Processus stochastiques : cryptanalyse

Stegen Thomas s154315
Adrien Minne s154340
Delaunoy Arnaud s153059

1 Première partie : chaines de Markov pour la modélisation du langage et MCMC

1.1 Chaine de Markov pour la modélisation du langage

Question 1

L'élément (i,j) de la matrice de transition correspond à la probabilité de passer de l'état i à l'état j. Il correspond donc à la probabilité que la lettre i soit suivie de la lettre j dans la séquence. Dès lors, soit θ l'élément (i,j) de la matrice de transition, θ est le paramètre d'une loi de Bernouilli avec comme possibilités :

— l'élément i est suivi de j (avec une probabilité θ)
— l'élément i n'est pas suivi de j (avec une probabilité $1-\theta$)

La méthode du maximum de vraisemblance consiste à maximiser $P(\mathbf{D_n}|\theta)$ avec $\mathbf{D_n}$ l'échantillon de donnée, ici seq1 et n le nombre de données. Pour une variable de Bernouilli, on a : Soit,

$$x_i = \begin{cases} 1 & \text{si i est suivi de j} \\ 0 & \text{si i n'est pas suivi de j} \end{cases}$$

et m le nombre d'occurrences de la lettre i.

$$\begin{aligned} P(\mathbf{D_n}|\theta) &= \prod_{i=1}^{m}(x_i\theta + (1-x_i)(1-\theta)) \\ &= \theta^{n_1}(1-\theta)^{n_0} \end{aligned}$$

Avec n_0 le nombre de fois où $x_i = 0$ et n_1 le nombre de fois où $x_i = 1$. Déterminons maintenant le θ maximisant cette fonction :

$$\begin{aligned} \frac{\partial P(\mathbf{D_n}|\theta)}{\partial\theta} &= n_1\theta^{n_1-1}(1-\theta)^{n_0} - n_1\theta^{n_1}(1-\theta)^{n_0-1} \\ &= \theta^{n_1-1}(1-\theta)^{n_0-1}(n_1(1-\theta) - n_0\theta) \\ &= \theta^{n_1-1}(1-\theta)^{n_0-1}(n_1 - \theta(n_1+n_0)) \end{aligned}$$

La valeur de θ maximisant la fonction $P(\mathbf{D_n}|\theta)$ est donc :

$$\theta_{i,j} = \frac{n_1}{n_0+n_1} = \frac{\text{nombre d'occurrences de i suivies de j}}{\text{nombre d'occurrences de i}}$$

Question 2

Cette question est résolue par la fonction `estimate_prob`. La t-ième puissance de Q est simplement calculée avec l'opérateur exposant de matlab.
Pour le calcul de $P(X_t = x)$, on calcule $\pi(t) = \pi(0) * Q^t$ avec $\pi(0) = (0.25\ 0.25\ 0.25\ 0.25)$ dans le cas ou la première lettre est choisie au hasard et $\pi(0) = (0\ 0\ 1\ 0)$ quand la première lettre est c. $P(X_t = x)$ est l'élément de $\pi(t)$ correspondant à x (le premier pour a, le deuxième pour b, ...).
L'évolution de la probabilité est reprise sur les graphiques 1, 2, 3 et 4. On constate qu'en t = 0, elle est uniquement dépendante de $\pi(0)$, ce qui est normal au vu de la définition de $\pi(0)$.

Quand t augmente, la probabilité est de moins en moins dépendante de $\pi(0)$ et dépend donc de plus en plus de la matrice de transition. C'est en accord avec la théorie car quand le temps augmente, la distribution tend vers la distribution stationnaire si Q^t converge vers une valeur limite pour $t \to +\infty$. C'est le cas ici car

$$Q^{1000} = Q^{1001} = \begin{bmatrix} 0.3518 & 0.0754 & 0.2161 & 0.3568 \\ 0.3518 & 0.0754 & 0.2161 & 0.3568 \\ 0.3518 & 0.0754 & 0.2161 & 0.3568 \\ 0.3518 & 0.0754 & 0.2161 & 0.3568 \end{bmatrix}$$

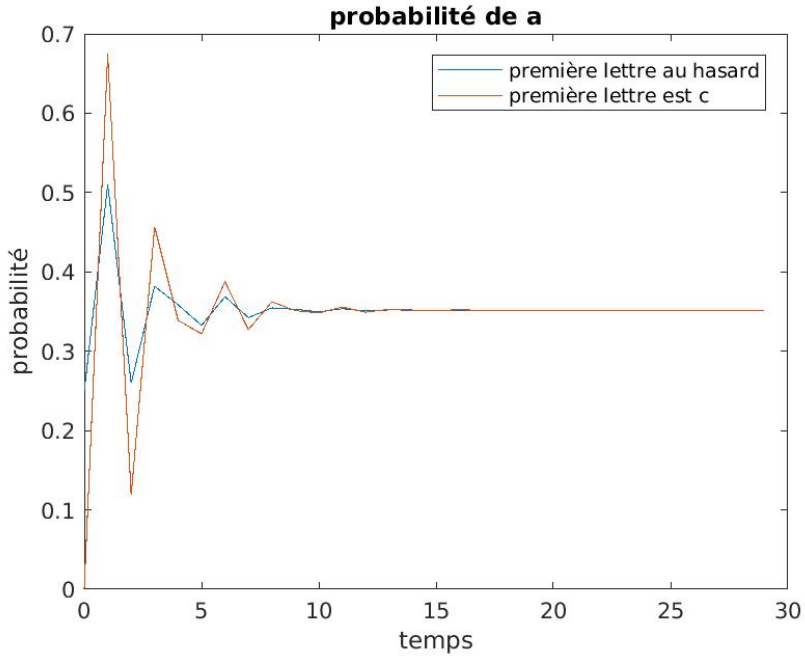

FIGURE 1

Question 3

Question 4

Question 5

1.2 Algorithme MCMC

Question 1

Pour prouver que π_0 est une distribution stationnaire de la chaine de Markov, il suffit de prouver que $\pi_0 = \pi_0 * Q$. On sait que les équations de balances détaillées $\pi_0(i)Q_{i,j} = \pi_0(j)Q_{j,i}$

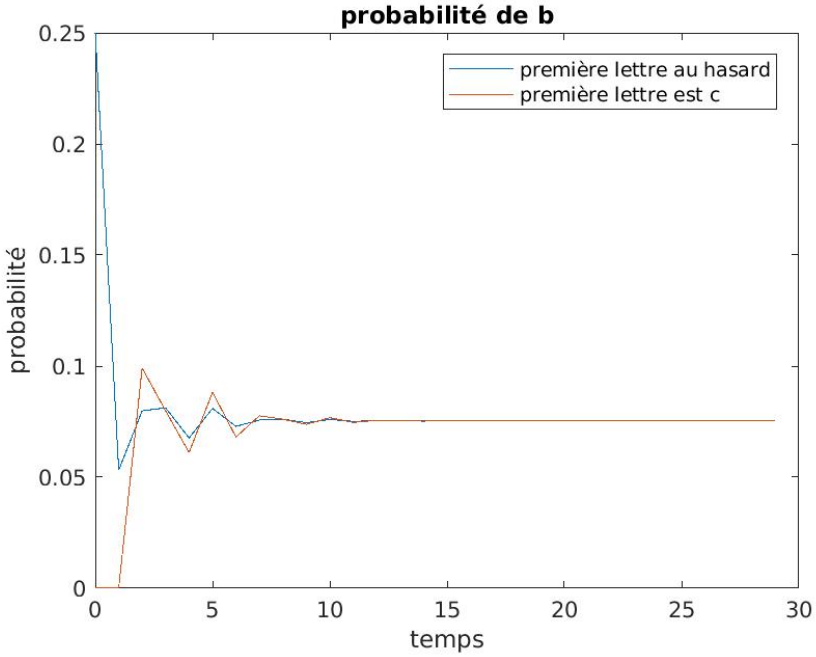

FIGURE 2

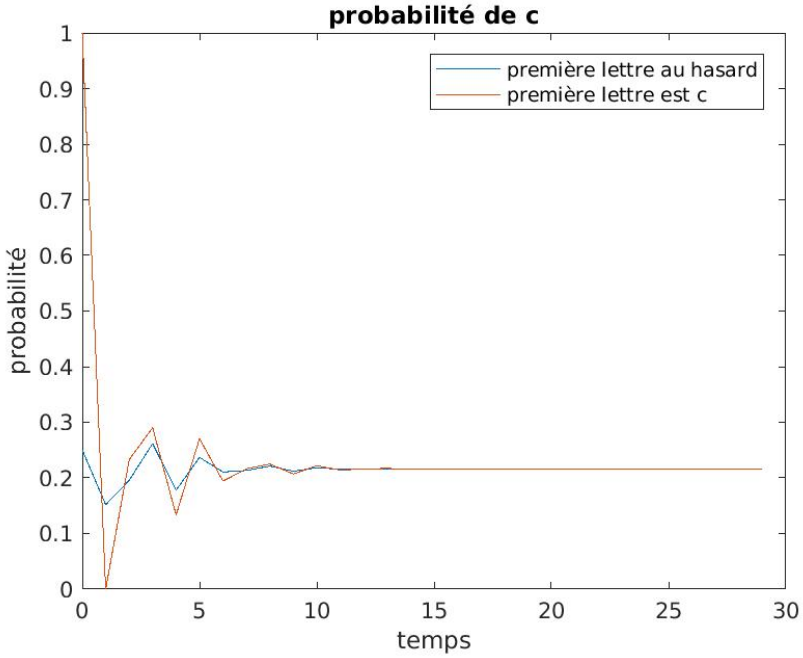

FIGURE 3

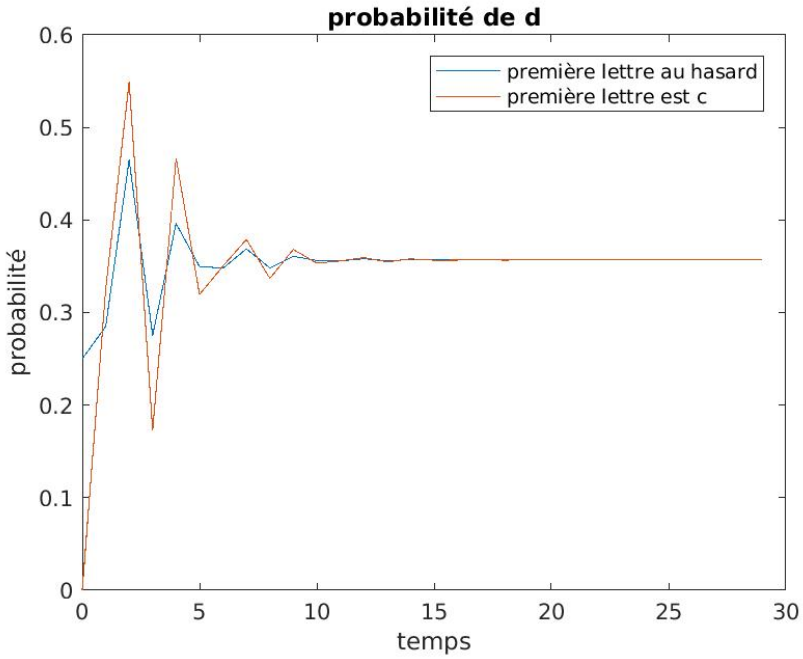

FIGURE 4

sont satisfaites. Passant en notation indicielle, on doit donc montrer que :

$$\begin{aligned}\pi_0(i) &= \sum_{k=0}^{N} \pi_0(k) * Q_{k,i} \\ &= \sum_{k=0}^{N} \pi_0(i) * Q_{i,k} \\ &= \pi_0(i) \sum_{k=0}^{N} \pi_0(i) Q_{i,k} \\ &= \pi_0(i)\end{aligned}$$

Cette distribution stationnaire est unique si la matrice de transition Q est irréductible.

Question 2

Cette démonstration a été faite avec l'aide du pdf nommé MetropolisExplanation mis dans l'archive trouvé sur internet.
Étudions d'abord la probabilité de transition.
La probabilité d'obtenir un élément x_j sachant que l'élément précédent de la chaîne de Markov est x_i est pour $i \neq j$ la probabilité que cet élément soit généré selon la loi q et accepté.

$$P(x_j|x_i) = \alpha(x_j, x_i) q(x_j|x_i)$$

avec

$$\begin{aligned}\alpha(x_j, x_i) &= min\left\{1, \frac{f(x_j)}{f(x_i)}\frac{q(x_i|x_j)}{q(x_j|x_i)}\right\} \\ &= min\left\{1, \frac{cP_X(x_j)}{cP_X(x_i)}\frac{q(x_i|x_j)}{q(x_j|x_i)}\right\}\end{aligned}$$

La probabilité d'obtenir à nouveau l'élément x_i sachant que l'élément précédent de la chaîne de Markov est également x_i est la somme de la probabilité que l'élément x_i soit généré selon la loi q et accepté et de la probabilité que tout autre élément soit généré et refusé.

$$P(x_i|x_i) = \alpha(x_i, x_i)q(x_i|x_i) + \sum_k (1 - \alpha(x_k, x_i))q(x_k|x_i)$$

Dans le cas où l'élément généré est différent du précédent, on a :

$$\begin{aligned}P(x_j|x_i)\pi_0(x_i) &= \alpha(x_j, x_i)q(x_j|x_i)\pi_0(x_i) \\ &= min\left\{1, \frac{cP_X(x_j)}{cP_X(x_i)}\frac{q(x_i|x_j)}{q(x_j|x_i)}\right\} q(x_j|x_i)\pi_0(x_i) \\ &= \frac{\pi_0(x_i)}{cP_X(x_i)} min\left\{cP_X(x_i)q(x_j|x_i), cP_X(x_j)q(x_i|x_j)\right\} \\ &\text{en posant } x_i \leftarrow x_j \text{ et } x_j \leftarrow x_i \\ &= \frac{\pi_0(x_j)}{cP_X(x_j)} min\left\{cP_X(x_j)q(x_i|x_j), cP_X(x_i)q(x_j|x_i)\right\} \\ &= min\left\{1, \frac{cP_X(x_i)}{cP_X(x_j)}\frac{q(x_j|x_i)}{q(x_i|x_j)}\right\} q(x_i|x_j)\pi_0(x_j) \\ &= \alpha(x_i, x_j)q(x_i|x_j)\pi_0(x_j) \\ &= P(x_i|x_j)\pi_0(x_j)\end{aligned}$$

Dans le cas où l'élément généré est le même que le précédent, x_i étant égal à x_j il est évident que

$$P(x_i|x_j)\pi_0(x_j) = P(x_j|x_i)\pi_0(x_i)$$

car

$$P(x_i|x_i)\pi_0(x_i) = P(x_i|x_i)\pi_0(x_i)$$

2 Deuxième partie : décryptage d'une séquence codée

Question 1

Pour déterminer la cardinalité de l'ensemble Θ, il suffit de calculer le nombre possibles de permutations des caractères disponibles. On considère la langue anglaise avec 40 caractèresle nombre de permutations possibles, on considère qu'on a 40 permutatations possibles pour la première lettre, 39 pour la deuxième, ... On a donc :

$$|\Theta| = 40 * 39 * 38 * ... * 1!$$

.

Question 2

Dans le modèle π_0, Q, on peut trouver la vraisemblance de la séquence T' en mutlipliant les probabilité d'avoir la lettre n de T' en partant de la lettre $n-1$ de T'. Nos probabilités sont calculées comme suit :

— $P(lettre1 = T'(1))$ est simplement sa probabilité d'avoir cette lettre dans π_0.
— $P(lettre2 = T'(2))$: est la probabilité d'avoir cette lettre dans $\pi_0 * Q$.
— $P(lettre3 = T'(3))$: est la probabilité d'avoir cette lettre dans $\pi_0 * Q^2$.
— etc

La vraisemblance de la chaine T' est donc, notant N la taille de la chaine T' et $|\pi|_{T'(k)}$ la probabilité d'avoir la $k^{ième}$ lettre de T' selon la distribution π :

$$P(T') = \prod_{k=0}^{N} |\pi_0 * Q^k|_{T'(k)}$$

Pour trouver la vraisemblance de D, il faut d'abord calculer la matrice Q_θ exprimant notre matrice Q dans le cas de la permutation θ. Ensuite, la vraisemblance de D se trouve de la même facon que celle de T', à savoir :

$$P(D) = \prod_{k=0}^{N} |\pi_0 * Q_\theta^k|_{D(k)}$$

Question 3

Si il n'a que 10 codes candidats possible qui ont la même probabilité, il est assez simple de retrouver le texte de base. Il suffit en effet de calculer la vraisemblance de notre texte avec chaque code θ suivant l'algorithme décrit au point 2.2, et puis simplmenet sélectionner le θ menan t au texte ayant la plus grande vraisemblance.

Question 4

Question 5

Question 6