ETH zürich

High Performance Computing for Science and Engineering II

P. Koumoutsakos, P. Hadjidoukas
ETH Zentrum, HG D1.1/G3
CH-8092 Zürich

Spring semester 2018

Set 07 - Sampling: toward MCMC and TMCMC

Issued: April 16, 2018
Hand in: April 23, 2018

Question 1: Sampling Importance Resampling.

This exercise will bring you closer to the concept of Transitional Markov Chain Monte Carlo. The goal here is to get samples from the distribution f having already in hand samples from a distribution g. Usually sampling g is easier than sampling f.

Assume we have samples $\{Y_i\}_{i=1}^N$ from g. We define the weights

$$\omega_i = \frac{f(Y_i)/g(Y_i)}{\sum_{i=1}^N f(Y_i)/g(Y_i)}\,, \tag{1}$$

and a random variable X with $\Pr(X = y_i \,|\, y_1, \ldots, y_N) = \omega_i$.

a) Show that in the limit $N \to \infty$, the random variable X is distributed according to f.

b) Let $\{X_i\}_{i=1}^N$ be i.i.d. samples following the same distribution as X. Show that

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N h(X_i)\right] = \mathbb{E}\left[\sum_{i=1}^N \omega_i h(Y_i)\right]\,. \tag{2}$$

Question 2: MCMC for optimization

Markov Chain Monte Carlo (MCMC) is a sampling technique used in Bayesian inference to obtain samples from the posterior. It can also be used in discrete and continuous optimization problems.

An optimization problem has the following form:

$$\min_{\vartheta\in\Theta} E(\vartheta),\ E : \mathcal{E} \to \mathcal{V}, \Theta \subset \mathcal{E}. \tag{3}$$

The set $\mathcal{V}$ should be totally ordered. Usually it is the set of real or natural numbers. If $\Theta = \mathcal{E}$, the minimization is *unconstrained*, and if $\Theta \subsetneq \mathcal{E}$, the problem is called *constrained*.

If $\mathcal{V}$ is numeric, the problem can be reformulated as follows:

$$\min_{\vartheta\in\Theta} E(\vartheta) \Leftrightarrow \max_{\vartheta\in\Theta} \exp\{-E(\vartheta)/T\},\ T > 0. \tag{4}$$

$$\Leftrightarrow \max_{\vartheta\in\Theta} \frac{\exp\{-E(\vartheta)/T\}}{\int_\vartheta \exp\{-E(\vartheta)/T\}\mathrm{d}\vartheta} := \max_{\vartheta\in\Theta} P(\vartheta; T). \tag{5}$$

where we silently ignore any integrability issues and assume E behaves nicely enough. The quantity P in (5) is a valid probability density function. If $P(\vartheta; T)$ had its mass highly concentrated around the minima of E, we could just sample ϑ from this distribution. The sampling is done by the evolution of a Markov chain using appropriate transition probabilities as explained below. The high concentration of mass around the minima is achieved by a technique known as *simulated annealing* which reduces the value of T over time as new samples are being generated.

Overview of MCMC

As we know from Exercise 5, a Markov chain is characterized by its *transition probability* $\mathbb{P}[x_{k+1} \mid x_k] = t(x_{k+1}, x_k)$. The *marginal distribution* of state $k+1$ can be written as:

$$\mathbb{P}[x_{k+1}] = \int_{x_k} \mathbb{P}[x_{k+1} \mid x_k] \, \mathbb{P}[x_k] \, \mathrm{d}x_k. \tag{6}$$

If we can let $k \to \infty$, then we obtain a distribution $p(x)$ which satisfies the following relation:

$$p(x) = \int_y t(x, y) p(y) \mathrm{d}y. \tag{7}$$

As an aside, you can check that if the Markov Chain is discrete, the finding of the stationary distribution corresponds to solving an eigenvector problem.

There are in general no guarantees that such a distribution exists. If it does, the marginal distributions will eventually converge to it. When *detailed balance* holds:

$$p(x) t(x, y) = t(y, x) p(y), \tag{8}$$

then the fixed point exists almost trivially:

$$\int_y t(x, y) p(y) \mathrm{d}y = \int_y t(y, x) p(x) \mathrm{d}y = p(x) \underbrace{\int_y t(y, x) \mathrm{d}y}_{=1} = p(x). \tag{9}$$

$P(\vartheta; T)$ is what we want our stationary distribution to be. Various MCMC algorithms are thus concerned with picking the appropriate transition probability $t(\cdot, \cdot)$.

One of the popular MCMC algorithms is the Metropolis-Hastings algorithm. For it we need a symmetric proposal distribution $q(x \mid y) = q(y \mid x)$ (note: this is not the same as $t(x, y)$):

1. Sample $\vartheta \sim q(\cdot \mid \vartheta_{k-1})$.
2. Set $h = \min\left(1, \exp\{-(E(\vartheta) - E(\vartheta_{k-1}))/T\}\right)$
3. Sample $u \sim$ Unif$(0, 1)$
4. If $u \leq h$, set $\vartheta_k = \vartheta$, else set $\vartheta_k = \vartheta_{k-1}$.

Questions

a) Show that the Metropolis Hasting algorithm induces a transition probability $t(x, y)$ that satisfies the detailed balance condition.

In the next questions you will implement the Metropolis Hastings algorithm maximum to find the maximum likelihood solution the linear regression problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2. \tag{10}$$

b) Implement a routine `metropolis_step`. It takes the current sample $\boldsymbol{\beta}_k$ (and other needed values, such as X, y, T) and generates the $\boldsymbol{\beta}_{k+1}$ according to the Metropolis Hastings algorithm. It also returns the new value of the loss function $E(\boldsymbol{\beta}_{k+1})$.

   Pick an appropriate symmetric proposal distribution, such as $q(\boldsymbol{\beta}_{k+1} \mid \boldsymbol{\beta}_k) = \mathcal{N}(\boldsymbol{\beta}_{k+1}; \boldsymbol{\beta}_k, \lambda)$.

c) Implement the `mcmc_routine` which runs the `metropolis_step` K times and keeps track of:

   1. the $\boldsymbol{\beta}^*_{k^*}$ and k^* for which $E(\cdot)$ is the smallest among all the generated samples,
   2. all the values $k, E(\boldsymbol{\beta}_k)$.

   The routine should take as parameter the function $T(k)$ used to control the parameter T in the Metropolis Hastings step.

d) Find experimentally how many iterations $\bar{k}$ are needed *on average* to reach the minimum within a given tolerance: $\|\boldsymbol{\beta}^*_{k^*} - \hat{\boldsymbol{\beta}}\| < \varepsilon$. Repeat the experiment for multiple values of ε. The $\hat{\boldsymbol{\beta}}$ can be obtained via least squares solutions. Produce a plot of $\bar{k}(\varepsilon)$ vs. ε. Do this for different decreasing functions $T(k)$, such as $T_0/\log(k), T_0/k, T_0\exp(-cT)$.

Let $X \in \mathbb{R}^{n\times p}$. Then set $n = 2^{15}, p = 30$. You may generate $\boldsymbol{\beta}$ and X as you wish, but let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon,\ \epsilon \sim \mathcal{N}(0, 0.1)$.