S2: Statistical Methods for Data Science

Example Sheet 2

The goal of this example sheet it to give you a thorough understanding of the "why" of stochastic sampling in general, and Markov chain Monte Carlo in particular.

Rosenbrock's "banana" function in n dimensions is defined as

$$f_n(\mathbf{x}) = \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right],$$

where $\mathbf{x} \in \mathbb{R}^n$.

1. Show that $f_n(\mathbf{x})$ has a single *global* minimum at $f_n(1, 1, \ldots, 1) = 0$. (Depending on the value of n, there may be several other *local* minima.)

Consider the target probability distribution over the space $\mathbf{x} \in \mathbb{R}^n$ with a log probability density function given by

$$\log P_n(\mathbf{x}) = -f_n(\mathbf{x}) + \text{const.}$$

2. Plot the $n = 2$ dimensional version of the distribution $P_2(x_1, x_2)$ as a function of x_1 and x_2.

The Rosenbrock function (or the distribution defined from it) is used to test optimisation and sampling algorithms. Despite its relatively simple analytic expression, the narrow, curving, ridge-like structure of the function makes it challenging to sample from even in relatively low numbers of dimensions.

For the rest of this question you should consider the case of $n = 3$ dimensions.

Your task is to calculate the expectation of $|\mathbf{x}|$ or, in other words, evaluate

$$\begin{aligned}\mathrm{E}_{\mathbf{x}}\left[|\mathbf{x}|\right] &= \int \mathrm{d}\mathbf{x}\ |\mathbf{x}| P_n(\mathbf{x}), \\ &= \int_{-\infty}^{\infty} \mathrm{d}x_1 \int_{-\infty}^{\infty} \mathrm{d}x_2 \int_{-\infty}^{\infty} \mathrm{d}x_3\ \sqrt{x_1^2 + x_2^2 + x_3^2} P_3(x_1, x_2, x_3).\end{aligned}$$

You should attempt to do this in three different ways:

3. First, you should attempt to calculate this integral using a deterministic method of your choice. E.g. a simple rectangle rule, quadrature method, or something similar. You will find that this doesn't work. You are not expected to get a good answer from this method, the goal is instead for you to understand the need for Monte Carlo methods. (If you are somehow able to get an accurate answer by this method, then think about the case $n = 4$!)

4. Second, code up your own simple version of the Metropolis-Hastings algorithm (or, if you prefer, any of the other sampling algorithms described in the lectures). Use this to draw stochastic samples from the target distribution and use these samples to evaluate the integral.

5. Third, use a stochastic sampling software package to draw samples from the target distribution and evaluate the integral. There are many such packages available, it does not have to use an algorithm described in lectures. You might use a software package you are using in your project, or treat this as an opportunity to learn a new tool. (If you don't know of any suitable packages, one option is `zeus`; see this link).

With your stochastic samples you should (i) make a corner plot, (ii) estimate $\mathrm{E}_{\mathbf{x}}[|\mathbf{x}|]$, and (iii) estimate the uncertainty in your answer for $\mathrm{E}_{\mathbf{x}}[|\mathbf{x}|]$.

Think about when it is important to "thin" your Markov chains[1] and what changes would be needed to adapt your solution work in the case $n = 4$. These points will be discussed further in the examples class.

[1]The `zeus` package has a easy-to-use function for calculating the autocorrelation length which can be used to find an appropriate thinning factor