

Probabilités conditionnelles

Comment une IA choisit le mot suivant dans une phrase

À savoir avant de commencer

Ce document utilise les arbres pondérés et la formule des probabilités totales (Première), puis les matrices de transition et l'état stationnaire d'une chaîne de Markov (Terminale). Il peut se lire indépendamment du document 0, mais réutilise l'idée générale de « paramètres que l'on ajuste » pour faire le lien avec l'entraînement d'une IA.

Niveau : Progressif : Seconde → Première → Terminale (chaque partie indique son niveau)

Introduction : deviner le mot suivant, un jeu de probabilités

Quand un modèle de langage comme Claude génère une réponse, il ne « réfléchit » pas au sens plein du terme : à chaque étape, il calcule une probabilité pour chaque mot possible du vocabulaire, sachant tout ce qui a déjà été écrit, puis choisit (ou tire au sort) un mot selon ces probabilités.

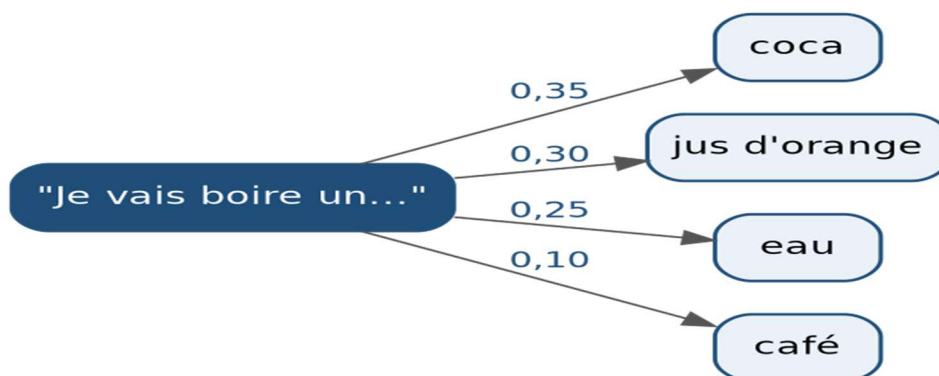
On appelle cela une probabilité conditionnelle : la probabilité qu'un évènement se réalise, sachant qu'un autre évènement est déjà réalisé.

Dans ce document, on étudie ce principe sur des exemples très simplifiés, avec un vocabulaire de quelques mots seulement, pour comprendre la mécanique mathématique qui se cache derrière la génération de texte.

Partie A — Une première prédiction (probabilité simple)

Niveau : Seconde

Une IA très simplifiée a appris, à partir d'un grand nombre de phrases, qu'après « Je vais boire un... », les mots suivants apparaissent avec les probabilités suivantes :



1. Vérifie que la somme de ces probabilités vaut 1. Pourquoi est-ce nécessaire ?
2. Si l'IA choisit toujours le mot le plus probable, quel mot complètera la phrase ?
3. Quelle est la probabilité que le mot choisi soit « coca » ou « eau » ?
4. Si l'on répète cette prédiction 200 fois (200 phrases différentes se terminant ainsi), combien de fois peut-on s'attendre à voir apparaître le mot « coca » ?

En réalité, une IA ne choisit pas toujours le mot le plus probable : elle « explore » parfois d'autres possibilités, pour éviter de répéter toujours les mêmes phrases. Imaginons qu'elle choisisse le mot le plus probable 70 % du temps, et un mot au hasard parmi les 4 (choisi équitablement) les 30 % du temps restants.

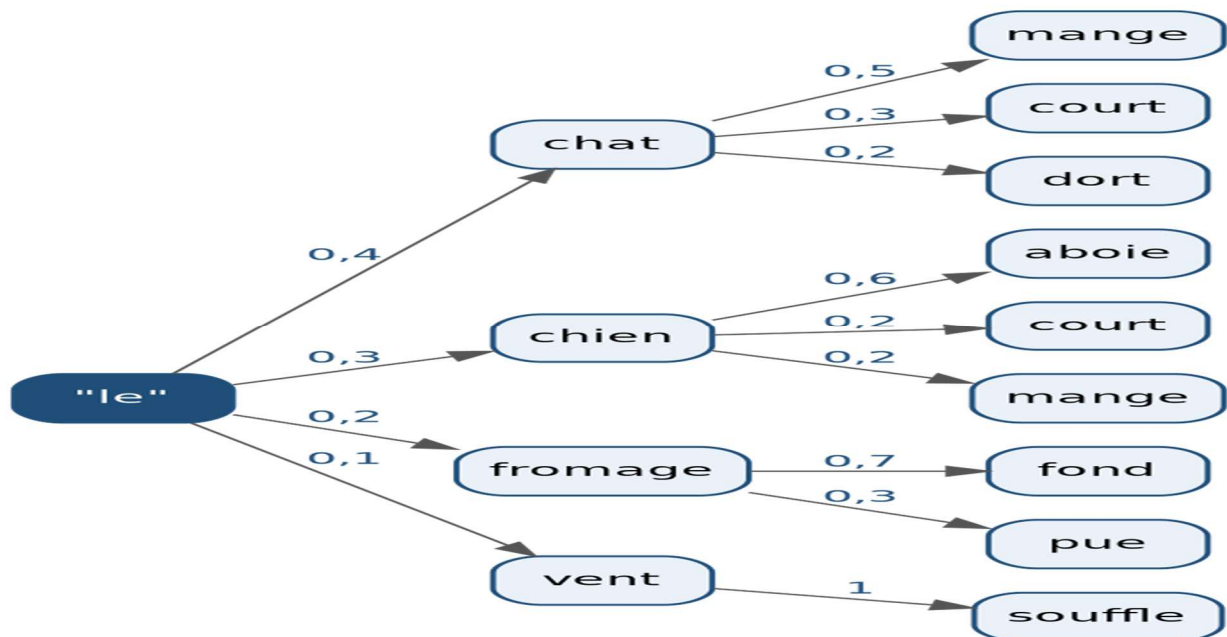
5. Calcule la nouvelle probabilité que le mot choisi soit « coca » dans ce cas.

Partie B — Enchaîner plusieurs mots : arbres pondérés

Niveau : Première

Pour générer plusieurs mots à la suite, l'IA doit tenir compte de chaque mot déjà choisi pour prédire le suivant. On peut représenter cela par un arbre pondéré, où chaque branche porte une probabilité conditionnelle.

Une IA simplifiée a appris, à partir d'un corpus de textes en français, les probabilités suivantes pour compléter une phrase commençant par « le » :



Ces probabilités sont conditionnelles : par exemple, $P(\text{mange} \mid \text{chat}) = 0,5$ signifie que, sachant que le deuxième mot est « chat », la probabilité que le troisième soit « mange » vaut 0,5.

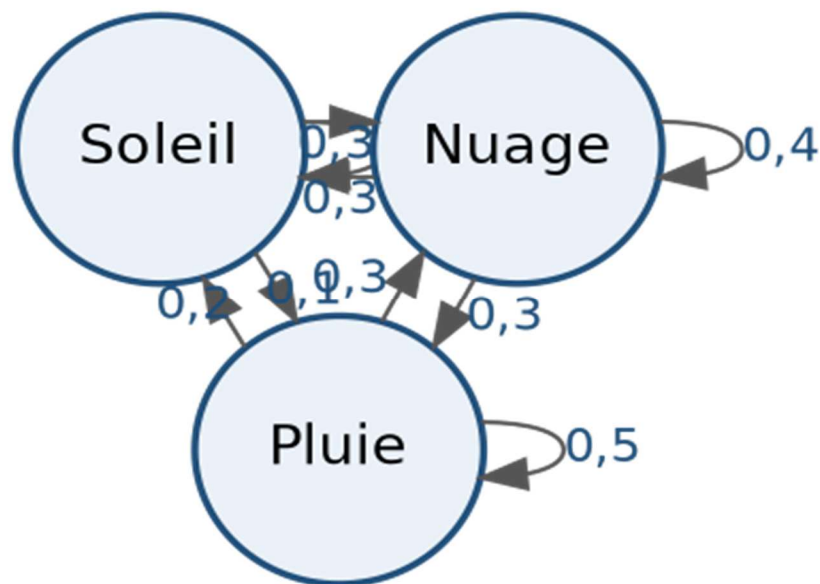
6. Pour chacun des quatre chemins de l'arbre, calcule la probabilité de la séquence de 3 mots correspondante, en multipliant les probabilités le long des branches : « le chat mange », « le chien court », « le fromage fond », « le vent souffle ».
7. Vérifie que la somme de toutes les probabilités des séquences complètes de l'arbre (les 8 chemins possibles, pas seulement les 4 ci-dessus) vaut bien 1. (Indice : utilise la formule des probabilités totales à chaque nœud.)
8. Quelle est, parmi les 4 séquences de la question 1, celle qui a la plus grande probabilité d'être générée par l'IA ?
9. Si l'IA choisit, à chaque étape, le mot le plus probable localement (on parle de choix « glouton » ou greedy), quelle phrase de 3 mots va-t-elle générer ? Compare avec la réponse à la question 3.
10. Pourquoi ce type de modèle peut-il donner l'illusion de « comprendre » une langue, alors qu'il ne fait qu'appliquer des probabilités apprises sur des exemples ?

Partie C — Modéliser une suite de prédictions par une matrice de transition

Niveau : Terminale (spécialité + option mathématiques expertes)

Pour une IA qui ne regarde que l'élément immédiatement précédent pour prédire le suivant (un modèle très simplifié, appelé chaîne de Markov d'ordre 1), on peut résumer toutes les probabilités conditionnelles dans un tableau appelé matrice de transition.

Prenons un exemple météo, plus simple à manipuler que des mots : un système à 3 états — Soleil (S), Nuage (N), Pluie (P) — où le temps de demain ne dépend que du temps d'aujourd'hui.



Ce diagramme se traduit par la matrice de transition suivante (chaque ligne donne les probabilités de transition à partir d'un état) :

	vers Soleil	vers Nuage	vers Pluie
depuis Soleil	0,6	0,3	0,1
depuis Nuage	0,3	0,4	0,3
depuis Pluie	0,2	0,3	0,5

- Vérifie que chaque ligne de la matrice P somme bien à 1. Pourquoi cette propriété est-elle indispensable pour une matrice de transition ?
- On suppose qu'il fait soleil aujourd'hui (jour 0). Donne le vecteur de probabilité initial π_0 (sous la forme $(P(\text{Soleil}), P(\text{Nuage}), P(\text{Pluie}))$).
- Calcule le vecteur $\pi_1 = \pi_0 \times P$, qui donne la probabilité du temps qu'il fera demain (jour 1).
- Calcule de la même façon $\pi_2 = \pi_1 \times P$ (le temps du jour 2). Vérifie que la somme des trois probabilités reste égale à 1.

Après un grand nombre de jours, les probabilités finissent par se stabiliser : on parle d'état stationnaire. Cet état, noté $\pi = (x ; y ; z)$, vérifie l'équation $\pi \times P = \pi$, avec $x + y + z = 1$.

- En écrivant les trois équations issues de $\pi \times P = \pi$ (une par colonne), montre que l'on peut se ramener au système : $0,4x = 0,3y + 0,2z$ et $2y = x + z$.
- Résous ce système, avec la condition $x + y + z = 1$, pour trouver l'état stationnaire $(x ; y ; z)$. Arrondis à 10^{-3} près.
- Interprète ce résultat : que représente, concrètement, cet état stationnaire pour un système météorologique modélisé de cette façon ?

Partie D — Bonus : le piège du choix « glouton »

Niveau : Terminale

On a vu en partie B que choisir, à chaque étape, le mot le plus probable localement (choix « glouton ») menait à la même séquence que la séquence globalement la plus probable. Ce n'est pas toujours le cas ! Voici un contre-exemple.

Un modèle simplifié doit générer 2 mots. Le premier mot est A avec une probabilité de 0,6, ou B avec une probabilité de 0,4. Ensuite :

- si le premier mot est A, le second mot est X avec une probabilité de 0,5, ou Y avec une probabilité de 0,5 ;
- si le premier mot est B, le second mot est X avec une probabilité de 0,9, ou Y avec une probabilité de 0,1.

- Quelle séquence de 2 mots un algorithme « glouton » (qui choisit toujours le mot le plus probable à chaque étape) va-t-il générer ? Quelle est sa probabilité ?
- Calcule la probabilité de chacune des 4 séquences possibles (A-X, A-Y, B-X, B-Y). Quelle est la séquence réellement la plus probable ?
- Que peux-tu en conclure sur les limites d'un algorithme purement « glouton » pour générer du texte ?

Ce que ce modèle simplifie, et ce que fait réellement une IA comme Claude

Ce que ce modèle simplifie	Ce que fait réellement un modèle comme Claude
Un vocabulaire de 4 ou 5 mots seulement. Une mémoire très courte : la prédiction ne dépend que du dernier mot (chaîne de Markov d'ordre 1). Des probabilités fixes, écrites une fois pour toutes dans un tableau. Un choix simplifié : soit le mot le plus probable (glouton), soit un tirage uniforme au hasard.	Un vocabulaire de dizaines de milliers de « tokens » (mots ou fragments de mots). Un contexte de plusieurs milliers de tokens pris en compte simultanément, grâce à un mécanisme appelé attention — bien plus riche qu'une mémoire d'ordre 1. Des probabilités calculées à la volée par un immense réseau de neurones (des milliards de paramètres, ajustés par descente de gradient — voir document 0), et non lues dans un tableau fixe. Des stratégies de génération plus fines (température, tirage parmi les k ou p mots les plus probables, recherche en faisceau...) qui limitent justement le piège du choix « glouton » vu en partie D.

L'idée centrale — calculer une distribution de probabilité conditionnelle sur le mot suivant, sachant le contexte — reste, elle, tout à fait juste : c'est exactement ce que fait un modèle de langage à chaque étape de sa génération de texte.

Exercices d'entraînement

Exercice 1 — Probabilité simple (arbre à un niveau)

Niveau : Seconde

Une IA complète la phrase « Il fait très chaud, je vais boire un... » avec les probabilités suivantes : coca 0,35 ; jus d'orange 0,30 ; eau 0,25 ; café 0,10.

21. Quelle est la probabilité que le mot choisi soit « jus d'orange » ou « eau » ?
22. Sur 200 phrases générées, combien de fois peut-on s'attendre à voir apparaître « café » ?
23. L'IA choisit le mot le plus probable 70 % du temps, et un mot au hasard parmi les 4 les 30 % restants. Calcule la nouvelle probabilité que le mot choisi soit « coca ».

Exercice 2 — Arbre à deux niveaux

Niveau : Première

Après le mot « un », une IA simplifiée prédit : chat (0,5) ou chien (0,5). Puis : si chat, miaule (0,6) ou dort (0,4) ; si chien, aboie (0,7) ou dort (0,3).

24. Représente cette situation par un arbre pondéré.
25. Calcule la probabilité de chacune des 4 séquences de 2 mots possibles.
26. Vérifie que la somme de ces 4 probabilités vaut 1.
27. Quelle est la séquence la plus probable ? Est-ce la même que celle obtenue par un choix « glouton » à chaque étape ?

Exercice 3 — Matrice de transition et état stationnaire

Niveau : Terminale

On modélise un lecteur qui, chaque jour, choisit de lire soit un article Sport (S), soit un article Actu (A), selon son choix de la veille, d'après la matrice de transition suivante :

	vers Sport	vers Actu
depuis Sport	0,8	0,2
depuis Actu	0,3	0,7

28. Vérifie que P est bien une matrice de transition.
29. Le lecteur a lu un article Sport aujourd'hui. Donne π_0 , puis calcule π_1 et π_2 .
30. Détermine l'état stationnaire $(x ; y)$ de cette chaîne, avec $x + y = 1$.
31. Interprète ce résultat en une phrase.

Corrigé

Partie A

A.1. $0,35 + 0,30 + 0,25 + 0,10 = 1$. C'est nécessaire car ces 4 mots forment l'ensemble de toutes les possibilités envisagées par le modèle : la probabilité totale de choisir « un mot quelconque parmi ces 4 » doit valoir 1 (certitude).

A.2. Le mot le plus probable est « coca » (0,35).

A.3. $P(\text{coca ou eau}) = P(\text{coca}) + P(\text{eau}) = 0,35 + 0,25 = 0,60$ (événements incompatibles, on additionne).

A.4. Nombre attendu = $200 \times 0,35 = 70$ fois.

A.5. $P'(\text{coca}) = 0,7 \times 0,35 + 0,3 \times (1/4) = 0,245 + 0,075 = 0,32$, soit 32 %.

Partie B

Séquence	Calcul	Probabilité
le chat mange	$0,4 \times 0,5$	0,20
le chien court	$0,3 \times 0,2$	0,06
le fromage fond	$0,2 \times 0,7$	0,14
le vent souffle	$0,1 \times 1$	0,10

B.2. Les 8 séquences complètes sont : le chat mange (0,20), le chat court ($0,4 \times 0,3 = 0,12$), le chat dort ($0,4 \times 0,2 = 0,08$), le chien aboie ($0,3 \times 0,6 = 0,18$), le chien court (0,06), le chien mange ($0,3 \times 0,2 = 0,06$), le fromage fond (0,14), le fromage pue ($0,2 \times 0,3 = 0,06$), le vent souffle (0,10). Somme : $0,20 + 0,12 + 0,08 + 0,18 + 0,06 + 0,06 + 0,14 + 0,06 + 0,10 = 1,00$.
✓ (C'est la formule des probabilités totales appliquée aux deux niveaux de l'arbre : à chaque nœud, la somme des probabilités des branches vaut 1, et le produit se conserve globalement.)

B.3. La séquence la plus probable parmi les 4 de la question 1 est « le chat mange », avec 0,20.

B.4. Un choix glouton sélectionne d'abord le mot le plus probable après « le », c'est-à-dire « chat » (0,4 contre 0,3 ; 0,2 ; 0,1), puis le mot le plus probable après « chat », c'est-à-dire « mange » (0,5). On obtient donc aussi « le chat mange ». Ici, le choix glouton coïncide avec la séquence globalement la plus probable — mais ce n'est pas toujours vrai (voir partie D).

B.5. Le modèle ne fait qu'appliquer des statistiques apprises sur d'énormes quantités de texte : il reproduit des régularités du langage sans « comprendre » le sens des mots. Il paraît intelligent car ces régularités correspondent, la plupart du temps, à des phrases sensées pour un lecteur humain.

Partie C

C.1. Ligne Soleil : $0,6 + 0,3 + 0,1 = 1$. Ligne Nuage : $0,3 + 0,4 + 0,3 = 1$. Ligne Pluie : $0,2 + 0,3 + 0,5 = 1$. C'est indispensable car, en partant de n'importe quel état, on est certain d'arriver « quelque part » le lendemain : la somme des probabilités de toutes les transitions possibles doit valoir 1.

C.2. $\pi_0 = (1 ; 0 ; 0)$.

C.3. $\pi_1 = \pi_0 \times P = (0,6 ; 0,3 ; 0,1)$ (on lit simplement la première ligne de P).

C.4. $\pi_2 = \pi_1 \times P$: $P(\text{Soleil}) = 0,6 \times 0,6 + 0,3 \times 0,3 + 0,1 \times 0,2 = 0,36 + 0,09 + 0,02 = 0,47$; $P(\text{Nuage}) = 0,6 \times 0,3 + 0,3 \times 0,4 + 0,1 \times 0,3 = 0,18 + 0,12 + 0,03 = 0,33$; $P(\text{Pluie}) = 0,6 \times 0,1 + 0,3 \times 0,3 + 0,1 \times 0,5 = 0,06 + 0,09 + 0,05 = 0,20$. Somme : $0,47 + 0,33 + 0,20 = 1,00$ ✓

C.5. En notant $\pi = (x ; y ; z)$, l'équation $\pi \times P = \pi$ donne, pour la colonne Soleil : $x = 0,6x + 0,3y + 0,2z$ soit $0,4x = 0,3y + 0,2z$. Pour la colonne Nuage : $y = 0,3x + 0,4y + 0,3z$ soit $0,6y = 0,3x + 0,3z$, c'est-à-dire $2y = x + z$.

C.6. En combinant $0,4x = 0,3y + 0,2z$, $2y = x + z$ et $x + y + z = 1$, on trouve (par substitution) : $x = 7/18 \approx 0,389$; $y = 1/3 \approx 0,333$; $z = 5/18 \approx 0,278$.

C.7. Cet état stationnaire signifie qu'après un grand nombre de jours, quel que soit le temps qu'il faisait au départ, la probabilité qu'il fasse soleil un jour donné se stabilise autour de 38,9 %, nuage autour de 33,3 %, et pluie autour de 27,8 % : le système « oublie » sa condition initiale et se stabilise vers ces fréquences à long terme.

Partie D

D.1. Un algorithme glouton choisit d'abord A ($0,6 > 0,4$). Puis, pour le second mot sachant A, X et Y sont à égalité (0,5 chacun) : on peut supposer qu'il choisit X par convention. La séquence générée est donc A-X, avec une probabilité de $0,6 \times 0,5 = 0,30$.

D.2. Les 4 probabilités : $A-X = 0,6 \times 0,5 = 0,30$; $A-Y = 0,6 \times 0,5 = 0,30$; $B-X = 0,4 \times 0,9 = 0,36$; $B-Y = 0,4 \times 0,1 = 0,04$. La séquence réellement la plus probable est B-X, avec $0,36 > 0,30$.

D.3. L'algorithme glouton, en choisissant A dès la première étape car il semblait le plus prometteur localement, est passé à côté de la séquence globalement la plus probable (B-X). Cela montre qu'un choix optimal à chaque étape ne garantit pas un résultat optimal sur l'ensemble de la séquence : c'est pour cette raison que les vrais modèles de langage utilisent des stratégies plus élaborées (comme la recherche en faisceau, ou « beam search », qui explore plusieurs chemins prometteurs à la fois avant de choisir).

Exercice 1

1. $P(\text{jus ou eau}) = 0,30 + 0,25 = 0,55$. 2. $200 \times 0,10 = 20$ fois. 3. $P'(\text{coca}) = 0,7 \times 0,35 + 0,3 \times (1/4) = 0,245 + 0,075 = 0,32$.

Exercice 2

2. $P(\text{chat, miaule}) = 0,5 \times 0,6 = 0,30$; $P(\text{chat, dort}) = 0,5 \times 0,4 = 0,20$; $P(\text{chien, aboie}) = 0,5 \times 0,7 = 0,35$; $P(\text{chien, dort}) = 0,5 \times 0,3 = 0,15$.

3. Somme : $0,30 + 0,20 + 0,35 + 0,15 = 1,00$ ✓.

4. La séquence la plus probable est « un chien aboie » (0,35). Le choix glouton choisit d'abord chat ou chien à égalité (0,5 chacun) — supposons chien par convention — puis aboie (0,7, le plus probable pour chien) : on retombe bien sur « un chien aboie », la séquence globalement la plus probable ici.

Exercice 3

1. Ligne Sport : $0,8 + 0,2 = 1$. Ligne Actu : $0,3 + 0,7 = 1$. ✓

2. $\pi_0 = (1 ; 0)$. $\pi_1 = (0,8 ; 0,2)$. $\pi_2 = (0,8 \times 0,8 + 0,2 \times 0,3 ; 0,8 \times 0,2 + 0,2 \times 0,7) = (0,64 + 0,06 ; 0,16 + 0,14) = (0,70 ; 0,30)$.

3. $\pi \times P = \pi$ donne $x = 0,8x + 0,3y$ soit $0,2x = 0,3y$, donc $x = 1,5y$. Avec $x + y = 1$: $1,5y + y = 1 \rightarrow 2,5y = 1 \rightarrow y = 0,4$; $x = 0,6$. État stationnaire : $(0,6 ; 0,4)$.

4. À long terme, le lecteur lit un article Sport environ 60 % des jours, et un article Actu environ 40 % des jours, quel que soit son choix de départ.

Mathématiques du programme mobilisées dans ce document

Seconde

Probabilités : événements incompatibles, calcul de probabilité d'une réunion d'événements.

Proportions et effectifs attendus dans une expérience répétée.

Première (spécialité mathématiques)

Arbres pondérés, règle du produit le long d'un chemin, formule des probabilités totales.

Variations aléatoires discrètes.

Terminale (spécialité + option mathématiques expertes)

Matrices : produit d'un vecteur ligne par une matrice, matrices de transition.

Chaînes de Markov, état stationnaire, résolution de systèmes d'équations linéaires.