SIMULATION-BASED LEARNING: THEORY AND APPLICATIONS
M2 MVA 2014-2015
Homework 3

Part A is due by 14:00 on **March 3rd 2015**
and Part B is due by 14:00 on **March 10th 2015**
The codes should be submitted by email to `gersende.fort@telecom-paristech.fr`.
The report has to be either submitted by email in pdf format (same address) or given to one of the instructor.
They can be done in groups of two students, in which case we ask that both students submit the homework.
The writeup may be in french or in english.
Please, name the file as follows: `MVA_HW3_⟨your_name⟩.pdf` if you worked alone and `MVA_HW3_⟨your_name1⟩_⟨your_name2⟩.pdf` if you worked in a group of two.

1 Part A

1.1 Exercise 1

Choose a target distribution π on the integers $\mathsf{X} = \{1, 2, \cdots, 6\}$ such that $\pi(x) > 0$ for all $x \in \mathsf{X}$.

1. Let $\gamma > 0$. Write a Hastings-Metropolis algorithm with proposal distribution:

   given the current point X_n, $Y_{n+1} \sim \mathcal{U}\left(\{X_n - \gamma, \cdots, X_n - 1, X_n + 1, \cdots, X_n + \gamma\}\right)$.

2. Does this algorithm satisfy the uniform Doeblin condition ? If such, provide an upper bound on the rate of convergence in total variation norm of the distribution of X_n to π.

3. Run the above algorithm for different values of γ (for example: $\gamma = 1$, $\gamma = 4$ and $\gamma = 50$) and illustrate the convergence of the chain $\{X_n, n \geq 0\}$ to the target distribution π.
   - Do you observe any influence of γ on the rate of convergence of the algorithm ?
   - Comment the results.

1.2 Exercise 2

Let a family of transition matrix $\{P_t, t \in]0, 1[\}$ on $\{0, 1\}$ defined as follows

$$P_t = \begin{bmatrix} t & 1-t \\ 1-t & t \end{bmatrix}$$

1. For $t \in]0, 1[$, what is the invariant distribution π_t of the transition matrix P_t ? Show that the distribution of a Markov chain with transition matrix P_t converges to π_t in total variation norm, whatever the initial value of the chain.

2. Fix $t_0, t_1 \in]0, 1[$. Let $\{X_n, n \geq 0\}$ be a chain defined as follows: $X_0 = x \in \{0, 1\}$; given X_n,

if $X_n = 0$, then $X_{n+1} \sim P_{t_0}(X_n, \cdot)$

otherwise, $X_{n+1} \sim P_{t_1}(X_n, \cdot)$.

Show that $\{X_n, n \geq 0\}$ is an (homogeneous) Markov chain; write its transition matrix $P_\star$, compute its invariant distribution $\pi_\star$ and show that the chain is uniformly ergodic.

1.3 Exercise 3

The target distribution is a centered multivariate Gaussian distribution with dimension $d = 200$. Its covariance matrix Γ_π is diagonal with eigenvalues regularly spaced in the interval $[10^{-2}, 10^3]$.

1. Run a symmetric random walk Hastings-Metropolis algorithm with proposal distribution

$$\text{given } X_n,\ Y_{n+1} \sim \mathcal{N}_d\left(X_n, cI\right)$$

for some constant $c > 0$; I denotes the $d \times d$ identity matrix. For different values of c,

- display the trace plot of a path of the first component of the chain.
- plot the evolution of the mean acceptance rate $\hat{\alpha}_n$ along the path as a function of the number of iterations. $\hat{\alpha}_n$ is defined by

$$\hat{\alpha}_n := \frac{1}{n - n_0} \sum_{k=n_0+1}^{n} \mathbb{1}_{X_k = Y_k}\,, \qquad n \geq n_0\,,$$

where n_0 is the burn-in time.

2. Run an adaptive symmetric random walk Hastings-Metropolis algorithm defined as follows: at iteration n, given the current sample X_n and a current estimate of the covariance matrix Γ_n,

- do an iteration of a symmetric random walk Hastings-Metropolis algorithm with proposal distribution $\mathcal{N}_d\left(X_n, (2.38)^2\Gamma_n/d\right)$. Obtain X_{n+1}.

- update the estimation of the mean

$$\mu_{n+1} = \frac{1}{n+1-n_0} \sum_{k=n-n_0+1}^{n+1} X_k \left(= \mu_n + \frac{1}{n+1-n_0}(X_{n+1} - \mu_n) \text{ for } n \geq n_0 \right).$$

- update the estimation of the covariance matrix, given for $n \geq n_0$ by[1]

$$\Gamma_{n+1} = \Gamma_n + \frac{1}{n+1-n_0}\left(\left(X_{n+1} - \mu_{n+1}\right)\left(X_{n+1} - \mu_{n+1}\right)' - \Gamma_n\right).$$

It is advocated to choose $n_0 > 0$ and start the adaptation at iteration n_0 (when $n \leq n_0$, run a classical Hastings-Metropolis algorithm). In case n_0 is not large enough, numerical problems may occur when sampling a Gaussian distribution with covariance $\propto \Gamma_n$; if such, it is advocated to modify the proposal distribution in order to propose from the mixture

$$(1 - 0.05)\ \mathcal{N}_d\left(X_n, \frac{(2.38)^2}{d}\Gamma_n\right) + 0.05\ \mathcal{N}_d\left(X_n, \frac{0.1}{d} I\right)$$

- display the trace plot of a path of the first component of the chain.
- plot the evolution of the mean acceptance rate $\hat{\alpha}_n$ along the path as a function of the number of iterations.

[1] by convention, vectors are in column and x' denotes the transpose of the matrix x

- plot the evolution of the *suboptimality factor* defined by

$$n \mapsto d \frac{\sum_{i=1}^{d} \lambda_{i,n}^{-2}}{\left(\sum_{i=1}^{d} \lambda_{i,n}^{-1}\right)^2}$$

  where $\lambda_{1,n}, \cdots, \lambda_{d,n}$ are the eigenvalues of $\Gamma_n^{1/2}\Gamma_\pi^{-1/2}$.

3. Comment the results.

On the suboptimality factor. Consider a multi-dimensional random-walk Hastings-Metropolis algorithm with proposal covariance matrix $(2.38)^2/d\,\Gamma_p$ acting on a normal target distribution with covariance matrix Γ_π. Theorem 5 of Roberts & Rosenthal (2001) [2] proves that it is optimal to take $\Gamma_p = \Gamma_\pi$; for other choice, the mixing rate will be slower by a suboptimality factor of

$$d \frac{\sum_{i=1}^{d} \lambda_i^{-2}}{\left(\sum_{i=1}^{d} \lambda_i^{-1}\right)^2}$$

where $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of the matrix $\Gamma_p^{1/2}\Gamma_\pi^{-1/2}$.

[2] G.O. Roberts and J.S. Rosenthal (2001), Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. 16 , 351–367

2 Part B (Problem 1 of DM1, 2nd visit)

2.1 Step 1.

Let $\Theta \subseteq \mathbb{R}^d$. Consider the optimization problem $\quad \mathrm{argmin}_{\theta\in\Theta}\left(f(\theta)+g(\theta)\right) \quad$ where $f:\Theta\to\mathbb{R}$ is a continuously differentiable function with Lipschitz gradient:

$$\exists L>0,\ \forall \theta,\theta'\in\Theta \qquad \|\nabla f(\theta)-\nabla f(\theta')\| \leq L\|\theta-\theta'\|\,; \tag{1}$$

and $g:\Theta\to]0,+\infty]$ is a convex function, not identically equal to $+\infty$ and lower semi-continuous.

For $\gamma>0$, define

$$Q_\gamma(\vartheta;\theta) := f(\theta)+\langle \nabla f(\theta), \vartheta-\theta\rangle + \frac{1}{2\gamma}\|\vartheta-\theta\|^2 + g(\vartheta)$$

1. Show that for any $0<\gamma\leq 1/L$ and any $\theta,\vartheta\in\Theta$, $f(\vartheta)+g(\vartheta)\leq Q_\gamma(\vartheta;\theta)$ and $f(\theta)+g(\theta) = Q_\gamma(\theta;\theta)$.

2. For a sequence $\{\gamma_n; n\geq 0\}$ such that $\gamma_n\in]0,1/L]$, define the sequence $\{\theta_n, n\geq 0\}$ by induction: $\theta_0\in\Theta$; given θ_n,
$$\theta_{n+1} = \mathrm{argmin}_{\vartheta\in\Theta} Q_{\gamma_{n+1}}(\vartheta;\theta_n)\,. \tag{2}$$
Show that $(f+g)(\theta_{n+1})\leq (f+g)(\theta_n)$.

3. Show that it also holds
$$Q_\gamma(\vartheta;\theta) = f(\theta)+\frac{1}{2\gamma}\,\|\vartheta-(\theta-\gamma\nabla f(\theta))\|^2 - \frac{\gamma}{2}\|\nabla f(\theta)\|^2 + g(\vartheta).$$

4. In the case $g(\theta)=\lambda\sum_{i=1}^n|\theta_i|$ for some $\lambda>0$, show that (2) gets into
$$\theta_{n+1} = P_{\gamma_{n+1}}(\theta_n-\gamma_{n+1}\nabla f(\theta_n))$$
with for any $1\leq i\leq d$,
$$\left(P_\gamma(u)\right)_i = \begin{cases} u_i-\gamma\lambda & \text{if } u_i\geq \gamma\lambda\,,\\ u_i+\gamma\lambda & \text{if } u_i\leq -\gamma\lambda\,,\\ 0 & \text{if } u_i\in(-\gamma\lambda,\gamma\lambda)\,.\end{cases}$$

2.2 Step 2.

We model binary responses $Y_i\in\{0,1\}$ for $i=1,\cdots,N$ as N conditionally independent realizations of a random effect logistic regression model, [3]

$$Y_i|\mathbf{U} \overset{ind.}{\sim} \mathsf{Ber}\left(s(x_i'\beta+\sigma z_i'\mathbf{U})\right),\ \ 1\leq i\leq N\,, \tag{3}$$

where $x_i\in\mathbb{R}^p$ is the vector of (known) covariates, $z_i\in\mathbb{R}^q$ are (known) loading vector, $\mathsf{Ber}(\alpha)$ denotes the Bernoulli distribution with parameter $\alpha\in(0,1)$, $s(x)=\exp(x)/(1+\exp(x))$ is the cumulative distribution function of the standard logistic distribution. The random effect $\mathbf{U}$ is assumed to be standard Gaussian $\mathbf{U}\sim\mathcal{N}_q(0,I)$. Set $\theta=(\beta,\sigma)\in\Theta:=\mathbb{R}^p\times(0,\infty)$.

1. Give the expression of the log-likelihood of the observations $(Y_1,\cdots,Y_N)$, $\theta\mapsto\ell(\theta)$ - the dependance upon the observations is omitted in the notation.

[3] By convention, the vectors are column-vectors and x' denotes the transpose of the matrix x

2. Show that the gradient of the log-likelihood is given by

$$\nabla \ell(\theta) = \int \left\{ \sum_{i=1}^{N} (Y_i - s(x_i'\beta + \sigma z_i'\mathbf{u})) \begin{bmatrix} x_i \\ z_i'\mathbf{u} \end{bmatrix} \right\} \pi_\theta(\mathbf{u}) \, \mathrm{d}\mathbf{u} \,, \tag{4}$$

   where $\pi_\theta(\mathbf{u}) := \exp\left(\ell_c(\theta|\mathbf{u}) - \ell(\theta)\right) \phi(\mathbf{u})$, ϕ is the density of a standard $\mathbb{R}^q$-Gaussian distribution and

$$\ell_c(\theta|\mathbf{u}) = \sum_{i=1}^{N} \left\{ Y_i \left(x_i'\beta + \sigma z_i'\mathbf{u} \right) - \ln\left(1 + \exp\left(x_i'\beta + \sigma z_i'\mathbf{u}\right)\right) \right\} \,.$$

We would like to compute the maximum likelihood estimator under a constraint of sparsity on the vector β and solve $\operatorname{argmin}_{\theta \in \Theta} \left(-\ell(\theta) + \lambda \sum_{i=1}^{d} |\beta_i| \right)$ for some $\lambda > 0$. To that goal, a solution consists in applying the algorithm described in Section 2.1. Unfortunately, the gradient $\nabla f(\theta_n)$ is untractable and we propose to substitute this quantity by a Monte Carlo approximation.

2.3 Step 3

From (4), we have $\nabla \ell(\theta) = \int H_\theta(\mathbf{u}) \, \pi_\theta(\mathbf{u}) \, \mathrm{d}\mathbf{u}$ where $\pi_\theta(\mathbf{u}) \, \mathrm{d}\mathbf{u}$ is a probability distribution. In this section, we describe a Gibbs sampler with invariant distribution $\pi_\theta(\mathbf{u}) \, \mathrm{d}\mathbf{u}$. For $c \in \mathbb{R}$, define the density $w \mapsto \overline{\pi}(w; c)$ on $\mathbb{R}^+$ by

$$\overline{\pi}(w; c) := Z \, \cosh(c/2) \, \exp\left(-wc^2/2\right) \, \rho(w) \, \mathbb{1}_{\mathbb{R}^+}(w) \,, \qquad \rho(w) := w^{-3/2} \sum_{k \geq 0} (-1)^k (2k{+}1) \exp(-(2k{+}1)^2/(8w))$$

for a constant Z which does not depend on c. For $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{w} = (w_1, \cdots, w_N) \in \mathbb{R}^N$, set

$$\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) := \left(\prod_{i=1}^{N} \overline{\pi}\left(w_i; x_i'\beta + \sigma z_i'\mathbf{u}\right) \right) \pi_\theta(\mathbf{u}) \,.$$

Note that we have $\nabla \ell(\theta) = \int H_\theta(\mathbf{u}) \, \pi_\theta(\mathbf{u}) \, \mathrm{d}\mathbf{u} = \int \int H_\theta(\mathbf{u}) \, \tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) \, \mathrm{d}\mathbf{u} \, \mathrm{d}\mathbf{w}$.

1. Show that

$$\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) = C_\theta \, \phi(\mathbf{u}) \prod_{i=1}^{N} \exp\left(\sigma(Y_i - 1/2) z_i'\mathbf{u} - w_i (x_i'\beta + \sigma z_i'\mathbf{u})^2/2\right) \, \rho(w_i) \mathbb{1}_{\mathbb{R}^+}(w_i) \,,$$

   and give the expression of the constant C_θ as a function of Z, N, θ, Y_i and x_i.

2. Show that the conditional distribution of $\mathbf{u}$ given $\mathbf{w}$ associated to $\tilde{\pi}_\theta$ is Gaussian distribution with mean $\mu_\theta(\mathbf{w})$ and covariance matrix $\Gamma_\theta(\mathbf{w})$ given by

$$\Gamma_\theta(\mathbf{w}) = \left(I + \sigma^2 \sum_{i=1}^{N} w_i z_i z_i' \right)^{-1} \,, \qquad \mu_\theta(\mathbf{w}) = \sigma \Gamma_\theta(\mathbf{w}) \sum_{i=1}^{N} \left((Y_i - 1/2) - w_i x_i'\beta \right) z_i \,.$$

3. Show that the conditional distribution of $\mathbf{w}$ given $\mathbf{u}$ associated to $\tilde{\pi}_\theta$ is $\prod_{i=1}^{N} \overline{\pi}(w_i; |x_i'\beta + \sigma z_i'\mathbf{u}|)$.

4. How to sample from $\overline{\pi}$: it can be shown [4] that if W is returned by `Homework1`[5] run with $z \leftarrow c/2$ then $W/4 \sim \overline{\pi}(\cdot; c)$.

[4] it is not required to prove it

[5] see the code written for the first Homework

5. The functions `Homework1` and `randn`[6] are available. Write the pseudo-code of an algorithm `GibbsHomework3` to sample a Markov chain of length $N_{\max}$ with invariant distribution $\tilde{\pi}_\theta$, which uses calls to `Homework1` and `randn`.

6. Write the pseudo-code of an algorithm `GradSto` with input: $N_{\max}$ and θ; and output: a Monte Carlo approximation of $\nabla \ell(\theta)$ computed from a chain of length $N_{\max}$.

2.4 Step 4

The goal of this step is to run the following stochastic optimization algorithm:

$$\theta_{n+1} = P_{\gamma_{n+1}}(\theta_n + \gamma_{n+1} H_{n+1}) \tag{5}$$

where H_{n+1} is a Monte Carlo approximation of $\nabla \ell(\theta_n)$ computed from a chain of length m_{n+1} obtained by a call to `GradSto`. Roughly speaking, it is expected that $\{\theta_n, n \geq 0\}$ converges almost-surely to a solution of

$$\mathrm{argmin}_{\theta \in \Theta} \left(-\ell(\theta) + \lambda \sum_{i=1}^{p} |\beta_i| \right) .$$

The method is illustrated on a simulated data set.

1. Obtain the data set: Choose $N = 500$, $p = 1\,000$ and $q = 5$.
   Generate the $N \times p$ covariates matrix $X = [x_1; \cdots ; x_p]$ columnwise, by sampling a stationary $\mathbb{R}^N$-valued autoregressive model with parameter $\rho = 0.8$ and Gaussian noise $\sqrt{1-\rho^2}\,\mathcal{N}_N(0, I)$: $x_{i+1} = \rho x_i + \sqrt{1-\rho^2}\,\mathcal{N}_N(0, I)$.
   Generate the vector of regressors β_{true} from the uniform distribution on $[1, 5]$ and randomly set 98% of the coefficients to zero. The variance of the random effect is set to $\sigma^2_{\text{true}} = 0.1$.
   Consider a repeated measurement setting so that $z_i = e_{\lceil iq/N \rceil}$ where $\{e_j, j \leq q\}$ is the canonical basis of $\mathbb{R}^q$ and $\lceil \cdot \rceil$ denotes the upper integer part.

2. In this question, σ is assumed to be known. Run the algorithm (5) for different strategies of (γ_n, m_n):

   (i) $\gamma_n = \gamma$ for a small enough value (for example: $\gamma = 0.005$). We do NOT ask to compute the constant L given by (1). m_n increases linearly (for example: $m_n = 200 + n$).

   (ii) γ_n is decreasing (for example: $\gamma_n = 0.05/\sqrt{n}$) and m_n slowly increases (for example: $m_n = 200 + \lceil \sqrt{n} \rceil$).

   and different values of λ: explore two or three different values (including $\lambda = 30$). Since σ is assumed to be known, note that the algorithm only returns a sequence of $\mathbb{R}^p$-valued vectors $\{\beta_n, n \geq 0\}$.

   - If the algorithm converges: display on the same graph the limiting value β_∞ and the true value β_{true}.
   - At each iteration n, compute the relative error $\mathcal{E}_n = \|\beta_n - \beta_\infty\| / \|\beta_\infty\|$, the sensitivity and the precision

$$\mathsf{SEN}_n = \frac{\sum_i \mathbb{I}_{\{|\beta_{n,i}|>0\}} \mathbb{I}_{\{|\beta_{\infty,i}|>0\}}}{\sum_i \mathbb{I}_{\{|\beta_{\infty,i}|>0\}}}, \quad \mathsf{PRE}_n = \frac{\sum_i \mathbb{I}_{\{|\beta_{n,i}|>0\}} \mathbb{I}_{\{|\beta_{\infty,i}|>0\}}}{\sum_i \mathbb{I}_{\{|\beta_{n,i}|>0\}}},$$

[6] `randn` returns a standard real valued Gaussian r.v.

where $\beta_n = (\beta_{n,1}, \cdots, \beta_{n,p})$. Display these quantities as a function of the total number of Monte Carlo samples up to the current iteration i.e. display the $\mathbb{R}^2$-valued sequences $\{(\sum_{i=1}^n m_i, \mathcal{E}_n), n \geq 1\}$, $\{(\sum_{i=1}^n m_i, \mathsf{SEN}_n), n \geq 1\}$ and $\{(\sum_{i=1}^n m_i, \mathsf{PRE}_n), n \geq 1\}$.

- What do you observe ? comment the results (role of λ; fixed stepsize $\gamma_n = \gamma$ vs decreasing stepsizes; $\cdots$).

3. Modify the code to address the case when σ is unknown and has to be estimated. Run the algorithm for a choice of the sequences $\{\gamma_n, m_n, n \geq 0\}$ and a penalty factor λ. Comment the results on the estimation of σ_{true}.