P. Koumoutsakos
ETH Zentrum, CLT E 13
CH-8092 Zürich

HW 3 (Part 1 of 2): Sampling Methods

Issued: March 18, 2019
Due Date: April 1, 2019 10:00am

Task 1: Inversion Method for Gaussian Sampling (20 Points)

Consider the random variable

$$X = \sqrt{2}\,\mathrm{erf}^{-1}(2U-1), \text{ where } U \sim \mathcal{U}(0,1).$$

The error function is defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,\mathrm{d}t.$$

a) (15 Points) Use the inversion method to show that $X \sim \mathcal{N}(0,1)$.

We apply the inversion method to the density of the standard normal distribution,

$$p_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

The cumulative distribution gives,

$$\begin{aligned} F_X(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right)\mathrm{d}t, \\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left(-\frac{t^2}{2}\right)\mathrm{d}t, \\ &= \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{x/\sqrt{2}} \exp\left(-t^2\right)\mathrm{d}t, \\ &= \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right). \end{aligned}$$

Substituting $x = \sqrt{2}\,\mathrm{erf}^{-1}(2u-1)$,

$$u = F_X(x),$$

which gives the desired result.

b) (5 Points) Explain why the above formula is not used to generate normally distributed random numbers in practice (2 lines).

In practice, the inverse of the error function is not available or not cheap to compute. The Box-Muller algorithm is much cheaper to perform.

Task 2: Importance Sampling (30 Points)

Taking advantage of the importance sampling method, we can rewrite the integral

$$\mathbb{E}_f[h(X)] = \int h(x)\, f(x)\, dx, \tag{1}$$

into

$$\mathbb{E}_g\Big[h(X)\frac{f(X)}{g(X)}\Big] = \int h(x)\, \frac{f(x)}{g(x)}\, g(x)\, dx\,, \tag{2}$$

for any density g that satisfies $\operatorname{supp}(f) \subseteq \operatorname{supp}(g)$. Then the estimator for Eq.1:

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(Y_i)\,, \tag{3}$$

,where Y_i are i.i.d. samples from f, turns into the estimator for Eq.2

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(X_i)\,\frac{f(X_i)}{g(X_i)}\,, \tag{4}$$

where X_i are i.i.d. samples from g.

a) (15 Points) Show that the optimal choice of g, in the sense that it minimizes the variance of estimator (Eq.4), is given by

$$g^{\star}(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)\, dx}\,. \tag{5}$$

Can you argue why this result is not in practice employed?
(hint: use Jensen's inequality: In the context of probability theory, if X is a random variable and ϕ is a convex function, then: $\phi(\mathbb{E}[\mathrm{X}]) \leq \mathbb{E}[\phi(\mathrm{X})]$).

In the following the variance of $\hat{I}$ is a function of g. We use the symbolism $\mathrm{Var}^{\hat{I}}(g)$. The variance of $\hat{I}$ is given by

$$\mathrm{Var}^{\hat{I}}(g) = \mathbb{E}_g\left[\frac{h^2(X)f^2(X)}{g^2(X)}\right] - \left(\mathbb{E}_g\left[\frac{h(X)f(X)}{g(X)}\right]\right)^2\,. \tag{6}$$

Notice that the second term does not depend on g, since the denominator cancels out with the density of the expectation. Moreover, using the Jensen's inequality we can identify a lower bound of this variance. We can write based on the Jensen's inequality:

$$\mathbb{E}_g\left[\frac{h^2(X)f^2(X)}{g^2(X)}\right] \geq \left(\mathbb{E}_g\left[\frac{h(X)f(X)}{g(X)}\right]\right)^2 = \left(\int h(x)f(x)\, dx\right)^2\,. \tag{7}$$

Observe that the lower bound is independent of g. In the following we show that this is a strict lower bound. The lower bound can be indeed attained if we choose g by Eq.(5), i.e., plug in $g^{\star}$ in the left hand side of the inequality (7) and conclude that the lower bound is attained at $g^{\star}$. This proves that $g^{\star}$ minimizes the variance. The result cannot be used directly because it involves the computation of the integral that we are looking for.

b) (15 Points) Write a program to estimate the probability $P(X > 4.5)$ for $X \sim \mathcal{N}(0,1)$ using:

1. the estimator (3), with $h \equiv 1$, if $(X > 4.5)$ and $h \equiv 0$ otherwise.
2. the estimator (4), where $h \equiv 1$, if $(X > 4.5)$ and $h \equiv 0$ otherwise, f the density of the normal distribution and g the density of an exponential distribution, truncated at 4.5 with scale 1.

Compare your results with the exact value that can be obtained using the cumulative distribution function of the normal distribution. What do you observe for $N = 10^4$? Can you argue on the performance of the two estimators?

The estimator for $g \equiv 1$ is, most of the time, equal to zero, while the importance sampling estimator is able to produce a reasonable result.

Task 3: MCMC: Hastings and application to coin-toss problem (40 Points)

In his 1970's paper Hastings proposed a general form of the acceptance probability $\alpha(x|y)$ for MCMC algorithms,

$$\alpha(x|y) = \frac{s(x|y)}{1 + \frac{q(x|y)p(y)}{q(y|x)p(x)}}, \tag{8}$$

where q represents the proposal distribution (in general non-symmetric), p the stationary distribution, and s any symmetric function ($s(x|y) = s(y|x)$) which can guarantee that $\alpha(x|y) \leq 1$ for all x, y.

a) (10 Points) Show that the transition probability $t(x|y) = \alpha(x|y)q(x|y)$ with α chosen as above satisfies the *detailed balance*. Note that you do not have to prove that the expression for the transition probability holds.

Detailed balance is the condition: $t(x|y)p(y) = t(y|x)p(x)$.

$$\begin{aligned} t(x|y)p(y) &= \alpha(x|y)q(x|y)p(y) = \\ &= \frac{s(x|y)q(x|y)p(y)}{1 + \frac{q(x|y)p(y)}{q(y|x)p(x)}} \cdot \frac{q(y|x)p(x)}{q(y|x)p(x)} \\ &= \frac{s(x|y)q(x|y)p(y)q(y|x)p(x)}{q(y|x)p(x) + q(x|y)p(y)} \\ &= \frac{s(x|y)q(y|x)p(x)}{\frac{q(y|x)p(x)}{q(x|y)p(y)} + \frac{q(x|y)p(y)}{q(x|y)p(y)}} \\ &= \frac{s(x|y)q(y|x)p(x)}{1 + \frac{q(y|x)p(x)}{q(x|y)p(y)}} \\ &= \frac{s(y|x)q(y|x)p(x)}{1 + \frac{q(y|x)p(x)}{q(x|y)p(y)}} \\ &= \alpha(y|x)q(y|x)p(x) = t(y|x)p(x). \end{aligned}$$

Grading scheme:

3 points for identifying the detailed balance condition. 7 points for the derivation from $t(x|y)p(y) \to t(y|x)p(x)$.

b) (10 Points) Show that the Metropolis choice: $\alpha_M(x|y) = \min\{1, \frac{q(y|x)p(x)}{q(x|y)p(y)}\}$ is a special case of the above formula. What is $s(x|y)$ in this case?

We show it by construction.

Denote $r(x|y) = \frac{q(y|x)p(x)}{q(x|y)p(y)}$. Then the Metropolis choice α_M and the general choice α may be rewritten as:

$$\alpha_M(x|y) = \min\{1, r(x|y)\},\ \alpha(x|y) = \frac{s(x|y)}{1 + \frac{1}{r(x|y)}}.$$

If $r(x|y) < 1$:

$$\alpha_M(x|y) = r(x|y) = r(x|y)\frac{1 + \frac{1}{r(x|y)}}{1 + \frac{1}{r(x|y)}} = \frac{1 + r(x|y)}{1 + \frac{1}{r(x|y)}} \Rightarrow s(x|y) = 1 + r(x|y).$$

If $r(x|y) \geq 1$:

$$\alpha_M(x|y) = 1 = \frac{1+\frac{1}{r(x|y)}}{1+\frac{1}{r(x|y)}} \Rightarrow s(x|y) = 1 + \frac{1}{r(x|y)}.$$

Hence:

$$s(x|y) = \begin{cases} 1 + r(x|y) & r(x|y) < 1 \\ 1 + \frac{1}{r(x|y)} & r(x|y) \geq 1 \end{cases}.$$

To finish the proof, we need to show that $s(x|y) = s(y|x)$. Note that $r(y|x) = \frac{1}{r(x|y)}$. Then:

$$\begin{aligned} s(y|x) &= \begin{cases} 1 + r(y|x) & r(y|x) < 1 \\ 1 + \frac{1}{r(y|x)} & r(y|x) \geq 1 \end{cases} = \begin{cases} 1 + \frac{1}{r(x|y)} & \frac{1}{r(x|y)} < 1 \\ 1 + r(x|y) & \frac{1}{r(x|y)} \geq 1 \end{cases} \\ &= \begin{cases} 1 + \frac{1}{r(x|y)} & r(x|y) \geq 1 \\ 1 + r(x|y) & r(x|y) < 1 \end{cases} = s(x|y). \end{aligned}$$

Grading scheme:
5 points for deriving the expression for $s(x|y)$. 5 points for showing that $s(x|y) = s(y|x)$.

c) (5 Points) It is often a good practice to convert your calculations for the MCMC algorithm in the logarithmic scale. Write down how the formulation of one step of the Metropolis-Hastings algorithm changes, if we convert all probabilities to the log-scale. Can you think of a reason why this formulation can be advantageous for numerical applications?

The acceptance criterion of the metropolis hastings (MH) algorithm is given by:

$$\alpha_M(x|y) = \min\left\{1, \frac{q(y|x)p(x)}{q(x|y)p(y)}\right\}$$

By taking the logarithm on both sides we get:

$$\begin{aligned} \log\big(\alpha_M(x|y)\big) &= \log\min\left\{1, \frac{q(y|x)p(x)}{q(x|y)p(y)}\right\} \overbrace{=}^{\log \text{ is strictly increasing}} \\ &= \min\log\left\{1, \frac{q(y|x)p(x)}{q(x|y)p(y)}\right\} \\ &= \min\left\{0, \log\left(\frac{q(y|x)p(x)}{q(x|y)p(y)}\right)\right\} \\ &= \min\Big\{0, \log q(y|x) + \log p(x) - \log q(x|y) - \log p(y)\Big)\Big\} \end{aligned}$$

For symmetric proposal distribution $q(y|x) - q(x|y) = 0$ so that

$$\log\big(\alpha_M(x|y)\big) = \min\Big\{0, \log p(x) - \log p(y)\Big\}$$

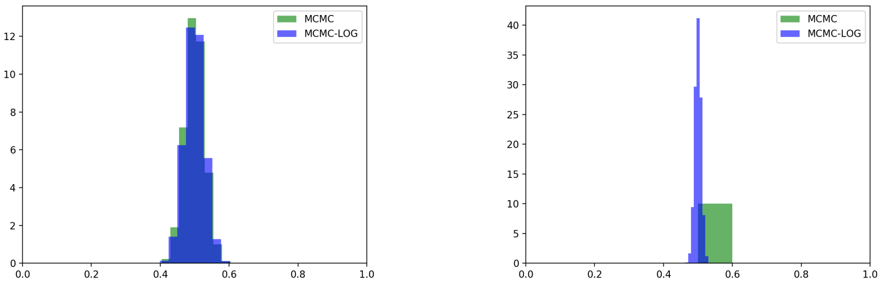
The steps of the MCMC in the log scale with a maximum number of iteration I_{max} are given by:

```
Set of samples $S = \emptyset$
Start with initial sample at $x_{old}$
$S = S \bigcup \{x_{old}\}$
while $i \leq I_{max}$ do
    Propose new sample $x_{new} \sim q(X|x_{old})$
    Sample from a uniform $u \sim U[0,1]$
    Calculate $\log(\alpha_M(x_{new}|x_{old})) = \min\big\{0, \log p(x_{new}) - \log p(x_{old})\big\}$
    if $\log(u) < \log(\alpha_M(x_{new}|x_{old}))$ then
        $S = S \bigcup \{x_{new}\}$
        $x_{old} = x_{new}$
    else
        $S = S \bigcup \{x_{old}\}$
    end if
end while
```

The logarithm of α_M can be used to accept/reject the sample. This is quite useful in practice as their use can avoid overflow and underflow errors.

d) (10 Points) We provide you with a **skeleton code** that implements MCMC based on the Metropolis-Hastings rejection criterion in order to sample from the posterior distribution $p(H|\boldsymbol{d})$ from the coin flip example in Section 2.2 of the Lecture Notes. Recall that from the Bayes theorem: $p(H|\boldsymbol{d}) = p(\boldsymbol{d}|H)p(H)$. As a proposal distribution from moving from state x to state y, $q(y|x)$, we use the **symmetric** Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.1$.
You are asked to implement the version using logarithmic scale of the probabilities involved in the provided code. Fill in the "TODO"s in the provided code.
Run your code and plot a histogram of both the samples drawn from the original MCMC and the modified MCMC in the logarithmic scale. Compare the two results.

After implementing the log version of MCMC and running the code we get the result depicted in Figure 1a. Both methods are sampling a distribution around 0.5 as expected, as a coin tossed 300 times, 150 of them showing heads is most probably a fair coin. The two methods are in theory equivalent.



(a) Result with 300 tosses, half of them heads.

(b) Result with 3000 tosses, half of them heads.

e) (5 Points) Change the number of tosses to 3000 and the number of heads to 1500. Try to plot a histogram of the samples drawn with the original MCMC and the modified MCMC in logarithmic scale. What do you observe?

The result in plotted in Figure 1b. In case we change the number of tosses to 3000 we observe that the original version without the log-scale encounters overflow problems and NaN values. This is due to the fact that in order to evaluate the posterior we have large exponents of probabilities (< 1) and underflow (or overflow problems in divisions) arise. The modified version in the log-scale alleviates these numerical stability problems. The histogram is more concentrated than the case of 300 tosses/150 heads, as we now have a larger experiment of 3000 toses/1500 heads, which means that we are more certain that the coin is fair. The certainty can be evaluated based on the width (how thin) the histogram is.

Guidelines for reports submissions:

- Submit a report in .pdf format and the two solution codes (recommended: in python) by Monday 01.04.
- The final version of your report should include the solutions to the second part of Homework 3, which will be handed out on Monday 25.03.2019.