f_k(\mathbf{x}) > 0.5$ ou, à défaut, avec $\arg\max_k f_k(\mathbf{x})$

$$\sum_{k=1}^{3} f_k(\mathbf{x}) = \sum_{k=1}^{3} \frac{e^{\mathbf{x}\mathbf{w}_k}}{\sum_{j=1}^{3} e^{\mathbf{x}\mathbf{w}_j}} = 1$$

Q 3.3 (0.5pt) Pour un couple observé $(\mathbf{x}_i, y_i)$, exprimer $P(Y = y_i | X = \mathbf{x}_i)$ en fonction des paramètres du modèle et des données.

$$P(Y = y_i | X = \mathbf{x}_i) = \frac{e^{\mathbf{w}_{y_i}\mathbf{x}_i}}{\sum_{j=1}^{3} e^{\mathbf{x}_i\mathbf{w}_j}}$$

Q 3.4 (1.5pt) On veut déterminer les paramètres optimaux par maximum de vraisemblance (discriminante). En faisant l'hypothèse que les décisions sont indépendantes pour les échantillons, donner la formulation du problème d'optimisation à résoudre (en passant au log).

On veut maximiser :

$$\arg\max_{\mathbf{w}} L = \arg\max_{\mathbf{w}} \prod_i p((y_i|\mathbf{x}_i)|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} \frac{e^{\mathbf{x}_i\mathbf{w}_{y_i}}}{\sum_j e^{\mathbf{x}_i\mathbf{w}_j}} = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \left(\mathbf{x}_i\mathbf{w}_{y_i} - \log(\sum_j e^{\mathbf{x}_i\mathbf{w}_j}) \right)$$

Q 3.5 (3.5pts) Calculer la dérivée de la log-vraisemblance par rapport aux $w_{k,\ell}$ et en déduire l'expression du gradient de la log-vraisemblance par rapport à chacun des $\mathbf{w}_k$.

Il faut calculer les dérivées par rapport aux $w_{k,\ell}$

$$\frac{\partial}{\partial w_{k,\ell}} = \sum_{i|y_i=k} x_{i,\ell} - \sum_{i=1}^{N} \frac{u'}{u}$$

avec :

$$u = \sum_j e^{\mathbf{x}_i\mathbf{w}_j} \qquad u' = x_{i,\ell} e^{\mathbf{x}_i\mathbf{w}_k}$$

soit :

$$\frac{\partial}{\partial w_{k,\ell}} = \sum_{i|y_i=k} x_{i,\ell} - \sum_{i=1}^{N} \frac{x_{i,\ell} e^{\mathbf{x}_i\mathbf{w}_k}}{\sum_j e^{\mathbf{x}_i\mathbf{w}_j}}$$

$$\nabla_{\mathbf{w}_k} \log L = \sum_{i|y_i=k} \mathbf{x}_i + \sum_i \mathbf{x}_i \frac{e^{\mathbf{x}_i\mathbf{w}_k}}{\sum_j e^{\mathbf{x}_i\mathbf{w}_j}}$$

1pt pour le premier terme avec la bonne somme
1.5pt pour le second terme
0.5pt pour le passage au gradient

Q 3.6 (1pt) Le résultat précédent, une fois obtenu, est assez logique. Expliquer brièvement pourquoi.

Pour maximiser $P(Y = k|X = \mathbf{x}_i)$ pour les points de la classe k, il faut spécifiquement maximiser $e^{\mathbf{x}_i \mathbf{w}_k}$. Une des solutions optimale correspond à avoir un $\mathbf{w}_k$ *ressemblant* aux points $\mathbf{x}_i$.
La formule de mise à jour renforce justement $\mathbf{w}_k$ dans la direction des $\mathbf{x}_i$ appartenant à la classe $+$ dans la direction des points qui ont une forte probabilité d'appartenir à la classe.

Q 3.7 (1pt) Donner l'algorithme d'optimisation de la vraisemblance en pseudo code (ou en python) en accordant une attention particulière aux dimensions des vecteurs mis à jour.

Fournir : $X, Y, \varepsilon, \gamma$

1. Initialiser les $\mathbf{w}_k$ (0 ou randn x epsilon)
2. tant que les itérations ne sont pas dépassées
   (a) $\forall k, \mathbf{w}_k^{(old)} \leftarrow \mathbf{w}_k$
   (b) $\forall k, \mathbf{w}_k \leftarrow \mathbf{w}_k^{(old)} + \varepsilon \nabla_{\mathbf{w}_k} \log L|_{\mathbf{w}_k^{(old)}}$
   (c) Si $\|W - W^{(old)}\|^2 < \gamma$, break
3. Retourner les $\mathbf{w}_k$

Note : pour un calcul optimisé, il faut précalculer les sommes des $\mathbf{x}$ des différentes classes une fois pour toutes.

Q 3.8 (1pt) Discuter brièvement les qualités de cette formulation par rapport à du un-contre-tous en régression logistique.

Les résultats sont plus facile à interpréter (car probabiliste). Il ne peut pas y avoir plusieurs probabilités d'affectation à différentes classes > 0.5.
On peut espérer que cette normalisation fasse que ça se passe un peu mieux dans les cas déséquilibrés. Les poids des $\mathbf{w}$ d'une classe étant mis à jour par rapport aux $\mathbf{w}$ des autres classes.

Exercice 4 (31 pts) – Ne parlez pas tous en même temps !

Soit $\tau = (\tau_1, \tau_2, ..., \tau_n)$ une retranscription d'une discussion entre deux individus, avec τ_i le i-ième mot de la séquence enregistrée. Pour chaque mot de τ, on ne sait pas quel individu en est l'émetteur (on observe uniquement une suite de mots). Soit un vocabulaire de 2 mots A et B. On suppose dans un premier temps un modèle où les probabilités d'émission de chaque individu ne dépendent pas de ce qui a été dit précédemment. On sait que l'individu 1 émet le mot A avec une probabilité de 0.7. L'individu 2 quant à lui émet B selon une probabilité de 0.8. Lorsqu'un individu parle, on sait que l'autre interlocuteur ne l'interrompt qu'avec une probabilité de 0.2. L'individu 1 commence la séquence avec une probabilité de 0.9.

Q 4.1 (2pt) Donner dans le cadre ci-dessous le modèle de markov caché correspondant (dessin automate ou matrices de transition)

$$A = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$$

$$b = \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix}$$

$$\pi = [0.9, 0.1]$$

Q 4.2 (2pt) Sur un générateur de nombres pseudo-aléatoires on tire la séquence 0.43, 0.62, 0.85, 0.17, 0.91, 0.99, 0.98, 0.13. Donner dans les cases ci-dessous la séquence de mots observée (0.5 pt par mot correct, -0.5pt par mot incorrect).

t	1	2	3	4
τ				

réponse attendue (si ils font comme en TD une alternance sampling s_t puis τ_t sur les pas de temps successifs) $AABA$ (états : 1 2 2 2)
autre reponse correcte (si ils samplent d'abord tous les s_t puis tous les τ_t) : $BBBA$ (états : 1 1 2 1)

Dans la suite on note A la matrice de transition, avec $a_{s,s'}$ la probabilité de transition de l'état s à l'état s', b la matrice d'émission, avec $b_s(w)$ la probabilité d'émettre le mot w dans l'état s et $b(w)$ le vecteur colonne des probabilités d'émission du mot w en fonction des différents états. π correspond aux probabilités initiales, avec π_s la probabilité de commencer dans l'état s. On note également $p(s_t)$ le vecteur ligne des probabilités de se trouver dans les différents états au temps t et $p(s_t = i)$ la probabilité de se trouver dans l'état particulier i au temps t. $p(\tau_t = w|s_t)$ est le vecteur colonne des probabilités d'émission de w selon les états possibles s_t.

Q 4.3 (2pt) Quelle est la probabilité que le troisième mot soit w (à 10^{-2} près) ? Cocher vrai ou faux pour chaque proposition ci dessous (0.5pt pour une réponse correcte, -0.5pt pour une réponse incorrecte).

Proposition	Vrai	Faux
$p(s1)p(w\|s_1)$		
$\pi AAb(w)$		
$\sum_{s_1} \pi_{s_1} \sum_{s_2} a_{s_1,s_2} \sum_{s_3} a_{s_2,s_3} b_{s_3}(w)$		
$p(s_3)b(w)$		

Tout est vrai.
Erratum : zut en fait je voulais mettre $p(s1)p(_{7_3} = w|s1)$ à la première, ce qui faisait qu'elle était vraie (on peut le faire avec n'importe quel état), mais là c'est faux...

Q 4.4 (2pt) Quelle est la probabilité que le quatrième mot d'une séquence soit un A sachant que le troisième mot a été émis par l'individu 1 (à 10^{-2} près) ? (2pt pour une réponse correcte, -0.5pt pour une réponse incorrecte)

Proposition	0.24	0.90	0.52	0.71	0.57	0.60	0.43	0.26	0.50	0.33
Réponse										

$p(t4|s3) = \sum_{s4} p(t4|s4)p(s4|s3) = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$
$p(t4 = A|s3 = 1) = 0.6$

Q 4.5 (2pt) Soit $p(s_3) = [0.644; 0.356]$. Quelle est la probabilité que le troisième et le quatrième mot d'une séquence soient identiques (à 10^{-2} près) ? (2pt pour une réponse correcte, -0.5pt pour une réponse incorrecte)

Proposition	0.24	0.90	0.52	0.71	0.57	0.60	0.43	0.26	0.50	0.33
Réponse										

$p(t4|s3) = \sum_{s4} p(t4|s4)p(s4|s3) = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$
$p(t3 = t4) = \sum_i p(t3 = i)p(t4 = i|t3 = i) = \sum_{s3} p(s_3) \sum_i p(t3 = i|s3)p(t4 = i|s3) = 0.644 * (0.7 * 0.6 + 0.3 * 0.4) + 0.356 * (0.2 * 0.3 + 0.8 * 0.7) = 0.56848$

Q 4.6 (2pt) Au bout d'un temps suffisamment long, la probabilité d'être dans l'un ou l'autre des deux états correspond à la distribution stationnaire du modèle de Markov considéré. Cocher la probabilité d'être dans l'état 1 pour tout temps $t >> 1$, parmi les propositions ci-dessous (2pt pour une réponse correcte, -0.5pt pour une réponse incorrecte)

Proposition	0.24	0.90	0.52	0.71	0.57	0.60	0.43	0.26	0.50	0.33
Réponse										

Distribution stationnaire pour s_t obtenue selon : $\mu A = \mu$
On a : $0.8\mu_1 + 0.2\mu_2 = \mu_1$
et : $\mu_1 + \mu_2 = 1$
Alors : $\mu_2 = 1 - \mu_1$
Donc : $0.8\mu_1 + 0.2(1 - \mu_1) = \mu_1$
Soit : $0.4\mu_1 = 0.2$
Et donc : $\mu_1 = 0.5$

Q 4.7 (2pt) Au bout d'un temps long (i.e., $t >> 1$), on observe un morceau de séquence de deux mots. Quelle est la probabilité d'observer la séquence A A (à 10^{-2} près) ? (2pt pour une réponse correcte, -0.5pt pour une réponse incorrecte)

Proposition	0.24	0.90	0.52	0.71	0.57	0.60	0.43	0.26	0.50	0.33
Réponse										

Distribution stationnaire pour s_t obtenue selon : $\mu A = \mu$
On a alors $mu = [0.5, 0.5]$
$p(t_{t+1}|s_t) = \sum_{s_{t+1}} p(t_{t+1}|s_{t+1})p(s_{t+1}|s_t) = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$
$p(t_{t+1} = t_t = A) = p(t_t = A)p(t_{t+1} = A|t_t = A) = \sum_{s_t} p(s_t)p(t_t = A|s_t)p(t_{t+1} = A|s_t) = 0.5 * (0.7 * 0.6 + 0.2 * 0.3) =$
0.24

Q 4.8 (1pt) On cherche la séquence d'états la plus probable correspondant à la séquence A A B B A. Pour cela on emploie l'algorithme de Viterbi, où il s'agit dans un premier temps de calculer pour tout t et tout i : $\delta_t(i) = \max_{s_1^{t-1}} p(s_1^{t-1}, s_t = i, \tau_1^t|\lambda)$. Le tableau δ est donné ci-dessous :

δ_t	1	2	3	4	5
$s = 1$	0.63	0.3528	0.0847	0.0203	0.0114
$s = 2$	0.02	0.0252	0.0564	0.0361	0.0058

Il s'agit également de définir $\Psi_t(j) = \underset{i \in [1,\ N]}{\operatorname{argmax}}\, \delta_{t-1}(i)a_{ij}$ pour chaque étape. $\Psi_t(j)$ est donné dans le tableau ci-dessous :

Ψ_t	1	2	3	4	5
$s = 1$	0	1	1	1	1
$s = 2$	0	1	1	2	2

Donner la séquence s^* la plus probable en fonction de ces informations :

	1	2	3	4	5
s^*					

```
tau
A A B B A
delta
0,63 0,3528 0,084672 0,02032128 0,0113799168
0,02 0,0252 0,056448 0,03612672 0,0057802752

phi
0 1 1 1 1
0 1 1 2 2

s*
1 1 1 1 1
```

Q 4.9 (4pt) Même question que la précédente mais sans aide pour δ et Ψ pour la séquence d'observations A B A B B. Compléter les tableaux δ, Ψ et s^* ci-dessous :

δ_t	1	2	3	4	5
$s = 1$	0.63				
$s = 2$	0.02				

Ψ_t	1	2	3	4	5
$s=1$	0				
$s=2$	0				

s^*	1	2	3	4	5

Viterbi

```
tau
A      B      A
delta
0,63 0,1512 0,084672 0,02032128 0,0048771072
0,02 0,1008 0,016128 0,01354752 0,0086704128

Psi
0      1      1
0      1      2      1      1
                     1      2

s*
1      1      1
                    2      2
```

Q 4.10 On souhaiterait maintenant augmenter le modèle de langue des individus : la probabilité $p(v|u,i)$ d'émission d'un mot v en un état ne dépend pas uniquement de l'individu i qui parle mais également du mot précédent u dans la transcription. Ces probabilités sont données dans les deux tableaux ci-dessous, 1 pour chaque individu.

$p(v\|u,i=1)$	v=A	v=B
u=A	0.7	0.3
u=B	0.5	0.5

Individu 1

$p(v\|u,i=2)$	v=A	v=B
u=A	0.5	0.5
u=B	0.2	0.8

Individu 2

Les probabilités de transitions entre individus émetteurs restent inchangées (indépendamment de ce qui est en train d'être dit). On suppose que la séquence commence nécessairement par un A (avec probabilités initiales sur l'individu émetteur identiques aux questions précédentes).

Q 4.10.1 (3pt) Sachant que cela peut se modéliser par 4 états cachés, donner dans les tableau ci-dessous les probabilités des transitions, sachant que les états sont nommés $< w, i >$, avec w le mot émis par l'état et i l'individu orateur en cet état (0.5pt par case correcte, -0.5pt par case incorrecte).

	A1	A2	B1	B2
π	0.9	0.1	0	0

Probabilités initiales

b	A	B
A1	1	0
A2	1	0
B1	0	1
B2	0	1

Probabilités d'émission

A	A1	A2	B1	B2
A1	0.56			
A2			0.06	
B1		0.04		
B2	0.1			

Probabilités de transition

```
    A1   A2   B1    B2
A1 0,56 0,1 0,24 0,1
A2 0,14 0,4 0,06 0,4
```

B1 0,4 0,04 0,4 0,16
B2 0,1 0,16 0,1 0,64

0.25 par reponse correcte

Q 4.11 Dans la question précédente, les individus interagissent. On souhaiterait maintenant produire un modèle qui permet de séparer les flux de deux individus qui ne s'écoutent pas (comme dans un débat politique par exemple), sachant que les retranscriptions sont mélangées. Les probabilités de passage d'un orateur à l'autre restent inchangées par rapport aux questions précédentes mais les modèles de langue des deux individus sont indépendants : pour chacun des deux individus i, la probabilité que i émette un mot v sachant que cet individu i a émis le mot u à sa dernière prise de parole est définie selon les deux tableaux $p(v|u,i)$ de la question précédente. La différence est que l'on doit modéliser une mémoire de ce que chacun des deux individus ont dit la dernière fois qu'ils ont parlé pour pouvoir utiliser ces probabilités de transition langagière.

Exemple : on observe la séquence A,A,B,A. On sait que la séquence d'orateurs qui se cache derrière ces observations est 1,1,1,2 et que c'est l'individu 1 qui reprend la parole pour le 5ième mot. La probabilité qu'il émette un A est de 0.5 car la dernière fois que 1 a parlé, il avait émis B : c'est la probabilité $p(v = A|u = B, i = 1)$ qui s'applique (et non $p(v = A|u = A, i = 1)$ comme cela aurait été le cas à la question précédente).

À noter qu'en début de séquence, un utilisateur peut ne pas avoir encore parlé précédemment. Les probabilités d'émission initiales (lorsque l'individu 1 ou 2 parle pour la première fois) sont égales aux valeurs de la question 1.

Q 4.11.1 (2pt) En utilisant le même formalisme HMM que précédemment et en observant également des séquences de mots émis par les deux individus, de combien d'états cachés aurions nous besoin au minimum pour modéliser ce système (2 points pour réponse correcte, -0.5 points pour réponse incorrecte) ?

Réponse	

état : <numero d'invidu qui parle, dernier mot émis par 1, dernier mot émis par 2> (avec dernier mot émis par i le mot émis par le noeud courant si c'est i qui parle)
ce qui donne (2 * 3 * 2)=12 car 2 individus, deux mots possibles + mot vide pour celui qui n'est pas en train de parler, deux mots possibles pour celui qui parle.
moitié des points si on répond 8 ou 18 (aucun vide ou vide possible sur les deux)

On voit qu'avec un vocabulaire plus grand, le nombre d'états cachés nécessaires tendrait à exploser rapidement pour prendre en compte la mémoire des derniers mots émis. Des algorithmes de décodage type Viterbi sont alors exclus. Connaissant une séquence de 5 mots observés, on souhaite connaître les distributions marginales conditionnelles de l'orateur à chaque pas de temps. On propose alors d'employer un mécanisme d'échantillonnage par Gibbs Sampling pour estimer ces distributions.

Q 4.11.2 (3pt) Soit la séquence observée $\tau = (A, A, B, A, B)$ et l'instanciation des orateurs à l'étape courante : $(1, 1, 2, 2, 1)$. Soit $p(i'|i)$ la probabilité d'un orateur i' suivant un orateur i, et $p(v|u, i)$ la probabilité que l'orateur i émette v sachant que c'est lui qui parle et qu'il avait émis u à sa dernière prise de parole. La probabilité conditionnelle que l'émetteur i_3 du 3-ième mot soit l'individu 1 sachant tous les autres orateurs est donnée par (cocher vrai ou faux pour chaque proposition dans le tableau ci-dessous, 0.25pt par bonne réponse cochée, -0.25pt par mauvaise réponse cochée) :

Proposition	1	2	3	4	5	6	7	8	9	10	11	12
Vrai												
Faux												

$p(i_3 = 1|i_1 = 1, i_2 = 1, i_4 = 2, i_5 = 1, \tau) =$

1. $\frac{p(i_1 = 1, i_2 = 1, i_4 = 2, i_5 = 1|\tau)}{p(i_3 = 2|i_1 = 1, i_2 = 1, i_4 = 2, i_5 = 1, \tau)}$
2. $\frac{p(i_1 = 1, i_2 = 1, i_3 = 1, i_4 = 2, i_5 = 1|\tau)}{p(i_1 = 1, i_2 = 1, i_4 = 2, i_5 = 1|\tau)}$
3. $\frac{p(i_1 = 1, i_2 = 1, i_3 = 1, i_4 = 2, i_5 = 1|\tau)}{p(i_1 = 1, i_2 = 1, i_3 = 1, i_4 = 2, i_5 = 1|\tau) + p(i_1 = 1, i_2 = 1, i_3 = 2, i_4 = 2, i_5 = 1|\tau)}$
4. $\frac{p(i_1 = 1, i_2 = 1, i_3 = 1, i_4 = 2, i_5 = 1, \tau)}{p(i_1 = 1, i_2 = 1, i_3 = 1, i_4 = 2, i_5 = 1, \tau) + p(i_1 = 1, i_2 = 1, i_3 = 2, i_4 = 2, i_5 = 1, \tau)}$
5. $\frac{p(i_1 = 1, i_2 = 1, i_4 = 2, i_5 = 1, \tau)}{p(i_3 = 2|i_1 = 1, i_2 = 1, i_4 = 2, i_5 = 1, \tau)}$
6. $\frac{p(i_1 = 1, i_2 = 1, i_3 = 1, i_4 = 2, i_5 = 1, \tau)}{p(i_1 = 1, i_2 = 1, i_4 = 2, i_5 = 1, \tau)}$
7. $\frac{p(v = A|u = \emptyset, i = 2)p(v = B|u = B, i = 1)p(i' = 1|i = 1)}{p(v = A|u = \emptyset, i = 2)p(v = B|u = B, i = 1)p(i' = 1|i = 1) + p(v = B|u = \emptyset, i = 2)p(v = A|u = B, i = 2)p(i' = 2|i = 2)}$
8. $\frac{p(v = B|u = A, i = 1)p(v = B|u = B, i = 1)p(i' = 1|i = 1)}{p(v = B|u = A, i = 1)p(v = B|u = B, i = 1)p(i' = 1|i = 1) + p(v = B|u = \emptyset, i = 2)p(v = A|u = B, i = 2)p(i' = 2|i = 2)}$
9. $\frac{p(v = B|u = A, i = 1)p(i' = 1|i = 1)}{p(v = B|u = A, i = 1)p(i' = 1|i = 1) + p(v = B|u = A, i = 2)p(i' = 2|i = 2)}$
10. $\frac{p(v = B|u = A, i = 1)p(i' = 1|i = 1)p(i' = 2|i = 1)}{p(v = B|u = A, i = 1)p(i' = 1|i = 1)p(i' = 2|i = 1) + p(v = B|u = A, i = 2)p(i' = 2|i = 1)p(i' = 2|i = 2)}$
11. $\frac{p(v = B|u = A, i = 1)p(i' = 1|i = 1)}{p(v = B|u = A, i = 1)p(i' = 1|i = 1) + p(v = B|u = \emptyset, i = 2)p(i' = 2|i = 1)}$
12. $\frac{p(v = B|u = A, i = 1)p(v = A|u = B, i = 1)p(i' = 1|i = 1)}{p(v = B|u = A, i = 1)p(v = A|u = B, i = 1)p(i' = 1|i = 1) + p(v = B|u = A, i = 2)p(v = A|u = B, i = 2)p(i' = 2|i = 1)}$

F, V, V, V, F, V, V, F, F, F, F, F

Q 4.11.3 (1pt bonus) La réponse la plus efficace à la question précédente (celle qui donne les éléments minimaux à calculer) est la réponse numéro :

Réponse	

7

Q 4.11.4 (4pt) Sachant que l'on s'occupe d'abord de l'individu ayant émis le 3ième mot, puis ensuite de celui ayant émis le 4ième, et que l'on a échantillonné les 2 nombres pseudo-aléatoires 0.5 et 0.5 pour les deux échantillonnages correspondants, compléter la séquence d'orateurs après deux étapes de Gibbs Sampling utilisant ces nombres, en considérant d'abord pour chaque tirage la probabilité que l'orateur soit l'individu 1 puis celle que ce soit l'individu 2 (2 points par réponse correcte, -2 points par réponse incorrecte) :

t	1	2	3	4	5
τ	A	A	B	A	B
s	1	1			1

$p(i_3 = 1 | i_1 = 1, i_2 = 1, i_4 = 2, i_5 = 1, \tau) = 0.3846$
donc avec 0.5, on choisit l'individu 2
$p(i_4 = 1 | i_1 = 1, i_2 = 1, i_3 = 2, i_5 = 1, \tau) = 0.7778$
donc avec 0.5 on choisit l'individu 1