M2 ISN – Chaînes de Markov et modélisation
Responsable : Adrien Hardy, email : `adrien.hardy@univ-lille.fr`

TP4

1 Monte Carlo : un exemple continu 1D

On considère la densité de probabilité sur $\mathbb{R}$ donnée par

$$f(x) \propto (1 + \sin(3x))\, \mathrm{e}^{-x^2/2}$$

à une constante de normalisation près.

1. Quelle est la valeur de la constante de normalisation ? Faire un graphe de f.
2. Simuler plusieurs réalisations de densité f à l'aide de la méthode des rejets et les présenter sous forme d'histogramme, sur lequel on rajoutera le graphe de f.
3. Simuler plusieurs réalisations de densité f à l'aide de l'algorithme de Metropolis-Hastings et faire de même que dans la question précédente.
4. Calculer la valeur théorique de $\int x^2 f(x)\, \mathrm{d}x$, puis donner une valeur approchée à l'aide des méthodes de Monte Carlo (rejet et MH) et les comparer.

2 Régression logistique (approche bayésienne)

On considère le modèle de régression logistique $\mathbb{P}(Y = 1|X) = \sigma(X, \beta)$ avec $X \in \mathbb{R}^d$ et où on a choisi la sigmoïde "logit"

$$\sigma(X, \beta) = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 X^{(1)} + \ldots + \beta_d X^{(d)})}}.$$

On rappelle que si les paramètres $\beta = (\beta_0, \ldots, \beta_d)$ ont pour loi *a priori* $\mathcal{N}(0, I_{d+1})$, la densité *a posteriori* de β sachant $X = [X_1, \ldots, X_N] \in (\mathbb{R}^d)^N$ et $Y \in \{0, 1\}^N$ est donnée à une constante multiplicative près par

$$p(\beta|X, Y) \propto \prod_{i=1}^{N} \sigma(X_i, \beta)^{y_i} (1 - \sigma(X_i, \beta))^{1-y_i}\, \mathrm{e}^{-\|\beta\|^2/2}.$$

1. Construire un jeu de données factice : simuler N vecteurs $X_1, \ldots, X_N$ de $\mathbb{R}^d$ et $Y = (y_1, \ldots, y_d) \in \{0, 1\}^d$ de la façon de votre choix.
2. Simuler à l'aide de Metropolis-Hastings une variable de densité $p(\beta|X, Y)$; on fera attention au nombre de pas acceptés dans l'algorithme, qu'on essaiera de maintenir aux environs de 25% en ajustant la variance de la loi des variables proposées.
3. Donner une approximation de l'estimateur

   $$\hat{\theta} = \int \theta\, p(\beta|X, Y) \mathrm{d}\theta.$$

   En déduire une prédiction du taux de succès de Y^* pour les données d'un nouvel individu $X^* \in \mathbb{R}^d$.