Approches Bayésiennes pour les problèmes inverses

Angèle Niclas

13ème école d'été de mécanique théorique - Quiberon
Problèmes Inverses en Mécanique - Septembre 2024

Table des matières

Références

- H. Kekkonen, Bayesian inverse problems (2019) → lien
- R. Scheichl, J. Zech, Numerical Methods for Bayesian Inverse Problems (2021) → lien
- A. Tarantola, Inverse Problem Theory (2005)

1 Introduction et Motivation

1.1 Oscillateur harmonique amorti

Introduisons pour commencer l'exemple de l'oscillateur harmonique amorti, qui servira de fil conducteur tout au long de ce cours. Il s'agit d'un système masse-ressort avec amortissement visqueux, où la force d'amortissement est proportionnelle à la vitesse de la masse. Les paramètres du système sont les suivants : k, la constante de raideur du ressort, m, la masse de l'objet, et τ, le temps de relaxation, caractérisant la décroissance de l'amplitude au cours du temps.

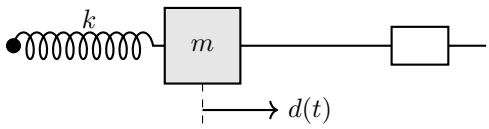

FIGURE 1: Oscillateur harmonique amorti

En définissant la pulsation propre par $\omega_0 = \sqrt{k/m}$, le déplacement $d(t)$ satisfait l'équation différentielle suivante :

$$\forall t \in \mathbb{R}_+ \qquad d''(t) + \frac{2}{\tau} d'(t) + {\omega_0}^2 d(t) = 0$$

On sait que, sous certaines conditions initiales, la solution prend la forme suivante :

$$\forall t \in \mathbb{R}_+ \qquad d(t) = Ae^{-t/\tau} \cos\left(\omega_0 \sqrt{1 - \frac{1}{\tau^2 {\omega_0}^2}} t + \phi\right)$$

où $A \in \mathbb{R}_+$ et $\phi \in \mathbb{R}$ dépendent des conditions initiales.

Objectif :

On mesure le déplacement aux instants $t \in 0, 0.5, 1, 1.5, 2, 2.5$ s, et on obtient les valeurs suivantes :

$$d(0) = 0.9, \quad d(0.5) = -0.6, \quad d(1) = -0.1, \quad d(1.5) = 0.3, \quad d(2) = -0.2, \quad d(2.5) = 0$$

On sait par ailleurs que

- Les capteurs de mesure présentent une incertitude de 0.05.
- La pulsation propre ω_0 est estimée proche de 4 rad/s.
- Le temps de relaxation τ est inférieur à 3 s.

Objectif : reconstruire les valeurs de ω_0 et τ afin de contrôler l'usure de l'oscillateur.

Nous devons alors résoudre les défis suivants :

- Sans tenir compte des indications, la solution n'est pas unique :
- Il est nécessaire de trouver comment incorporer des informations supplémentaires avec un degré de certitude variable
- Il faut prendre en compte le bruit des mesures
- Il faudrait réussir à quantifier la précision des paramètres obtenus

1.2 Cadre général

Le problème de l'oscillateur harmonique amorti peut être reformulé de manière plus générale comme un problème inverse en dimension finie :

Problème inverse (version déterministe) :

On s'intéresse à résoudre un problème inverse sous la forme $y = A(x)$, où :

- Les inconnues sont représentées par un vecteur $x \in \mathbb{R}^n$, où n est le nombre total d'inconnues

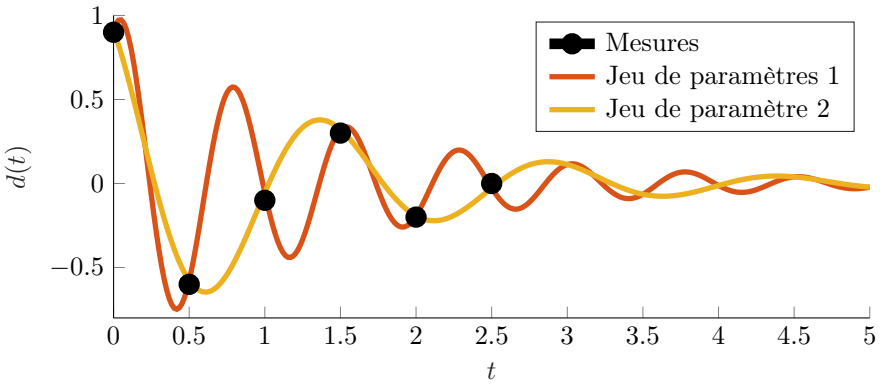

FIGURE 2: Oscillations amorties $d(t)$ pour deux jeux de paramètres différents. Jeu 1 : $A = 1$, $\phi = -0.44$, $\tau = 1.4$, $\omega_0 = 8.4$. Jeu 2 : $A = 1$, $\phi = 0.44$, $\tau = 1.4$, $\omega_0 = 4.2$.

- Les mesures observées sont stockées dans un vecteur $y \in \mathbb{R}^m$, où m représente le nombre de mesures
- Le modèle qui relie x à y est supposé connu et décrit par un opérateur $A : \mathbb{R}^n \to \mathbb{R}^m$.

Ce cadre est à la fois général et proche des situations rencontrées dans la pratique expérimentale. De plus, comme nous le verrons plus tard, n peut être très grand. Il est donc possible d'utiliser ce cadre pour reconstruire des fonctions continues : plutôt que de chercher directement une fonction $f : [a, b] \to \mathbb{R}$, on peut discrétiser l'intervalle $[a, b]$ en des points $t = (t_1, \ldots, t_n)$, puis chercher le vecteur d'inconnues $x = (f(t_1), \ldots, f(t_n))$, qui représente les valeurs de la fonction à ces points discrets.

Exercice 1 :

Dans le cas de l'oscillateur harmonique amorti, identifiez les vecteurs x, y et l'opérateur A.

Solution : $x = (A, \phi, \tau, \omega_0) \in \mathbb{R}^4$, $y = (0.9, -0.6, -0.1, 0.3, -0.2, 0) \in \mathbb{R}^6$,

$$d(x,t) = x_1 e^{-t/x_3} \cos\left(x_4 \sqrt{1 - \frac{1}{x_3^2 x_4^2}} t + x_2\right)$$

$$A(x) = (d(x,0), d(x,0.5), d(x,1), d(x,1.5), d(x,2), d(x,2.5))$$

1.3 Cadre déterministe

La méthode classique pour retrouver x à partir de y est déterministe et repose sur la minimisation des moindres carrés ;

$$x \in \operatorname{argmin}_{x \in \mathbb{R}^n} J(x), \qquad J(x) = \|A(x) - y\|$$

Cependant, plusieurs difficultés peuvent survenir lors de l'application de cette méthode :

- Que faire si J possède plusieurs minima ? Comment choisir entre eux ?
- Comment exploiter la forme spécifique de J ? On voit bien dans les Figures 3a-b) que deux distributions différentes de J peuvent malgré tout donner le même résultat pour x.
- Comment prendre en compte les erreurs de mesure dans le modèle ? On constate en Figure 3c) qu'à cause de bruit sur les mesures, x ne prend plus la valeur 1, alors qu'il y a une vaste région ou J est très proche du minimum.

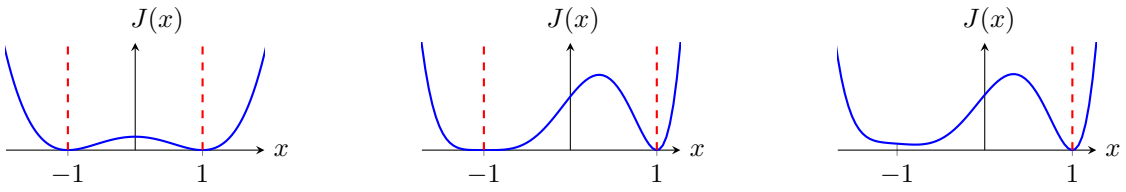

FIGURE 3: a) Deux minima globaux pour J. b) Deux minima globaux, avec des formes plus ou moins prononcées (piquées). c) Un unique minimum global, mais une vaste région où J reste proche de ce minimum.

Cette méthode ne prend pas non plus en compte certaines informations préalables sur le problème. Par exemple, dans le cas de l'oscillateur harmonique amorti, on sait que $\omega_0 \approx 4$. Une approche déterministe consiste alors à modifier la fonction de coût J en y ajoutant une pénalisation pour intégrer cette information (voir la méthode de Tikhonov). On minimise ainsi la fonction suivante :

$$\widetilde{J}(x) = J(x) + \lambda \|x_4 - 4\|$$

Cependant, plusieurs questions importantes subsistent :

- Comment interpréter physiquement le paramètre λ ? Quelle est son influence sur la solution ?
- Comment choisir les normes appropriées pour J et pour le terme de pénalisation ?

Dans ce cours, nous proposons de répondre à ces limitations en adoptant un cadre probabiliste et en modifiant le type de réponse attendue. Plutôt que de chercher une solution unique, nous autorisons plusieurs réponses possibles. Par exemple, dans la Figure 3, on pourrait interpréter les différents cas ainsi. Dans le cas a), il y aurait environ une chance sur deux que $x = 1$ ou $x = -1$. Dans le cas b), la probabilité que $x = -1$ est élevée, tandis qu'il y a une faible probabilité que $x = 1$. Dans le cas c), on pourrait dire que la zone $[-1.5, -0.8]$ est probable, mais il existe aussi une possibilité pour que x se situe autour de -1.

Une autre limitation du cadre déterministe réside dans les techniques d'optimisation, qui peuvent être mal adaptées à des problèmes de grande dimension. En effet, la fonctionnelle $\tilde{J}$ peut présenter plusieurs minima locaux, ce qui peut entraîner le blocage des algorithmes d'optimisation et ainsi empêcher une minimisation efficace. Par exemple, dans le cas de l'oscillateur harmonique amorti, on observe que la présence de ces minima locaux entrave la convergence des algorithmes d'optimisation classiques. Selon le point de départ de la minimisation, on obtient des jeux de paramètres très différents :

```
>>fminunc(tildeJ, [6, 0, 2, 4])
ans =
    2.4844   -1.2719    0.0503    4.0000
>>fminunc(tildeJ, [3, 0, 2, 4])
ans =
    1.0203    0.4753    1.3839    4.1362
```

Dans ce cours, nous introduirons également une méthode alternative basée sur des techniques d'intégration plutôt que sur l'optimisation, afin de fournir une estimation cohérente des paramètres x. Cette approche garantit l'unicité et la bonne définition du résultat du problème inverse.

2 Rappels de probabilité

2.1 Espace probabilisé

Pour travailler avec des probabilités, nous devons nous placer dans un espace probabilisé :

Définition 1 : Espace probabilisé

Un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ est constitué de

- Un ensemble Ω, appelé univers, qui représente l'ensemble des résultats possibles de l'ex-

périence considérée.

- Une famille $\mathcal{F}$ de parties de Ω, appelée σ-algèbre ou tribu, qui est stable par union, complémentaire, et contient Ω.
- Une fonction $\mathbb{P} : \mathcal{F} \to [0,1]$, appelée mesure de probabilité, telle que $\mathbb{P}(\Omega) = 1$ et qui vérifie la propriété suivante :

$$\forall (A_n)_{n\in\mathbb{N}} \in \mathcal{F}^{\mathbb{N}} \quad \text{tel que} \quad \forall i \neq j \quad A_i \cap A_j = \emptyset \quad \text{alors} \quad \mathbb{P}\left(\cup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$$

Exercice 2 :

Proposer un univers Ω adapté à l'exemple de l'oscillateur harmonique amorti.

Solution : On cherche à retrouver le vecteur de paramètres $x = (A, \phi, \tau, \omega_0)$. Une première proposition pour l'univers serait :

$$\Omega = \mathbb{R}_+ \times (-\pi, \pi) \times \mathbb{R}_+ \times \mathbb{R}_+$$

ce qui reflète les domaines possibles pour chacun des paramètres : A et τ sont positifs, ϕ est dans l'intervalle $(-\pi, \pi)$, et ω_0 est également positif. Cependant, on pourrait également définir l'univers de manière plus générale en prenant $\Omega = \mathbb{R}^4$, puis imposer des contraintes via la probabilité, par exemple en assignant une probabilité nulle à certaines régions non-physiques. Cela reviendrait à définir par exemple $\mathbb{P}(\mathbb{R}_- \times \mathbb{R}^3) = 0$.

Pour le choix de la tribu, si $\Omega = \mathbb{R}^n$ (ou un sous-ensemble de $\mathbb{R}^n$), la tribu la plus couramment utilisée est la tribu borélienne $\mathcal{B}(\mathbb{R}^n)$. Cette tribu est la plus petite tribu contenant, selon le choix, les intervalles, les ensembles ouverts ou les ensembles fermés.

Enfin, dans ce cours, nous nous concentrerons principalement sur des probabilités dites à densité :

Définition 2 : Densité de probabilité

Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction positive, mesurable (ou, plus simplement, continue par morceaux) et telle que son intégrale sur $\mathbb{R}^n$ soit égale à 1. On définit alors une probabilité $\mathbb{P}$ en posant

$$\forall A \in \mathcal{B}(\mathbb{R}^n) \qquad \mathbb{P}(A) = \int_A f(x)\mathrm{d}x$$

Une telle probabilité est dite à densité, et f est appelée la densité de probabilité de $\mathbb{P}$.

Exercice 3 :

- Donner la densité de probabilité pour une loi uniforme sur l'intervalle $[a, b]$, notée $\mathcal{U}(a, b)$.
- Donner la densité de probabilité pour une loi normale d'espérance μ et d'écart type σ, notée $\mathcal{N}(\mu, \sigma)$.
- Donner la densité de probabilité pour une loi normale multidimensionnelle avec espérance $\mu \in \mathbb{R}^n$ et matrice de covariance $\Sigma \in \mathcal{M}_n(\mathbb{R})$, notée $\mathcal{N}(\mu, \Sigma)$.

Solution : Loi uniforme :

$$f(x) = \frac{1}{b-a}\mathbf{1}_{a<x<b}$$

Loi normale :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Loi normale multidimensionnelle :

$$f(x) = \frac{1}{(2\pi)^{n/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

Dans un espace probabilisé, il est possible de définir des propositions qui sont vraies "presque partout" :

Définition 3 : Presque partout

Une proposition est dite vraie presque partout (abrégé p.p.) si elle est vraie sur l'ensemble $\Omega \setminus A$, où $A \in \mathcal{F}$ est un ensemble tel que la mesure de probabilité de A est nulle, c'est-à-dire $\mathbb{P}(A) = 0$.

Enfin, une fois l'espace probabilisé défini, nous pouvons travailler avec des variables aléatoires :

Définition 4 : Variable aléatoire

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et V un espace de Banach. Une fonction mesurable $X : \Omega \to V$ est appelée une variable aléatoire (VA).

Dans l'exemple de l'oscillateur, les variables aléatoires qui nous intéressent sont les coordonnées des paramètres du modèle. Par exemple, la variable aléatoire $A : \Omega \to \mathbb{R}$, définie par $A(\omega) = \omega \cdot (1, 0, 0, 0)$, correspond au paramètre d'amplitude du problème.

On utilise habituellement la notation X pour désigner la réalisation d'une variable aléatoire, plutôt que $X(\omega)$. La probabilité associée à X est aussi souvent notée $\mathbb{P}_X$ et est définie par :

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) \in B\})$$

On dit que X suit une loi μ (noté $X \sim \mu$) si $\mathbb{P}_X = \mu$. Si $\mathbb{P}_X$ est à densité, on note sa densité par π_X.

Pour deux variables aléatoires X et Y, la loi jointe de (X, Y) est définie par :

$$\mathbb{P}_{(X,Y)}(A \times B) = \mathbb{P}(X \in A, Y \in B)$$

Les lois marginales de X et Y sont les lois obtenues en considérant les distributions individuelles de X et Y respectivement. Enfin, pour mieux comprendre la distribution d'une variable aléatoire, il est intéressant de regarder son espérance et sa variance :

Définition 5 : Espérance et Variance

Soit $X : \Omega \to V$ une variable aléatoire. L'espérance de X existe si $\int_\Omega X(\omega) \mathrm{d}\mathbb{P}(\omega) < +\infty$. Dans ce cas, l'espérance de X, notée $\mathbb{E}(X)$, est définie par :

$$\mathbb{E}(X) = \int_\Omega X(\omega) \mathrm{d}\mathbb{P}(\omega) = \int_V x \, \mathrm{d}\mathbb{P}_X(x)$$

Si X^2 possède une espérance, la variance de X, notée $V(X)$ ou $\mathrm{Var}(X)$, est définie par :

$$V(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2$$

2.2 Probabilité conditionnelle

Pour utiliser le cadre bayésien, il est essentiel de définir les probabilités conditionnelles. En première approche, les probabilités conditionnelles sont généralement définies comme suit :

Définition 6 : Probabilité conditionnelle

Soient A et B deux événements dans $\mathcal{F}$ avec $\mathbb{P}(B) > 0$. La probabilité conditionnelle de A sachant B est définie par :

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Si $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, cela signifie que B n'a pas d'influence sur A, et on dit que A et B sont indépendants. Cette indépendance peut également être exprimée de manière symétrique :

Définition 7 : Indépendance

Deux événements A et B dans $\mathcal{F}$ sont indépendants si :

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Exercice 4 :

Soit Y une variable aléatoire qui prend les valeurs $1/3$ ou $2/3$ avec probabilité uniforme. Soit X une variable aléatoire telle que X prend la valeur 1 avec probabilité Y et la valeur 0 avec probabilité $1 - Y$. Que vaut $\mathbb{P}(X = 1 \mid Y = 1/3)$? Que dire de $\mathbb{P}(X = 1 \mid Y = 1/2)$? Et si $Y \sim \mathcal{U}(0, 1)$?

Solution : Par définition, $\mathbb{P}(X = 1 \mid Y = 1/3) = 1/3$. Mais la probabilité conditionnelle $\mathbb{P}(X = 1 \mid Y = 1/2)$ ne peut pas être déterminée directement, car $\mathbb{P}(Y = 1/2) = 0$. De même, si $Y \sim \mathcal{Y}(0, 1)$, il n'est possible de calculer aucunes des deux probabilités.

Pour résoudre le problème précédent, nous pouvons proposer une définition alternative pour la probabilité conditionnelle. Soit $X : \Omega \to V$ et $Y : \Omega \to W$ deux variables aléatoires. Avec la définition précédente et en supposant que $\mathbb{P}(Y = y) \neq 0$ pour $y \in Y(\Omega)$, on peut écrire pour tout $B \in \mathcal{B}(V)$:

$$\mathbb{P}(X \in B, Y = y) = \mathbb{P}(X \in B | Y = y)\mathbb{P}(Y = y)$$

En intégrant cette expression pour tout $y \in A$, où $A \in \mathcal{B}(W)$, on obtient :

$$\mathbb{P}(X \in B, Y \in A) = \int_A \mathbb{P}(X \in B | Y = y) \mathrm{d}\mathbb{P}_Y(y)$$

Cette relation permet de définir la version régulière de la probabilité conditionnelle :

Définition 8 : Version régulière de la probabilité conditionnelle

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et $X : \Omega \to V$, $Y : \Omega \to W$ deux variables aléatoires sur des espaces de Banach séparables V et W. Une fonction $\pi_{X|Y} : \mathcal{B}(V) \times W \to [0, 1]$ est appelée la version régulière de la probabilité conditionnelle de X sachant Y si elle satisfait les conditions suivantes :

- Pour tout $B \in \mathcal{B}(V)$, la fonction $y \mapsto \pi_{X|Y}(B, y)$ est mesurable par rapport à $\mathcal{B}(W)$.
- Pour tout $y \in Y(\Omega)$, la fonction $B \mapsto \pi_{X|Y}(B, y)$ est une distribution de probabilité sur $(V, \mathcal{B}(V))$.
- Pour tout $A \in \mathcal{B}(W)$ et tout $B \in \mathcal{B}(V)$, la relation suivante est vérifiée :

$$\mathbb{P}(X \in B, Y \in A) = \int_A \pi_{X|Y}(B, y) \mathrm{d}\mathbb{P}_Y(y)$$

On note $\mathbb{P}(X \in B \mid Y = y) = \pi_{X|Y}(B, y)$.

On peut démontrer qu'une telle fonction $\pi_{X|Y}$ existe et est unique sous certaines conditions techniques, notamment lorsque X et Y prennent leurs valeurs dans des espaces de Banach séparables. Cette existence et unicité garantissent la cohérence des probabilités conditionnelles dans le cadre probabiliste.

Pour conclure cette section, nous introduisons la formule fondamentale du cadre bayésien : la formule de Bayes.

Proposition 1 : Formule de Bayes

Soient $A, B \in \mathcal{F}$ deux événements tels que $\mathbb{P}(B) > 0$, alors :

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

Lorsque $\mathbb{P}(B) = 0$, on doit recourir à la version régulière de la probabilité conditionnelle pour définir correctement les probabilités conditionnelles.

Proposition 2 : Formule de Bayes (version régulière)

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et $X : \Omega \to \mathbb{R}^n$, $Y : \Omega \to \mathbb{R}^m$ deux variables aléatoires avec une densité jointe $\pi_{X,Y}$ et des densités marginales π_X et π_Y. Si $\pi_{X|Y}(\cdot \mid y)$ et $\pi_{Y|X}(\cdot \mid x)$ désignent les densités respectives de $\pi_{X|Y}(\cdot, y)$ et $\pi_{Y|X}(\cdot, x)$, alors $\mathbb{P}_Y$-presque partout :

$$\pi_{X|Y}(x \mid y) = \frac{\pi_{Y|X}(y \mid x)\pi_X(x)}{\pi_Y(y)}$$

3 Cadre bayésien

3.1 Prior, posterior, vraissemblance

Pour rappel, nous nous intéressons au problème inverse où nous cherchons à retrouver les paramètres x à partir de mesures y, en résolvant l'équation $y = Ax$. L'approche probabiliste nous permet de ne plus supposer que x est unique, mais de le considérer comme une variable aléatoire, dont nous cherchons la loi de probabilité. Cela reformule le problème inverse dans un cadre bayésien :

Problème inverse (version bayésienne) :

On considère le problème $Y = A(X) + E$, avec les éléments suivants :

- $X : \Omega \to V$ est une variable aléatoire représentant les paramètres inconnus du problème. On suppose que X suit une loi de densité π_X.
- $E : \Omega \to W$ est une variable aléatoire modélisant le bruit sur les mesures. Elle est supposée indépendante de X et suit une densité π_E, appelée bruit d'observation.
- $Y : \Omega \to W$ est une variable aléatoire correspondant aux mesures effectuées. Elle suit une densité π_Y.
- $A : V \to W$ est un opérateur, appelé opérateur direct, qui relie les paramètres X aux observations Y.

Dans le cadre bayésien, on observe une réalisation de Y (notée $Y(\omega)$ pour un $\omega \in \Omega$), et l'objectif est de déterminer la distribution de X, conditionnée par l'observation de $Y = y$. Cela revient à retrouver $\mathbb{P}(X \mid Y = y)$. Voici quelques terminologies usuelles dans le cadre bayésien :

Définition 9 : Prior, vraisemblance, et posterior

- Distribution a priori (prior) : $\mathbb{P}_X$ est appelée distribution a priori. Elle représente l'information disponible sur X **avant** d'effectuer les mesures ou les observations. Sa densité est notée $\pi_X(x)$.
- Vraisemblance : $\mathbb{P}(Y \mid X = x) = \pi_{Y|X}(\cdot, x)$ est appelée la vraisemblance. Elle modélise la probabilité d'observer les données Y en fonction des valeurs des paramètres X. On suppose qu'elle a pour densité $\pi_{Y|X}(y \mid x)$.
- Distribution a posteriori (posterior) : $\mathbb{P}(X \mid Y = y) = \pi_{X|Y}(\cdot, y)$ est appelée distribution a posteriori. Elle met à jour la distribution de X en tenant compte de la nouvelle observation $Y = y$. On suppose qu'elle a pour densité $\pi_{X|Y}(x \mid y)$.

La vraisemblance quantifie, pour un paramètre x fixé, la probabilité d'obtenir une observation $Y = y$. Si l'on fait varier x tout en gardant y constant, la valeur de x qui maximise $\pi_{Y|X}(y \mid x)$ est celle qui correspond le mieux aux données observées. C'est cette valeur qui "explique" au mieux les observations.

Rappelons que, selon la formule de Bayes, on a :

$$\pi_{X|Y}(x \mid y) = \frac{\pi_{Y|X}(y \mid x)\pi_X(x)}{\pi_Y(y)}$$

Autrement dit,

$$\text{posterior} \propto \text{vraissemblance} \times \text{prior}$$

La vraisemblance intègre les informations provenant des données observées et permet d'actualiser la distribution a priori. Ces informations sont ensuite combinées dans la distribution a posteriori, qui constitue la mise à jour de notre connaissance sur X après avoir observé Y.

3.2 Estimateurs

Ainsi, toute l'information disponible sur X, conditionnée à l'observation de $Y = y$, est contenue dans la distribution postérieure. Cependant, il peut être utile, voire nécessaire, de donner une seule valeur pour X au lieu de décrire l'ensemble de la distribution. Cette valeur pourrait être interprétée comme la meilleure estimation de X, étant donné les observations. Cela est particulièrement pertinent en grande dimension, où représenter une distribution complète peut être complexe. De plus, certaines contraintes extérieures peuvent exiger une solution déterministe, c'est-à-dire une valeur fixe pour x.

Pour répondre à cette nécessité, plusieurs estimateurs peuvent être considérés comme des candidats pour représenter la "valeur la plus probable" de X :

Définition 10 : Estimateurs

- Maximum de vraisemblance (Maximum Likelihood, ML) : Il s'agit de l'estimation de X qui maximise la vraisemblance, autrement dit, celle qui rend l'observation $Y = y$ la plus probable :
$$x_{ML} \in \operatorname{argmax}_{x \in V} \pi_{Y|X}(y \mid x)$$
- Maximum a posteriori (MAP) : Cet estimateur maximise la distribution postérieure. Il prend en compte non seulement la vraisemblance, mais aussi l'information a priori sur X :
$$x_{MAP} \in \operatorname{argmax}_{x \in V} \pi_{X|Y}(x \mid y)$$
- Moyenne conditionnelle (Conditional Mean, CM) : Cet estimateur correspond à l'espérance conditionnelle de X, donnée par la moyenne pondérée de X sous la distribution postérieure :
$$x_{CM} = \mathbb{E}(X \mid Y = y) = \int_V x\, \pi_{X|Y}(x \mid y) \mathrm{d}x$$

Exercice 5 :

Soit la fonction $\phi(x)$ définie par :

$$\phi(x) = \mathbf{1}_{0 \leq x \leq 1}(1 - x) + \mathbf{1}_{-1 \leq x < 0}(1 + x)$$

et supposons que, pour un paramètre $y \in (0, 1)$ et des écarts types $\sigma_1, \sigma_2 \in (0, 0.5)$, la densité conditionnelle de X sachant $Y = y$ soit donnée par :

$$\pi_{X|Y}(x \mid y) = \frac{y}{\sigma_1}\phi\left(\frac{x}{\sigma_1}\right) + \frac{1-y}{\sigma_2}\phi\left(\frac{x-1}{\sigma_2}\right)$$

Calculer x_{CM} et x_{MAP} en fonction de y, σ_1, et σ_2. Calculer la variance σ^2 de la distribution postérieure $\pi_{X|Y}$.

Solution :

$$x_{CM} = \frac{1-y}{\sigma_2}\int_{1-\sigma_2}^{1+\sigma_2} x\phi\left(\frac{x-1}{\sigma_2}\right)\mathrm{d}x = (1-y)\int_{-1}^{1}(x+1)\phi(x)\mathrm{d}x = 1 - y$$

$$x_{MAP} = \begin{cases} 0 & \text{si } y/\sigma_1 > (1-y)/\sigma_2 \\ 1 & \text{sinon} \end{cases}$$

On a ensuite

$$\begin{aligned}
\sigma^2 &= \int_{\mathbb{R}} (x - x_{CM})^2 \pi_{X|Y}(x \mid y) \mathrm{d}x && (1)\\
&= \int_{\mathbb{R}} x^2 \pi_{X|Y}(x \mid y) \mathrm{d}x - {x_{CM}}^2 && (2)\\
&= y\sigma_1^2 \int_{-1}^{1} x^2 \phi(x) \mathrm{d}x + (1-y) \int_{-1}^{1} (\sigma_2 x + 1)^2 \phi(x) \mathrm{d}x - (1-y)^2 && (3)\\
&= \frac{y\sigma_1^2 + (1-y)\sigma_2^2}{6} + (1-y) - (1-y)^2 && (4)
\end{aligned}$$

Les solutions de l'exercice précédent conduisent aux figures suivantes, qui illustrent la forme des densités conditionnelles $\pi_{X|Y}(x \mid y)$ ainsi que les valeurs de x_{CM} et x_{MAP} pour différents paramètres y, σ_1 et σ_2 :

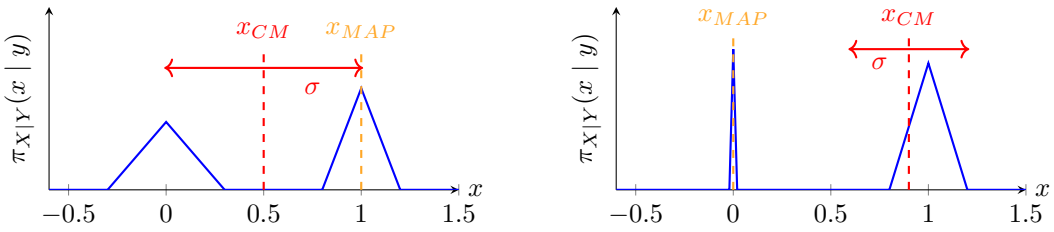

FIGURE 4: a) $\sigma_1 = 0.3$, $\sigma_2 = 0.2$, $y = 0.5$. b) $\sigma_1 = 0.02$, $\sigma_1 = 0.2$, $y = 0.1$. On représente les estimateurs CM et MAP et la variance σ de la distribution à posteriori.

On remarque également que :

- Le calcul numérique des différents estimateurs est très distinct : le maximum de vraisemblance (ML) et le maximum a posteriori (MAP) nécessitent la résolution d'un problème d'optimisation, tandis que la moyenne conditionnelle (CM) requiert le calcul d'une intégrale (voire deux).
- Les estimateurs MAP et CM s'appuient sur la distribution postérieure, alors que l'estimateur ML ne considère que la vraisemblance.
- Les estimateurs ML et MAP ne sont pas nécessairement uniques, tandis que l'estimateur CM est toujours unique.
- L'estimateur CM est plus stable que les autres et est moins sensible aux petites variations de la distribution postérieure, contrairement à l'estimateur MAP qui peut réagir fortement à ces changements.
- L'estimateur CM ne correspond pas forcément à un point de forte probabilité dans la distribution postérieure. Dans ce cas, une variance élevée signale que la confiance dans l'estimation est faible.

Ces différentes quantités sont toutes interconnectées et peuvent être analysées à l'aide du théorème de Bayes

Proposition 3 : Expression de la loi postérieure

Avec les hypothèses de la Proposition 2, on a $\mathbb{P}_Y$-presque partout :

$$\pi_{X|Y}(x \mid y) = \frac{1}{Z(y)} \pi_E(y - A(x)) \pi_X(x)$$

où

$$Z(y) = \int_{\mathbb{R}^n} \pi_E(y - A(x)) \pi_X(x) \mathrm{d}x$$

Preuve : Par le théorème de Bayes, nous avons déjà, $\mathbb{P}_Y$-presque partout :

$$\pi_{X|Y}(x \mid y) = \frac{\pi_{Y|X}(y \mid x)\pi_X(x)}{\pi_Y(y)}$$

Montrons que $\pi_{Y|X}(y \mid x) = \pi_E(y - A(x))$. Pour cela, considérons deux ensembles $B_1 \in \mathcal{B}(W)$ et $B_2 \in \mathcal{B}(V)$. On a :

$$\begin{aligned}
\mathbb{P}(Y \in B_1, X \in B_2) &= \mathbb{P}(A(X) + E \in B_1, X \in B_2) && (5)\\
&= \int_V \int_W \mathbf{1}_{B_1}(A(x) + e)\mathbf{1}_{B_2}(x)\pi_E(e)\pi_x(x)\mathrm{d}e\mathrm{d}x && (6)\\
&= \int_V \mathbf{1}_{B_2}(x) \left(\int_W \mathbf{1}_{B_1}(A(x) + e)\pi_E(e)\mathrm{d}e \right) \pi_X(x)\mathrm{d}x && (7)\\
&= \int_{B_2} \left(\int_W \mathbf{1}_{B_1}(\varepsilon)\pi_E(\varepsilon - A(x))\mathrm{d}\varepsilon \right) \mathrm{d}\mathbb{P}_X(x) && (8)
\end{aligned}$$

Par définition, $\pi_{Y|X}(B_1, x)$ est l'unique fonction qui satisfait :

$$\mathbb{P}(Y \in B_1, X \in B_2) = \int_{B_2} \pi_{Y|X}(B_1, x)\mathrm{d}\mathbb{P}_X(x)$$

ce qui implique que

$$\pi_{Y|X}(B_1, x) = \int_{B_1} \pi_E(e - A(x))\mathrm{d}e$$

De plus, par définition, $\pi_{Y|X}(y \mid x)$ est l'unique fonction telle que :

$$\pi_{Y|X}(B_1, x) = \int_{B_1} \pi_{Y|X}(y \mid x)\mathrm{d}y$$

Ainsi, nous en déduisons que $\pi_{Y|X}(y \mid x) = \pi_E(y - A(x))$. Ensuite, pour obtenir $\pi_Y(y)$, nous avons :

$$\begin{aligned}
\mathbb{P}(Y \in B_1) &= \mathbb{P}(Y \in B_1, X \in \mathbb{R}^n) = \int_{\mathbb{R}^n} \pi_{Y|X}(B_1, x)\mathrm{d}\mathbb{P}_X(x) && (9)\\
&= \int_{\mathbb{R}^n} \int_{B_1} \pi_{Y|X}(y \mid x)\mathrm{d}y\pi_X(x)\mathrm{d}x && (10)\\
&= \int_{B_1} \int_{\mathbb{R}^n} \pi_E(y - A(x))\pi_X(x)\mathrm{d}x\mathrm{d}y && (11)
\end{aligned}$$

Par conséquent, $\mathbb{P}(Y \in B_1) = \int_{B_1} Z(y)\mathrm{d}y$, et donc, par définition de la densité, $\pi_Y(y) = Z(y)$. ■

On remarque que l'expression du postérieur donnée n'est valable que si $Z(y) \neq 0$. En pratique, une valeur de y pour laquelle $Z(y) = 0$ indique une incohérence entre le modèle et les données observées. Ainsi, lorsque $Z(y)$ est nul, il est important de vérifier la compatibilité des observations avec les hypothèses du modèle ou d'explorer des ajustements possibles aux distributions pour éviter de telles situations.

Exercice 6 :

Soit $E \sim \mathcal{N}(0, I_m)$ et supposons que l'information a priori sur X est $X \sim \mathcal{N}(0, I_n/\alpha)$, où $\alpha > 0$. Déterminez les expressions des estimateurs x_{ML}, x_{MAP} et x_{CM}.

Solution : Pour le maximum de vraisemblance, on constate que

$$\pi_{Y|X}(y \mid x) = \pi_E(y - A(x)) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} \|A(x) - y\|^2 \right)$$

Donc

$$x_{ML} = \mathrm{argmax}_{x \in \mathbb{R}^n} \pi_{Y|X}(y \mid x) = \mathrm{argmin}_{x \in \mathbb{R}^n} \|A(x) - y\|^2$$

On sait aussi que

$$\pi_X(x) = \frac{\alpha^{n/2}}{(2\pi)^{n/2}} \exp\left(-\frac{\alpha \|x\|^2}{2} \right)$$

Donc

$$x_{MAP} = \text{argmax}_{x\in\mathbb{R}^n} \pi_{X|Y}(x \mid y) = \text{argmax}_{x\in\mathbb{R}^n} \pi_E(y - A(x))\pi_X(x) = \text{argmin}_{x\in\mathbb{R}^n} \left(\|A(x) - y\|^2 + \alpha\|x\|^2\right)$$

Enfin,

$$x_{CM} = \int_{\mathbb{R}^n} x\pi_{X|Y}(x \mid y)\mathrm{d}x = \frac{\alpha^{n/2}}{Z(y)(2\pi)^{m/2+n/2}} \int_{\mathbb{R}^n} x \exp\left(-\frac{1}{2}\|A(x) - y\|^2 - \frac{\alpha}{2}\|x\|^2\right) \mathrm{d}x$$

On observe que l'estimation ML n'est pas de nature bayésienne : la solution x_{ML} est indépendante du choix de la loi a priori et correspond à la minimisation des moindres carrés. En revanche, l'estimateur MAP s'apparente à une régularisation de Tikhonov, où l'information apportée par la loi a priori est interprétée comme une forme de régularisation. Par ailleurs, l'estimateur CM semble plus complexe à calculer directement, car il nécessite le calcul de deux intégrales sur $\mathbb{R}^n$.

3.3 Choix de priors

Pour conclure, nous proposons de discuter quelques choix usuels de priors. On distingue principalement deux types de priors : les priors informatifs et les priors peu informatifs.

Priors informatifs : Ces priors reposent sur une connaissance préalable spécifique de l'objet étudié.

- Un choix fréquent est le prior gaussien, car il est facile à manipuler et permet des calculs explicites. De plus, selon le théorème central limite, il constitue une bonne approximation lorsque l'on dispose d'un grand nombre d'observations indépendantes. Par exemple, pour modéliser le bruit, on utilise presque toujours une distribution gaussienne.
- Le prior Gamma est souvent utilisé pour les variables à valeurs strictement positives, notamment dans le cadre de modélisation de taux. Il est défini par $\pi_X(x) \propto x^{\alpha-1}\exp(-x/\beta)$
- Le prior de Wishart est utilisé pour modéliser les matrices de covariance, par exemple dans les modèles de régression.

Priors non (ou peu) informatifs : Ces priors sont utilisés lorsque l'on dispose de peu ou pas d'informations préalables sur les paramètres à estimer.

- Un prior uniforme est souvent choisi lorsqu'on souhaite rester neutre quant à la valeur d'un paramètre. Cependant, il est nécessaire de spécifier des bornes, car il est impossible de définir une distribution uniforme sur tout $\mathbb{R}$. De plus, ce prior peut introduire une forme de biais, car si $X \sim \mathcal{U}(0,1)$, alors X^2 ne suit pas une distribution uniforme.
- Le prior de Jeffreys est construit à partir de la fonction de vraisemblance de manière à être invariant par transformation des paramètres.

Autres considérations : Les priors conjugués sont également souvent utilisés, car ils permettent de simplifier les calculs tout en offrant une plus grande flexibilité que la loi normale en termes de paramétrisation.

Exercice 7 :

Proposer des priors dans le cas de l'oscillateur harmonique amorti.

Solution : Comme nous n'avons aucune information préalable sur la phase, un prior uniforme sur l'intervalle $[-\pi, \pi]$ est approprié :

$$\pi_\phi(\phi) \propto \mathbf{1}_{-\pi\leq\phi\leq\pi}$$

D'après la première mesure à $t = 0$ qui donne $d(0) = 0.9$, nous pouvons supposer que l'amplitude A est située dans un intervalle raisonnable, par exemple entre 0.5 et 10 :

$$\pi_A(A) \propto \mathbf{1}_{0.5\leq A\leq 5}$$

Sachant que τ est positif et que $\tau \leq 3$, un prior uniforme sur cet intervalle semble approprié :

$$\pi_\tau(\tau) \propto \mathbf{1}_{0<\tau\leq 3}$$

Nous savons que ω_0 est positif et proche de 4. Un prior gaussien centré sur 4 avec une décroissance exponentielle semble adapté :

$$\pi_{\omega_0}(\omega_0) \propto \mathbf{1}_{\omega_0>0} \exp\left(-\frac{(\omega_0 - 4)^2}{2}\right)$$

4 Méthodes numérique MCMC

Nous disposons désormais d'une expression explicite de la distribution a posteriori, notée désormais Π, avec pour densité π. Dans certains cas, la densité π n'est connue qu'à une constante multiplicative près. Dans ce contexte, on désigne par μ la mesure non normalisée. En dimension 1, le calcul numérique de x_{CM} et la visualisation de μ sont relativement simples, car il suffit d'estimer une intégrale discrétisée :

$$\int_{\mathbb{R}} f(x)\mu(x)\mathrm{d}x \approx \sum_{j=1}^{N} \omega_j f(x_j)\mu(x_j)$$

où les x_j sont des points judicieusement choisis dans $\mathbb{R}$ (généralement de manière régulière) et les ω_j sont des poids d'intégration associés. Toutefois, des défis numériques non négligeables apparaissent dès que l'on passe en grande dimension, surtout si l'on applique une approche naïve :

- Le calcul de x_{CM} nécessite l'évaluation de deux intégrales sur $\mathbb{R}^n$.
- La visualisation de la distribution exige un grand nombre d'évaluations de μ, ce qui peut rapidement devenir coûteux.

L'objectif de cette section est donc de proposer une méthode plus efficace, réduisant le nombre d'évaluations nécessaires de μ pour à la fois visualiser la distribution et calculer x_{CM}. Nous adopterons une approche probabiliste : au lieu de quadriller uniformément l'espace $\mathbb{R}^n$, nous chercherons à positionner intelligemment les points x_j de manière à obtenir une approximation précise de l'intégrale avec un nombre réduit de points (voir une illustration en Figure 5).

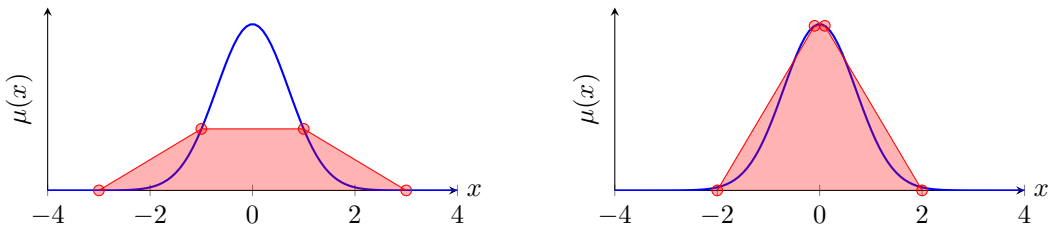

FIGURE 5: À gauche, les points x_j sont uniformément distribués, tandis qu'à droite, les points x_j sont répartis de manière à suivre la distribution cible.

Si l'on suppose que les points x_j sont générés selon la loi Π, on obtient alors avec une convergence assez rapide l'approximation suivante :

$$\int_{\mathbb{R}^n} f(x)\pi(x)\mathrm{d}x \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j)$$

où les x_j sont des échantillons tirés de la distribution Π. Cette approche est connue sous le nom de méthode de Monte-Carlo et est couramment utilisée pour estimer des intégrales lorsque la forme analytique exacte est difficile à obtenir. La question qui se pose est donc la suivante : comment générer une séquence de points qui suit la loi Π ? Pour cela, il est nécessaire de rappeler quelques notions de probabilités, en particulier celles concernant les chaînes de Markov.

4.1 Chaîne de Markov

Définition 11 : Chaîne de Markov

Une chaîne de Markov sur $\mathbb{R}^n$ est une suite de variables aléatoires $X_j : \Omega \to \mathbb{R}^n$ telle que, pour tout $j \in \mathbb{N}$ et pour tout $B \in \mathcal{B}(\mathbb{R}^n)$, on a :

$$\mathbb{P}(X_{j+1} \in B \mid X_1, \cdots, X_j) = \mathbb{P}(X_{j+1} \in B \mid X_j) \qquad \mathbb{P} - \text{p.p.}$$

Une chaîne de Markov décide du prochain état uniquement en fonction de l'état actuel, ce qui en fait un processus sans mémoire. Nous nous intéresserons particulièrement au cas des chaînes de Markov homogènes :

Définition 12 : Noyau de transition

Une fonction $K : \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0,1]$ est appelée noyau de transition si elle satisfait les deux propriétés suivantes :

- Pour tout $B \in \mathcal{B}(\mathbb{R}^n)$, la fonction $x \mapsto K(x,B)$ est mesurable.
- Pour tout $x \in \mathbb{R}^n$, la fonction $B \mapsto K(x,B)$ définit une distribution de probabilité sur $\mathcal{B}(\mathbb{R}^n)$.

Une chaîne de Markov $(X_j)_{j\in\mathbb{N}}$ est dite homogène s'il existe un noyau de transition K tel que, pour tout $j \in \mathbb{N}$, $x \in \mathbb{R}^d$ et $B \in \mathcal{B}(\mathbb{R}^d)$, on ait :

$$\mathbb{P}(X_{j+1} \in B \mid X_j = x) = K(x,B) \qquad \mathbb{P}-p.p.$$

Dans la suite, on suppose que le noyau K admet une densité k, telle que :

$$\forall B \in \mathcal{B}(\mathbb{R}^d) \quad K(x,B) = \int_B k(x,y)\mathrm{d}y$$

Le noyau de transition K décrit comment passer d'un état actuel à un autre, selon une certaine probabilité. Nous introduisons également les notations suivantes :

Définition 13 :

Soit K un noyau de transition et Π une mesure de probabilité sur $\mathbb{R}^n$. On définit la probabilité ΠK sur $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ par :

$$\forall B \in \mathcal{B}(\mathbb{R}^d) \quad (\Pi K)(B) = \int_{\mathbb{R}^d} K(x,B)\pi(x)\mathrm{d}x$$

On définit également, pour $j > 0$, le noyau de transition K^j par récurrence :

$$K^j(x,B) = \int_{\mathbb{R}^d} K^{j-1}(y,B)\, k(x,y)\mathrm{d}y$$

De manière équivalente, si $X_1 \sim \Pi$, alors $X_j \sim \Pi K^{j-1}$.

À titre d'information, l'expression ΠK est historiquement utilisée à la place de $K\Pi$, car dans le cas discret, les lois étaient représentées par des vecteurs-lignes, et K était simplement une matrice de transition. On introduit ensuite quelques propriétés fondamentales concernant les noyaux de transition et les chaînes de Markov :

Définition 14 : Invariance, irréductibilité, périodicité

- Une mesure Π est dite invariante par rapport à un noyau de transition K si elle vérifie :

  $$\Pi = \Pi K$$

  c'est-à-dire que Π reste inchangée après application de K.
- Un noyau K est dit irréductible par rapport à Π si, pour tout $x \in \mathbb{R}^n$ et tout $B \in \mathcal{B}(\mathbb{R}^n)$ tel que $\Pi(B) > 0$, il existe un entier $j \in \mathbb{N}$ tel que :

  $$K^j(x,B) > 0.$$

  Cela signifie que, quel que soit le point de départ de la chaîne, elle peut atteindre tout B avec probabilité non nulle après un certain nombre d'itérations.
- Un noyau irréductible K est dit périodique s'il existe un entier $k \geq 2$ et des ensembles disjoints non vides $\{E_1, \ldots, E_m\} \subset \mathbb{R}^n$ tels que :

  $$\forall j = 1, \ldots, m, \quad \forall x \in E_j, \quad K(x, E_{(j+1) \mod m}) = 1.$$

  Cela signifie que la chaîne de Markov évolue de manière cyclique à travers les ensembles $E_1, \ldots, E_m$.

Ces notions permettent d'énoncer un théorème fondamental pour les méthodes MCMC :

Proposition 4 :

Soit Π une mesure de probabilité de densité π sur $\mathbb{R}^n$, et soit $\{X_j\}_{j\geq 1}$ une chaîne de Markov homogène de noyau K. Supposons que Π est invariante par rapport à K et que K est irréductible et apériodique. Alors, pour tout $X \in \mathbb{R}^n$ et tout $B \in \mathcal{B}(\mathbb{R}^n)$, on a :

$$\lim_{N\to+\infty} K^N(X, B) = \Pi(B)$$

De plus, pour toute fonction $f \in \mathrm{L}^1_\Pi(\mathbb{R}^n)$,

$$\lim_{N\to+\infty} \frac{1}{N}\sum_{j=1}^{N} f(X_j) = \int_{\mathbb{R}^n} f(x)\pi(x)\mathrm{d}x \qquad \text{p.s.}$$

L'objectif est donc de trouver un noyau de transition K tel que la mesure Π soit invariante sous K, afin d'appliquer le résultat précédent. Une manière classique de construire un tel noyau repose sur la notion de réversibilité, définie comme suit :

Définition 15 : Réversibilité

Un noyau de transition K est dit Π-réversible s'il satisfait la condition de balance détaillée suivante :

$$\forall x, y \in \mathbb{R}^n, \quad k(x, y)\pi(x) = k(y, x)\pi(y)$$

Cette condition garantit que, sous l'hypothèse que $X_j \sim \Pi$, la probabilité de transition de x vers y est égale à celle de la transition de y vers x. En d'autres termes, le processus est symétrique en termes de probabilité. On peut également reformuler cette condition sous la forme suivante :

$$\forall A, B \in \mathcal{B}(\mathbb{R}^n) \quad \mathbb{P}(X_j \in A, X_{j+1} \in B) = \mathbb{P}(X_j \in B, X_{j+1} \in A)$$

Proposition 5 :

Si K est un noyau de transition Π-réversible, alors Π est invariante par rapport à K.

Preuve : Si $A \in \mathcal{B}(\mathbb{R}^n)$, alors

$$\begin{aligned}
(\Pi K)(A) &= \int_{\mathbb{R}^n} K(x, A)\pi(x)\mathrm{d}x \\
&= \int_{\mathbb{R}^n}\int_A q(x, y)\pi(x)\mathrm{d}y\mathrm{d}x \\
&= \int_{\mathbb{R}^n}\int_A q(y, x)\pi(y)\mathrm{d}y\mathrm{d}x \\
&= \int_A\int_{\mathbb{R}^n} q(y, x)\pi(y)\mathrm{d}x\mathrm{d}y \\
&= \int_A K(x, \mathbb{R}^n)\pi(y)ddy = \int_A 1\pi(y)\mathrm{d}y = \Pi(A)
\end{aligned}$$

L'objectif est donc de construire un noyau de transition K qui soit à la fois irréductible, apériodique et Π-réversible, garantissant ainsi que la chaîne de Markov converge vers la distribution invariante Π.

4.2 Algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings, introduit en 1953 par Metropolis et ses collaborateurs, puis généralisé en 1970 par Hastings, est un outil fondamental en statistique computationnelle. Il est reconnu comme l'un des 10 algorithmes les plus influents du XXe siècle. L'idée centrale pour la construction d'une chaîne de Markov dont la loi tendrait vers Π consiste à partir d'un point x,

à proposer un nouveau point y selon une loi de transition $K(x, y)$, puis à décider d'accepter ou de rejeter cette proposition en fonction de la valeur de $\pi(y)$, la densité cible.

En détail, l'algorithme fonctionne comme suit :

- À l'étape j, on part de $x_j = x$ et on propose un nouveau point y en tirant selon la loi $K(x, \cdot)$.
- On calcule la probabilité d'acceptation donnée par :

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)k(y, x)}{\pi(x)k(x, y)}\right),$$

  où $k(x, y)$ est la densité de transition de la loi K.
- On tire un nombre $t \sim \mathcal{U}(0, 1)$, et on définit la nouvelle position x_{j+1} par :

$$x_{j+1} = \begin{cases} y & \text{si } t \leq \alpha(x, y), \\ x & \text{sinon.} \end{cases}$$

Proposition 6 :

Le noyau de la chaîne de Markov associée à cet algorithme, noté $\tilde{K} : \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0, 1]$, admet une densité $\tilde{k}$ définie par :

$$\tilde{k}(x, y) = \alpha(x, y)k(x, y) + \left(1 - \int_{\mathbb{R}^n} \alpha(x, z)k(x, z)\mathrm{d}z\right) \mathbf{1}_{x=y}$$

Le noyau $\tilde{K}$ est Π-réversible, et si K est irréductible et apériodique, alors $\tilde{K}$ l'est également.

Preuve : On remarque que la probabilité que $X_{j+1} = x$ sachant que $X_j = x$ (donc qu'on ait un rejet) est de $1 - \alpha(x, Y)$ avec $Y \sim K(x, \cdot)$. Ainsi,

$$\mathbb{P}(X_{j+1} = x \mid X_j = x) = \int_{\mathbb{R}^n} (1 - \alpha(x, y))k(x, y)\mathrm{d}y = 1 - \int_{\mathbb{R}^n} \alpha(x, y)k(x, y)\mathrm{d}y$$

A l'inverse, si $B \in \mathcal{B}(\mathbb{R})$ tel que $x \notin B$, alors pour que X_{j+1} soit dans B, on a eu une acceptation de probabilité $\alpha(x, Y)$ où $Y \sim K(x, \cdot)$, d'où

$$\mathbb{P}(X_{j+1} \in B \mid X_j = x) = \int_B \alpha(x, y)k(x, y)\mathrm{d}y$$

ce qui conclut l'expresion de $\tilde{k}$. Montrons ensuite que la condition de balance détaillée est vérifiée. Si $x = y$, alors on a trivialement la relation vérifiée. Sinon, pour $y \neq x$,

$$\begin{aligned} \tilde{k}(x, y)\pi(x) &= \alpha(x, y)k(x, y)\pi(x) \\ &= \min(p(x)k(x, y), p(y)k(y, x)) \\ &= \alpha(y, x)k(y, x)\pi(y) \\ &= \tilde{k}(y, x)\pi(y) \end{aligned}$$

Enfin, pour tout $x \in \mathbb{R}^n$ et $B \in \mathcal{B}(\mathbb{R}^n)$ avec $\Pi(B) > 0$, il existe $j > 0$ tel que $K^j(x, B) > 0$, donc au vue de l'expression de $\tilde{K}$, $\tilde{K}^j(x, B) > 0$, donc $\tilde{K}$ est irréductible. Enfin, si $\tilde{K}(x, E_i) = 1$, tous les points de $y \in E_i$ vérifient $\alpha(x, y) = 1$. Donc sur un ensemble périodique $\{E_1, \cdots, E_j\}$, on a uniquement des acceptations donc $\tilde{K}$ suit la même loi que K, qui n'est pas périodique ce qui est contradictoire. ■

L'algorithme de Metropolis-Hastings permet d'approcher la distribution cible Π et de calculer des intégrales relatives à π. Un avantage majeur est que l'algorithme ne nécessite que le ratio $\pi(y)/\pi(x)$, permettant ainsi de fonctionner avec une version non normalisée de la distribution, μ.

En pratique, on utilise fréquemment un noyau de densité k symétrique, ce qui simplifie le taux d'acceptation α en éliminant l'influence de k. Il est courant de choisir un noyau de transition uniforme ou gaussien. Par exemple, avec un noyau gaussien de la forme

$$k(x, y) = \frac{1}{\gamma\sqrt{2\pi}} \exp\left(-\frac{1}{2\gamma^2}\|x - y\|^2\right)$$

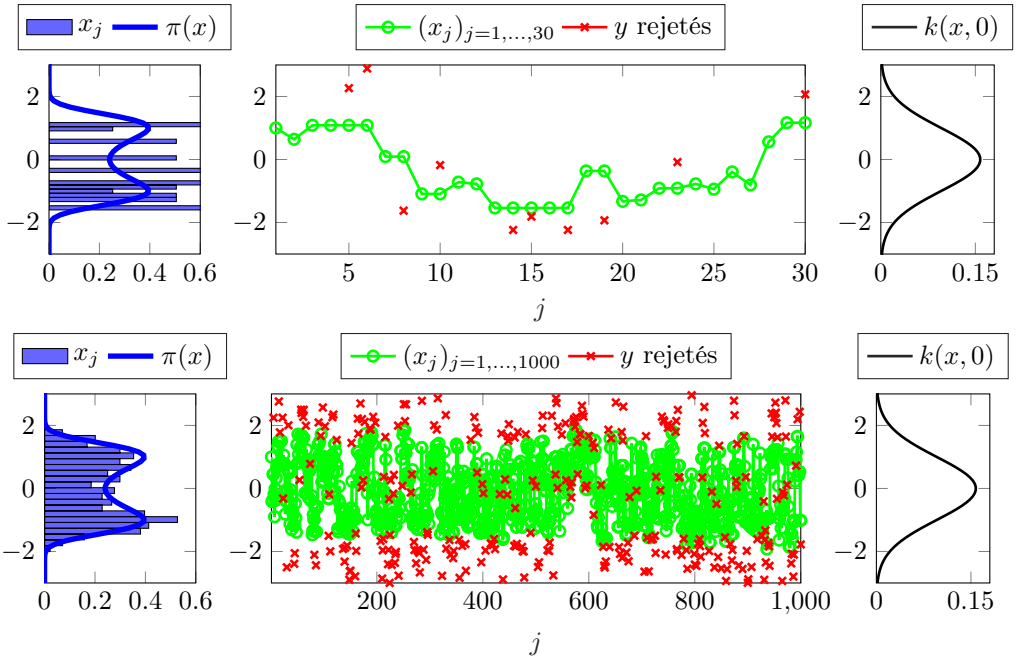

FIGURE 6: Deux réalisations de la chaîne de Markov $\{X_j\}_{j=1,\dots,N}$ pour $N = 30$ (en haut) et $N = 1000$ (en bas). À gauche, on représente la distribution cible π et l'histogramme des x_j. Au centre, on représente la chaîne de Markov et les points y rejetés au cours de l'algorithme. À droite, on représente le noyau k choisi, avec $\gamma = 1$.

et en cherchant à échantillonner une densité π telle que

$$\pi(x) \propto \mu(x) = \exp\left(-\frac{(1-x^2)^2}{2}\right)$$

on obtient plusieurs exemples de chaînes de Markov illustrée en Figure 6 en fonction de N. La Figure 7 montre également que le choix du noyau k (ou ici, du paramètre γ) est crucial pour que la chaîne de Markov converge rapidement vers la distribution cible.

Un exemple en dimension 2 est donné en Figure 8, avec une distribution cible définie par

$$\pi(x_1, x_2) \propto \mu(x_1, x_2) = \exp\left(-x_1^2 - x_2^2\right) + \frac{1}{2}\exp\left(-(x_1+5)^2 - (x_2-5)^2\right)$$

Cet exemple illustre l'efficacité du choix des x_j pour approcher l'intégrale de la distribution cible.

À partir des exemples présentés dans le TP, plusieurs bonnes pratiques peuvent être retenues pour optimiser l'algorithme :

- Pour bien ajuster le paramètre γ, il est empiriquement recommandé de viser un ratio d'acceptation d'environ 0,21, c'est-à-dire que le rapport entre les points acceptés et rejetés soit proche de cette valeur. Il est également possible de faire varier γ au cours des itérations, par exemple en fonction de l'indice j, afin d'implémenter une stratégie de recuit simulé.
- Pour éviter que la chaîne soit bloquée par un mauvais choix initial de x_1, il est conseillé de lancer plusieurs chaînes à partir de différentes valeurs initiales x_1, tirées aléatoirement.

En conclusion, on propose de résoudre enfin le problème posé en début de cours :

Exercice 8 :

Proposer une valeur de ω_0 et de τ pour dans le cadre de l'oscillateur harmonique amorti.

Solution : On utilise les priors mentionnés dans l'Exercice 7 et la vraissemblance obtenue dans l'Exercice 6 pour obtenir une expression de μ, proportionnelle à la distribution à posteriori cible. On fait ensuite tourner un algorithme de Metropolis-Hastings "bien" paramétré, pour calculer x_{CM}. On obtient avec $N = 10^5$ itérations que $x_{CM} \approx (1.00, 0.43, 1.44, 4.18)$. Ainsi, on propose $\tau = 1.44$ s et $\omega_0 = 4.18$ rad/s. Pour information, les données étaient générées avec $x = (1, 0.4, 1.5, 4.2)$.

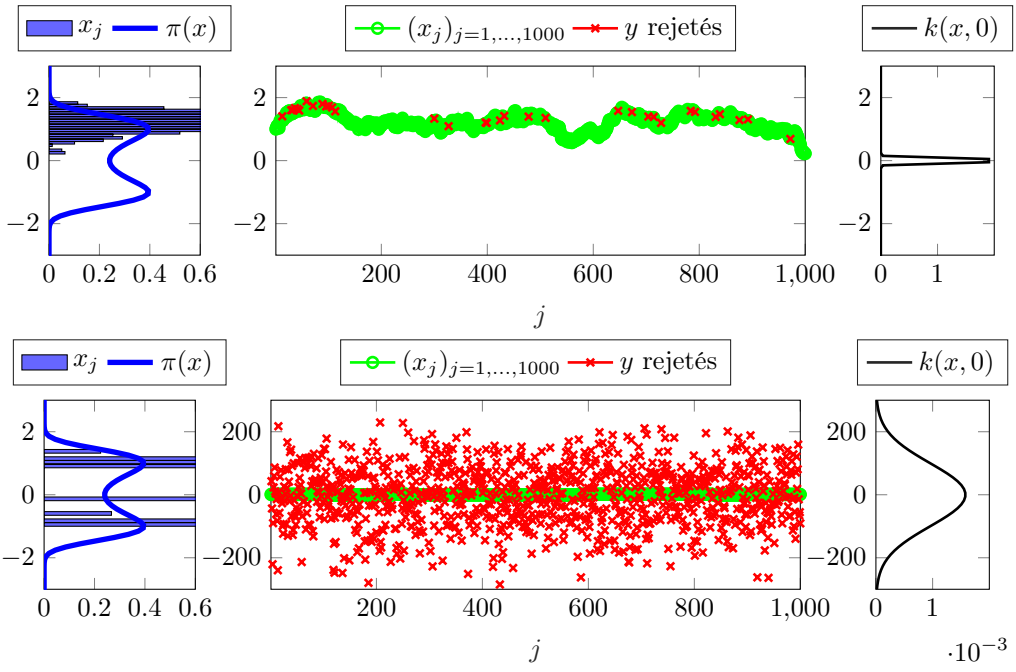

FIGURE 7: Deux réalisations de la chaîne de Markov $\{X_j\}_{j=1,\dots,N}$ pour $N = 1000$ et des noyaux avec $\gamma = 0.05$ (en haut) et $\gamma = 100$ (en bas). À gauche, on représente la distribution cible π et l'histogramme des x_j. Au centre, on représente la chaîne de Markov et les points y rejetés au cours de l'algorithme. À droite, on représente le noyau k choisi.

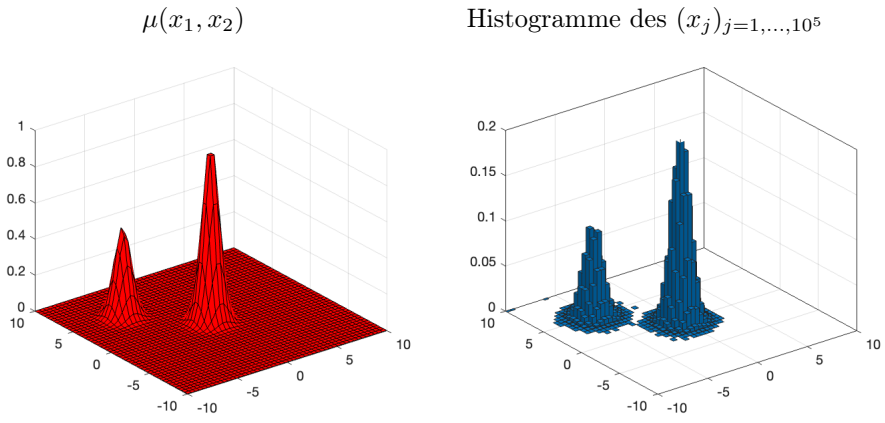

FIGURE 8: Distribution cible μ et histogramme des x_j, pour $N = 10^5$ et $\gamma = 2$.