MCMC project

A. Ben Friha, B. Cohen, V. Josse, L. Pascal

Décembre 2021

Rappels MCMC/ HMC

- **Méthodes de Monte Carlo** : On veut échantillonner à partir d'une densité de probabilité connue (à un facteur constant près par exemple). On se place dans le cadre où échantilonner de notre mesure de probabilité est impossible.
  - En remplaçant les échantillons réels par un ensemble de valeurs prises dans le support de nos valeurs d'intérêts (intégrale par exemple), et de la fonction à intégrer, nous pouvons obtenir une approximation de sa valeur.

Rappels MCMC/ HMC

- **Méthodes de Monte Carlo** : On veut échantillonner à partir d'une densité de probabilité connue (à un facteur constant près par exemple). On se place dans le cadre où échantilonner de notre mesure de probabilité est impossible.
  - En remplaçant les échantillons réels par un ensemble de valeurs prises dans le support de nos valeurs d'intérêts (intégrale par exemple), et de la fonction à intégrer, nous pouvons obtenir une approximation de sa valeur.
- **Caractère Markovien** : nos séquences forment des chaînes de Markov dont chaque nouveau terme ne dépend que de l'échantillon actuel.

Rappels MCMC/ HMC

- **Méthodes de Monte Carlo** : On veut échantillonner à partir d'une densité de probabilité connue (à un facteur constant près par exemple). On se place dans le cadre où échantilonner de notre mesure de probabilité est impossible.
  - En remplaçant les échantillons réels par un ensemble de valeurs prises dans le support de nos valeurs d'intérêts (intégrale par exemple), et de la fonction à intégrer, nous pouvons obtenir une approximation de sa valeur.
- **Caractère Markovien** : nos séquences forment des chaînes de Markov dont chaque nouveau terme ne dépend que de l'échantillon actuel.
- **Utilisations** :
  - Estimation d'intégrales par rapport à une distribution cible (calcul de moyenne/espérance empirique, ou de certains paramètres caractéristiques de distributions).
  - Calculs d'intégrales en dimensions plus grandes que 1 ou dont on ignore une primitive.

Exemple 1 : Détermination de la valeur de π.
On tire (X,Y) iid de lois uniformes sur [0,1] et on fait le rapport du nombre de points dans le quart de disque de rayon 1 et le nombre de tirages. On obtient alors une approximation de $\pi/4$ si le nombre de tirages est grand.

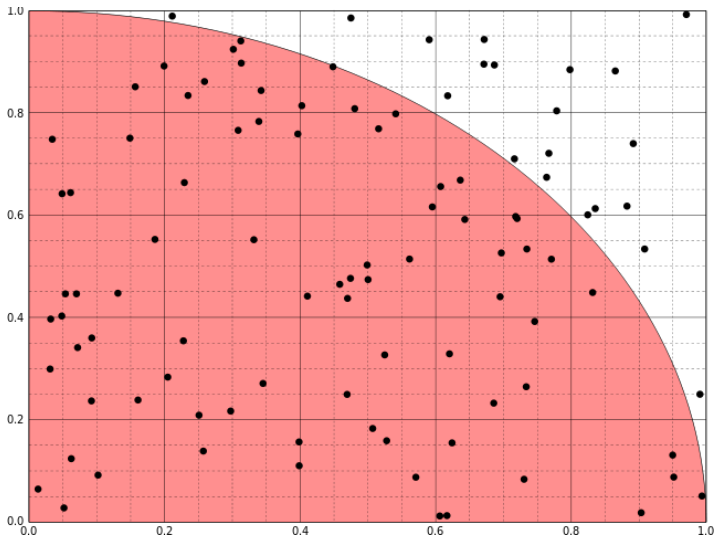

Figure: Monte Carlo et estimation de π

Exemple 2 : Jeu d'échecs.
L'approche consisterait ici à systématiquement explorer une branche de l'arbre jusqu'à sa position terminale. Etant donné le nombre grandiloquent de coups possibles, on optera pour un **Monte-Carlo Tree Search**. Cette méthode favorise les coups qui semblent meilleurs (plutôt que de choisir uniformément parmi tous les coups possibles) et augmente la probabilité de choisir un coup qui s'est avéré être bon dans les simulations précédentes.

Rappels MCMC/HMC : Metropolis-Hastings

Lorsqu'on utilise Monte Carlo pour échantillonner une densité de probabilité :

- Les méthodes de rejet souffrent du **fléau de la dimension** : la probabilité de rejet augmente exponentiellement avec le nombre de dimensions !
  - MCMC réduit l'impact de ce problème
- En grande dimension, difficile de trouver le pas de déplacement avec **Metropolis-Hastings**.
  - Chaque dimension peut avoir un comportement qui lui est propre, et il faut un pas adapté pour avoir le meilleur déplacement dans l'espace d'états car on choisit à chaque itération un nouvel échantillon dans cet espace grandissant exponentiellement avec la dimension.

Rappels MCMC/HMC : Monte Carlo Hamiltonien

- Par rapport à la distribution de proposition de marche aléatoire gaussienne utilisée dans l'algorithme de Metropolis-Hastings, la méthode hamiltonienne de Monte Carlo **réduit la corrélation entre les états d'échantillonnage consécutifs** en proposant une transition vers un état distant.
  - Réduire la corrélation entre les états signifie que **moins d'échantillons** de la chaîne de Markov sont nécessaires pour approximer l'intégrale de la distribution de probabilité cible pour une erreur de Monte Carlo donnée.
- Distribution cible : π sur $\mathbb{R}^d$, $\pi(q) \propto \mathrm{e}^{-U(q)}$.
- Distribution cible étendue : $\bar{\pi}$ sur $\mathbb{R}^d \times \mathbb{R}^d$,

$$\bar{\pi}(q,p) \propto \mathrm{e}^{-U(q)-p^T p/2}.$$

Rappel MCMC/HMC : Monte Carlo Hamiltonien

- En posant U **l'énergie potentielle**, K **l'énergie cinétique**, **l'hamiltonien** s'écrit :

$$H(p, q) = U(q) + K(p)$$

- La dynamique hamiltonienne se définit en supposant que q et p dépendent de t et que l'on souhaite annuler la dérivée de

$$H(p_t, q_t) = U(q_t) + K(p_t)$$

par rapport au temps.

- Après calculs on aboutit à :

$$dH/dq_{t,i}(p_t, q_t) = -dp_{t,i}/dt$$

et

$$dH/dp_{t,i}(p_t, q_t) = dq_{t,i}/dt$$

Cela va nous assurer des propriétés analogues pour un ensemble de même niveau (**level set**)

Intégrateur saute-mouton

Monte Carlo hamiltonien : instance de Metropolis-Hastings où les déplacements proposés dans l'espace d'états viennent d'un processus gouverné par une dynamique hamiltonienne et simulée à l'aide d'un integrateur numérique réversible et préservant le volume (généralement Leapfrog).

- **Méthode Leapfrog** pour $0 \leqslant \ell \leqslant L$:

$$p_{k+1}^{\ell+1/2} = p_{k+1}^{\ell} - \frac{h}{2}\nabla U(q_{k+1}^{\ell})$$

$$q_{k+1}^{\ell+1} = q_{k+1}^{\ell} + h p_{k+1}^{\ell+1/2}$$

$$p_{k+1}^{\ell+1} = p_{k+1}^{\ell+1/2} - \frac{h}{2}\nabla U(q_{k+1}^{\ell+1})$$

Intégrateur saute-mouton

Ajout d'une étape MH

- La **méthode Leapfrog** est **numérique** donc approximative et ne résout pas exactement les équations de la mécanique hamiltonienne, on ajoute une étape MH en complément avec un calcul de **taux d'acceptation** de transition pour améliorer les capacités d'exploration.
- Pour un **pas** égal à L, on itère le processus N fois en calculant à chaque itération L couples (q_i, p_i) Le **taux d'acceptation** de $q^{(k)}$ s'écrit alors pour chaque itération avec k variant entre 0 et N :

$$\alpha = \min(1, \exp(H(q_L, p_L) - H(q^{(k)}, p_0))$$

On tire ensuite U **uniforme** sur [0,1] que l'on compare à α pour accepter ou décliner un candidat.

Position du problème

Lorsqu'on utilise Monte Carlo pour échantillonner une densité de probabilité :

- Asympotiquement : convergence vers la bonne distribution.
- Mais : il faut beaucoup d'itérations pour avoir un résultat satisfaisant !
- *Exemple : distributions multi-modales (courantes en ML).*

On se propose d'utiliser des méthodes de MCMC pseudo-étendues pour pallier ce problème.

Position du problème

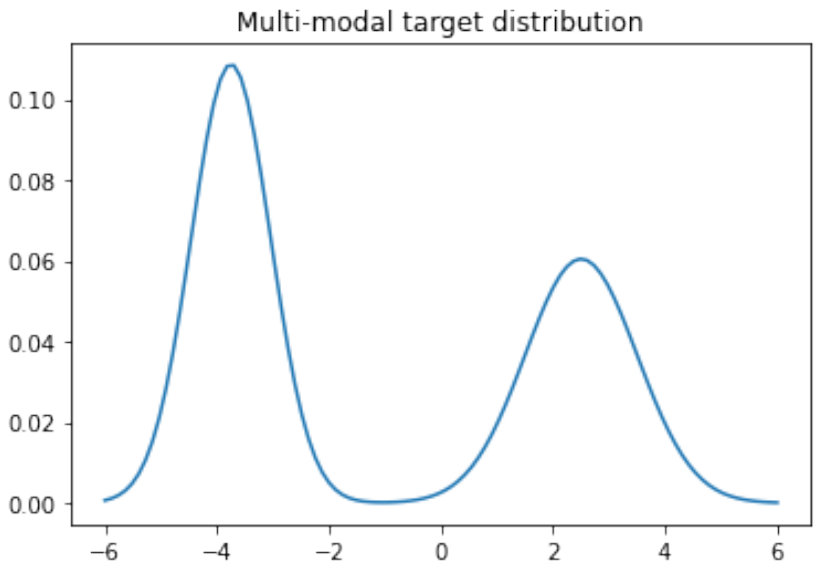

Figure: Exemple de distribution multi-modale simple (bi-gaussienne ici)

Position du problème

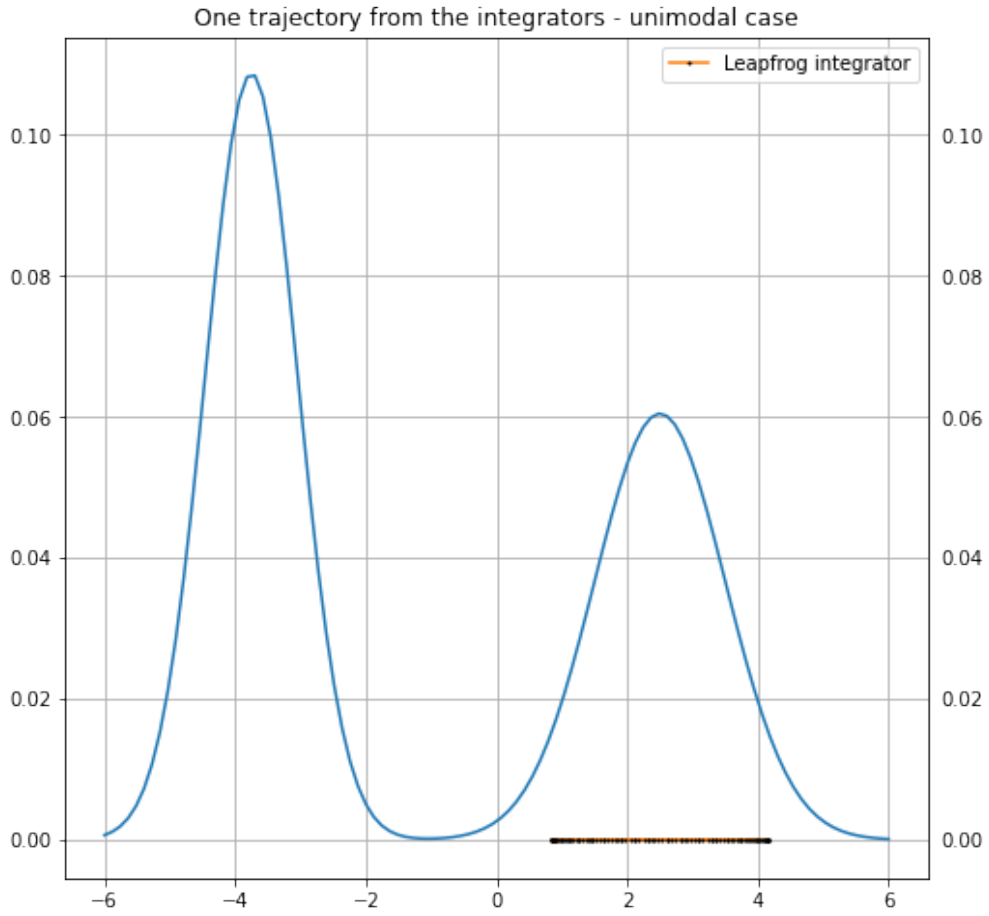

Figure: Avec les méthodes classiques, seul un mode est retrouvé

Pseudo-extended MCMC

Soit $\pi : \mathcal{X} \to \mathbb{R}^d$ une distribution de la forme :

$$\pi(\mathbf{x}) := \frac{\gamma(\mathbf{x})}{Z} = \frac{\exp(-\phi(\mathbf{x}))}{Z}$$

avec ϕ différentiable et Z constante.

Idée clef : réaliser à chaque étape N *pseudo-échantillons*, plutôt qu'un seul échantillon.

Pseudo-extended MCMC

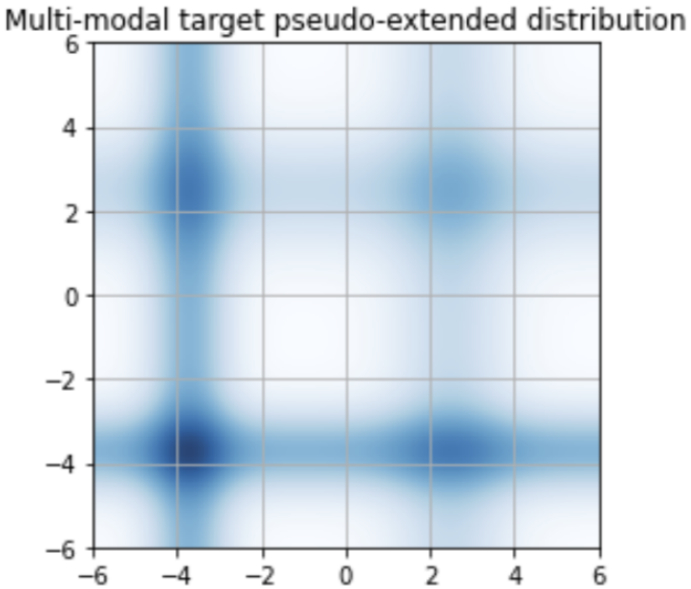

Figure: Exemple de distribution multi-modale simple étendue avec N = 2

Pseudo-extended MCMC

Principe :

- N tirages $\mathbf{x}_{1:N}$: variables auxiliaires
- Distribution instrumentale :

$$q(\mathbf{x}_i) \propto \exp(-\delta(\mathbf{x}_i)), \quad \delta \text{ une fonction.}$$

Pseudo-extended MCMC

Principe :

- N tirages $\mathbf{x}_{1:N}$: variables auxiliaires
- Distribution instrumentale :

$$q(\mathbf{x}_i) \propto \exp(-\delta(\mathbf{x}_i)), \quad \delta \text{ une fonction.}$$

- Distribution étendue :

$$\begin{aligned} \pi^N(\mathbf{x}_{1:N}) &:= \frac{1}{N} \sum_{i=1}^{N} \pi(\mathbf{x}_i) \prod_{j \neq i} q(\mathbf{x}_j) \\ &= \frac{1}{Z} \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{\gamma(\mathbf{x}_i)}{q(\mathbf{x}_i)} \right\} \prod_{i=1}^{N} q(\mathbf{x}_i) \end{aligned}$$

Pseudo-extended MCMC

En intégrant, les **distributions marginales** sont :

$$\pi^N(\mathbf{x}_i) = \frac{1}{N}\pi(\mathbf{x}_i) + \frac{N-1}{N}q(\mathbf{x}_i).$$

Comment retomber sur du $\pi(\mathbf{x})$ à partir de là ?

Pseudo-extended MCMC

En intégrant, les **distributions marginales** sont :

$$\pi^N(\mathbf{x}_i) = \frac{1}{N}\pi(\mathbf{x}_i) + \frac{N-1}{N}q(\mathbf{x}_i).$$

Comment retomber sur du $\pi(\mathbf{x})$ à partir de là ?

Processus de pondération pour trouver π. On a ce théorème :

Soient $\mathbf{x}_{1:N}$ distribuées selon $\pi^N(\mathbf{x}_{1:N})$. En utilisant des poids normalisés $\propto \frac{\gamma(\mathbf{x}_i)}{q((\mathbf{x}_i)}$, on a pour f intégrable :

$$\mathbb{E}_{\pi^N}\left[\frac{\sum_{i=1}^N \frac{f(\mathbf{x}_i)\gamma(\mathbf{x}_i)}{q((\mathbf{x}_i)}}{\sum_{i=1}^N \frac{\gamma(\mathbf{x}_i)}{q((\mathbf{x}_i)}}\right] = \mathbb{E}_\pi\left[f(\mathbf{x}_i)\right]$$

Pseudo-extended Hamiltonian Monte Carlo

- Pour échantillonner à partir de $\pi^N(\mathbf{x}_{1:N})$, on utilise un algorithme MCMC. Ici, on utilise HMC car le temps de calcul est avantageux.

Pseudo-extended Hamiltonian Monte Carlo

- Pour échantillonner à partir de $\pi^N(\mathbf{x}_{1:N})$, on utilise un algorithme MCMC. Ici, on utilise HMC car le temps de calcul est avantageux.
- **Problème :** méthode HMC basée sur une intégration par gradients
  - $\rightarrow$ simulation pouvant être piégée dans un mode local
    - ennuyeux quand on veut simuler à partir d'une distribution multi-modale.

Pseudo-extended Hamiltonian Monte Carlo

- Pour échantillonner à partir de $\pi^N(\mathbf{x}_{1:N})$, on utilise un algorithme MCMC. Ici, on utilise HMC car le temps de calcul est avantageux.
- **Problème :** méthode HMC basée sur une intégration par gradients
  - $\rightarrow$ simulation pouvant être piégée dans un mode local
    - ennuyeux quand on veut simuler à partir d'une distribution multi-modale.
- **Modèle de distribution étendue :** lie les modes entre eux dans un espace de dimension supérieure
  - $\rightarrow$ moins de risque de coincer dans un mode, car passage entre modes possible.

Modèle HMC

Paramètres $\mathbf{x} \in \mathbb{R}^d$ donnés.
Variables de moment : $\rho \in \mathbb{R}^d$, indépendantes de $\mathbf{x}$.

Hamiltonien $H(\mathbf{x}, \rho)$ **:** énergie totale du système

$$H(\mathbf{x}, \rho) = \phi(\mathbf{x}) + \frac{1}{2}\rho^\top M^{-1}\rho$$

- $\phi(\mathbf{x})$ énergie potentielle
- $\frac{1}{2}\rho^\top M^{-1}\rho$ énergie cinétique, avec M une matrice (souvent Id)

Modèle HMC

- Ensuite, échantillonner $(\mathbf{x}, \rho)$ de

$$\pi(\mathbf{x}, \rho) \propto \exp(H(\mathbf{x}, \rho)) = \pi(\mathbf{x})\mathcal{N}(\rho|0, M),$$

dont la loi marginale est exactement $\pi(\mathbf{x})$.

- Distribution étendue : hamiltonien de la forme :

$$H^N(\mathbf{x}_{1:N}, \rho) = -\log\left[\sum_{i=1}^{N} \exp(-\phi(\mathbf{x}_i) + \delta(\mathbf{x}_i))\right] + \sum_{i=1}^{N} \delta(\mathbf{x}_i)$$

$$+\frac{1}{2}\rho^\top M^{-1}\rho,$$

avec dorénavant $\rho \in \mathbb{R}^{d\times N}$, et $\delta(\mathbf{x})$ fonction potentielle arbitraire différentiable comme distribution instrumentale.

Simulation HMC

Généralement, on ne peut pas simuler du système Hamiltonien tel quel, on discrétise le temps en petits pas ϵ.

Liens avec pseudo-marginal MCMC

- Distribution pseudo-extended : peut être vu comme cas particulier de distribution pseudo-marginale
  - simulation réalisée avec un estimateur $\tilde{\pi}$ de π (et non π directement) dans l'algorithme MCMC choisi.

Liens avec pseudo-marginal MCMC

- Distribution pseudo-extended : peut être vu comme cas particulier de distribution pseudo-marginale
  - simulation réalisée avec un estimateur $\tilde{\pi}$ de π (et non π directement) dans l'algorithme MCMC choisi.

- Dans le cadre pseudo-marginal, on suppose π de la forme

$$\pi(\theta) = \int_{\mathcal{X}} \pi(\theta, x)\,\mathrm{d}x,$$

et ne pouvant pas être intégrée analytiquement.
On va estimer π.

Liens avec pseudo-marginal MCMC

- Estimateur non-biaisé $\tilde{\pi}(\theta)$ calculé par importance sampling, à partir de $\mathbf{x}_{1:N}$ tirés selon $q(\mathbf{x})$ pour estimer l'intégrale :

$$\tilde{\pi}(\theta) := \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(\theta, \mathbf{x}_i)}{q(\mathbf{x}_i)}$$

- La distribution pseudo-marginale cible est alors définie comme

$$\tilde{\pi}(\theta, \mathbf{x}) := \frac{1}{N} \sum_{i=1}^{N} \pi(\theta, \mathbf{x}_i) \prod_{j \neq i} q(\mathbf{x}_j)$$

qui admet $\pi(\theta)$ comme marginale.

Liens avec pseudo-marginal MCMC

Différences entre pseudo-marginal MCMC et pseudo-extended MCMC :

- la distinction entre variables latentes et paramètres
- pas d'importance sampling dans pseudo-extended : on simule directement à partir de la cible pseudo-extended avec un HMC.

Liens avec pseudo-marginal MCMC

Différences entre pseudo-marginal MCMC et pseudo-extended MCMC :

- la distinction entre variables latentes et paramètres
- pas d'importance sampling dans pseudo-extended : on simule directement à partir de la cible pseudo-extended avec un HMC.

$\rightarrow$ **L'avantage de pseudo-extended est donc qu'elle ne requiert pas de savoir échantillonner à partir de q.**

Liens avec pseudo-marginal MCMC

- Les deux méthodes visent deux problèmes différents :
  - le **cadre pseudo-marginal** permet d'échantillonner à partir de modèles dont on ne peut pas calculer la vraisemblance,
  - tandis que l'utilisation de la **méthode pseudo-extended** est adaptée pour échantillonner à partir de distributions complexes.
- La distribution instrumentale doit être choisie selon les propriétés désirées : dans le cas d'importance sampling, on a vu qu'on a besoin de pouvoir échantillonner à partir de q.
- **Souhait** : trouver une distribution instrumentale
  - calculable explicitement,
  - couvrant π,
  - reflétant sa multi-modalité.

Tempering targets avec les distributions intrumentales

- Avantage du cadre pseudo-extended : pas besoin de pouvoir échantillonner à partir de q.
  - Pouvoir l'évaluer ponctuellement à une constante multiplicative près (celle-ci disparaît lors du passage au Hamiltonien) suffit.

  $\rightarrow$ La contrainte est donc moindre sur le choix de la distribution et permet de mieux approcher la distribution cible en autorisant les modes multiples dans la distribution instrumentale.
- Pour améliorer l'échantillonnage avec une distribution instrumentale : méthode de tempering.
  Soit la famille des distributions approximées :

$$\Pi := \left\{ \pi_\beta(\mathbf{x}) = \frac{\gamma_\beta(\mathbf{x})}{Z(\beta)} \right\}$$

où $\gamma_\beta(\mathbf{x}) = \exp(-\beta\phi(\mathbf{x}))$ peut être évalué point par point, mais pas $Z(\beta)$.

Tempering targets avec les distributions intrumentales

- On construit la distribution cible étendue $\pi^N(\mathbf{x}_{1:N}, \beta_{1:N})$ sur $\mathcal{X}^N \times (0,1]^N$ avec N paires $(\mathbf{x}_i, \beta_i)$ pour $i = 1, \ldots, N$
  Nous allons construire la distribution cible de manière à ce que la distribution marginale des x_i soit une mixture d'éléments de Π, de manière à ce que la distribution marginale soit plus diffuse que π pour permettre la multi-modalité.
- Posons $q(\mathbf{x}, \beta) = \pi_\beta(\mathbf{x})q(\beta)$ avec $q(\beta) = \frac{Z(\beta)g(\beta)}{C}$, où g est une fonction qui peut être évaluée point par point, C une constante de normalisation. Alors les constantes de normalisation $Z(\beta)$, intractables, disparaissent dans la formule de $q(\mathbf{x}, \beta)$:

$$q(\mathbf{x}, \beta) = \frac{\gamma_\beta(\mathbf{x})g(\beta)}{C}$$

  La distribution instrumentale $q(\mathbf{x}, \beta)$ n'est pas analytique mais elle peut être évaluée point par point à une constante multiplicative près, ce qui suffit dans le cadre pseudo-extended.

Tempering targets avec les distributions intrumentales

- On définit la distribution cible pseudo-extended comme suit :

$$\pi^N(\mathbf{x}_{1:N}, \beta_{1:N}) = \frac{1}{N}\sum_{i=1}^{N}\pi(\mathbf{x}_i)\pi(\beta_i)\prod_{j\neq i} q(\mathbf{x}_j, \beta_j) \quad (1)$$

$$= \frac{1}{ZC^{N-1}}\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\gamma(\mathbf{x}_i)\pi(\beta_i)}{\gamma_{\beta_i}(\mathbf{x}_i)g(\beta_i)}\right\}\prod_{j=1}^{N}\gamma_{\beta_j}(\mathbf{x}_j)g(\beta_j) \quad (2)$$

où $\pi(\beta)$ est une distribution arbitraire que l'on choisit pour β. Les constantes Z et C sont indépendantes des variables et donc s'annulent lors de l'application de Metropolis-Hastings.

Intuition derrière *tempered MCMC*

- L'idée de *tempered MCMC* est d'échantillonner à partir d'une séquence de cibles $\pi_k(\mathbf{x}) \propto \exp(-\beta_k \phi(\mathbf{x}))$, $k = 1, \ldots, K$
- Quand β_k est petit, les différents modes de π sont "aplatis", de telle sorte que l'échantillonneur MCMC peut traverser plus facilement les régions de faible densité qui séparent les modes.

Simulation

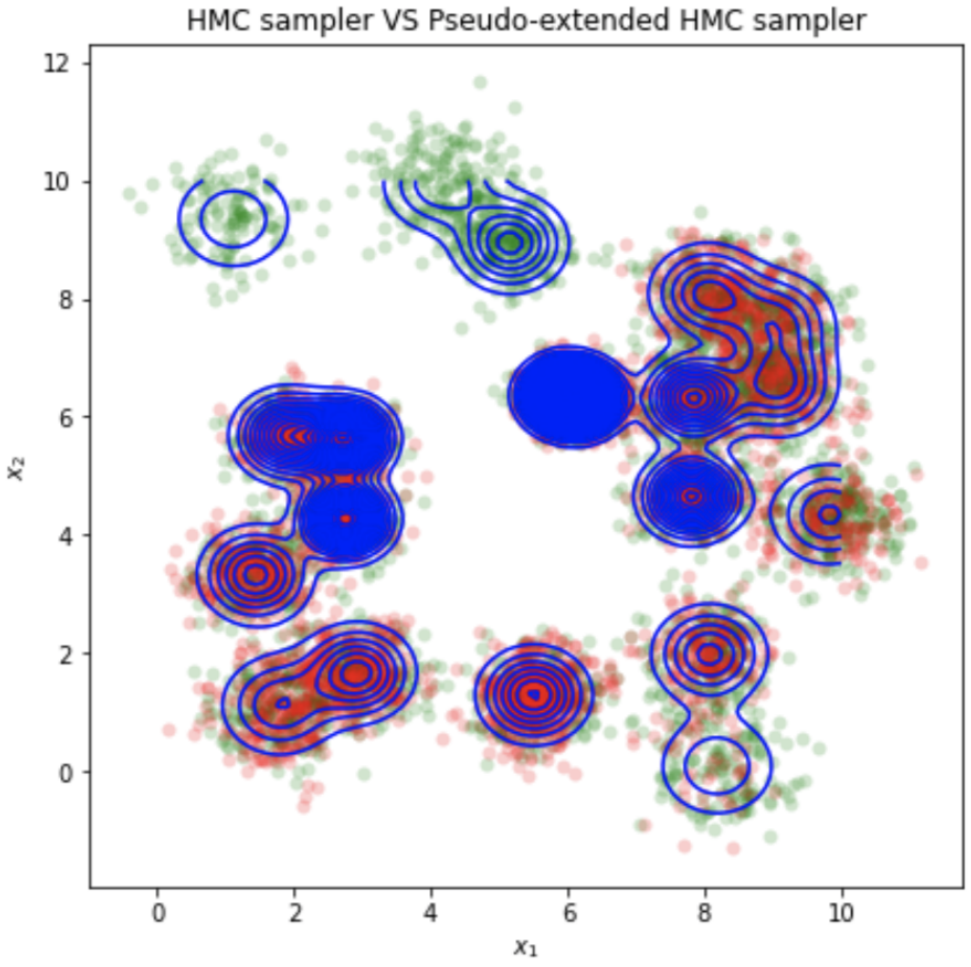

Figure: HMC vs. Pseudo-extended HMC

Conclusion

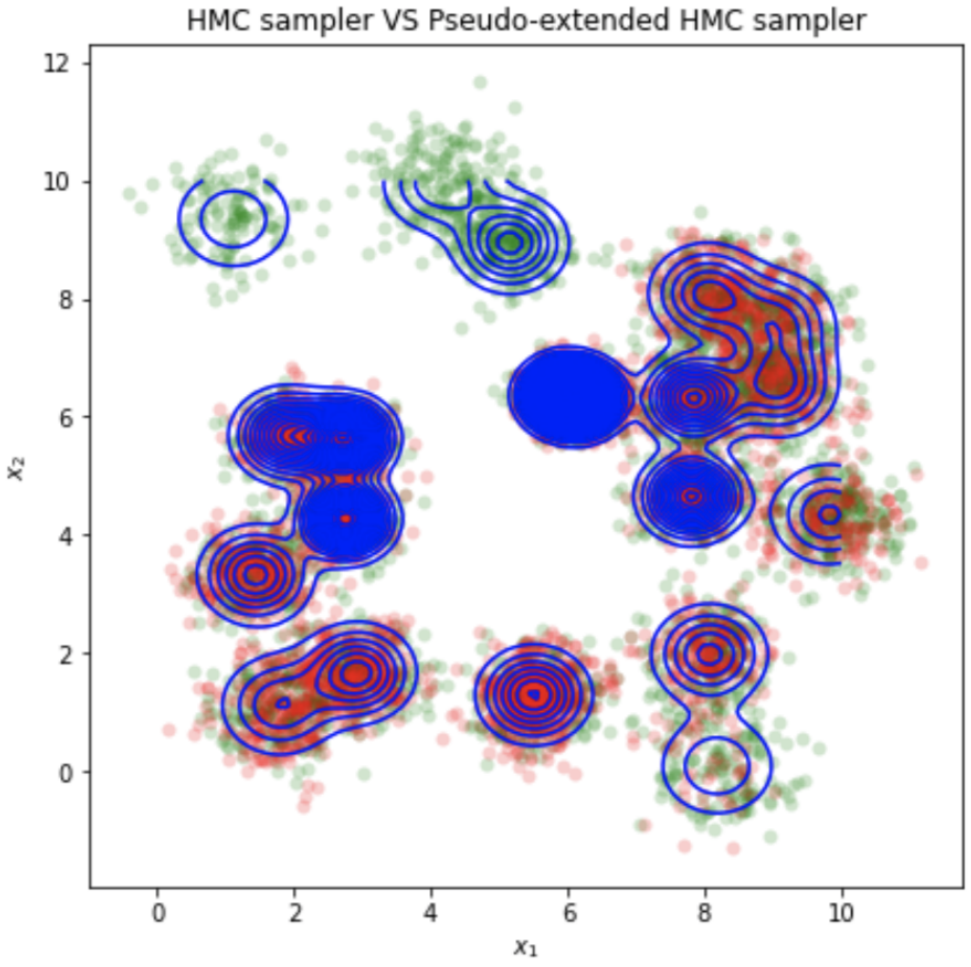

Figure: HMC vs. Pseudo-extended HMC