Annealed Importance Sampling Report

Usama Kamran, Patrick Ghallager, and Tallis Bowers

December 15, 2018

Introduction

Importance sampling and Markov Chain Monte Carlo (MCMC) sampling are methods we have looked at in detail this semester. Both of these procedures provide methods for estimating expectations of functions with respect to some underlying distribution from which it is impossible or infeasible to sample directly. We saw that both of these methods have limitations, however. Most notably, importance sampling provides a very poor, albeit consistent, estimator when the target distribution is on a high dimensional space, and MCMC has trouble converging to stationary distribution when the target distribution is multi-modal.

The method of annealed importance sampling (AIS) was originally designed as an alternative sampler for target distributions which are not easily sampled from via the methods mentioned in the previous paragraph. In some sense, AIS combines the ideas behind importance sampling and MCMC. At a very basic level AIS creates many Markov Chains, each of which provides a single sample point. The collection of "transition" points in a Markov Chain, are then used to create a weight associated with the final sample point coming from this chain. The details of this process are discussed in the proceeding section.

Theory

Given an intractable target distribution p_0, AIS seeks to find a tractable proposal distribution p_n as well as intermediate distributions p_i for $i = n-1, n-2, ..., 2, 1$ so that each distribution is in some sense "close" to its preceeding distribution. The only necessary constraint on this intermediate distributions is that the support of p_j is always covered by p_{j+1}. In order to move through these distributions from the proposal to the target with a Markov Chain, we need to define transition probabilities $T_j(x, x')$ which will allow us to move from a point sampled from p_{j+1} to a point sampled from p_j. Just as in MCMC we require that each of these transition probabilities be ergodic and have p_j as their stationary distributions.

Because only the probability density p_n need be tractable in the above set-up, we will let f_j be a function such that $p_j \propto f_j$ for all $j = 0, 1, ..., n-1$. The annealed importance sampling algorithm then proceeds as follows:

$$\begin{cases} \text{Sample } x_{n-1} \text{ from } p_n(\cdot) \\ \text{Sample } x_{n-2} \text{ from } T_{n-1}(x_{n-1}, \cdot) \\ \vdots \\ \text{Sample } x_0 \text{ from } T_1(x_1, \cdot). \end{cases}$$

The point x_0 in the above algorithm needs an associated weight for calculating expectations with respect to the target distribution p_0, and this weight is given by

$$w = \frac{f_{n-1}(x_{n-1})}{f_n(x_{n-1})} \frac{f_{n-2}(x_{n-2})}{f_{n-1}(x_{n-2})} \cdots \frac{f_1(x_1)}{f_2(x_1)} \frac{f_0(x_0)}{f_1(x_0)}.$$

In order to see that this process indeed produces an "importance sample", let us consider the joint distribution of the points $(x_{n-1}, x_{n-2}, ..., x_1, x_0)$ from the algorithm above. By construction, this joint density is proportional to

$$g(x_{n-1}, x_{n-2}, ..., x_1, x_0) = f_n(x_{n-1}) T_{n-1}(x_{n-1}, x_{n-2}) \cdots T_2(x_2, x_1) T_1(x_1, x_0).$$

Now, we want to consider the joint density of $(x_{n-1}, x_{n-2}, ..., x_1, x_0)$ if x_0 had been sampled from p_0 directly, and the other points are sampled from the reversals of the transition kernels. This is, in some sense, the "target" distribution for $(x_{n-1}, x_{n-2}, ..., x_1, x_0)$. We can define the reversals of the above transitions as

$$\tilde{T}_j(x^*, x) = T_j(x, x^*) \frac{f_j(x)}{f_j(x^*)}$$

for all $j = 1, 2, ..., n-1$. Then, we can see that the "target" joint density is proportional to

$$f(x_{n-1}, x_{n-2}, ..., x_1, x_0) = f_0(x_0) \tilde{T}_1(x_0, x_1) \cdots \tilde{T}_{n-2}(x_{n-3}, x_{n-2}) \tilde{T}_{n-1}(x_{n-2}, x_{n-1})$$

$$= \frac{f_0(x_0)}{f_1(x_0)} T_1(x_1, x_0) \frac{f_1(x_1)}{f_2(x_1)} T_1(x_2, x_1) \cdots \frac{f_{n-2}(x_{n-2})}{f_{n-1}(x_{n-2})} T_{n-1}(x_{n-1}, x_{n-2}) f_{n-1}(x_{n-1}).$$

Finally, using the basic idea of importance sampling, we see that we can find an expectation of a function with respect to f using samples from g by calculating a weighted average of the function with weights given by

$$\frac{f(x_{n-1}, x_{n-2}, ..., x_1, x_0)}{g(x_{n-1}, x_{n-2}, ..., x_1, x_0)} = \frac{f_{n-1}(x_{n-1})}{f_n(x_{n-1})} \frac{f_{n-2}(x_{n-2})}{f_{n-1}(x_{n-2})} \cdots \frac{f_1(x_1)}{f_2(x_1)} \frac{f_0(x_0)}{f_1(x_0)} = w,$$

which are the weights defined in the algorithm above. Therefore, the annealed importance sampling process with give a consistent estimator of any expectation we want to calculate. In the proceeding sections, we will look at how AIS is often used in practice.

Method

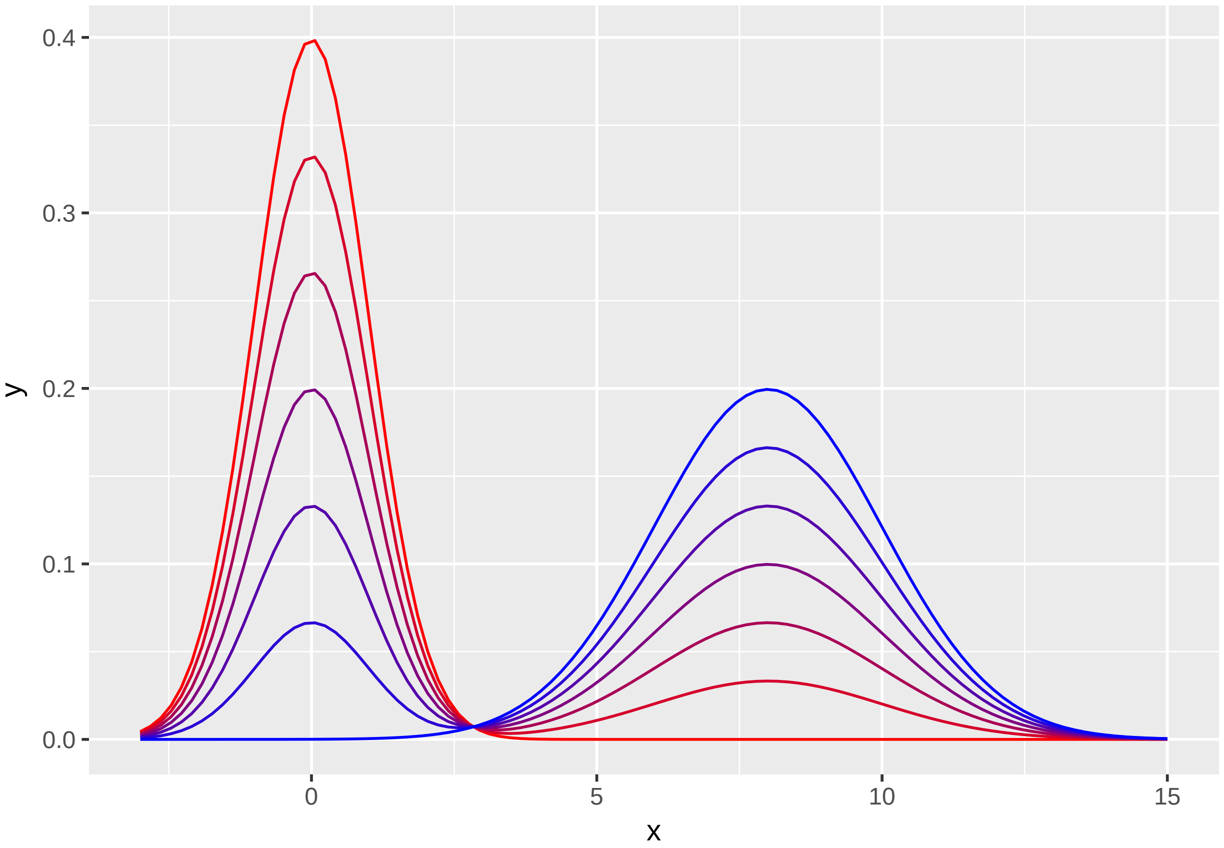
We have our algorithm, but the problems of intermediate distributions and transition proposals remain. In practice, these functions are defined in a simple way. Intermediate distributions are often set as annealing between our initial proposal distribution, $f_n(x)$, and our target disbtribution, $f_0(x)$:

$f_j(x) = f_0(x)^{\beta_j} f_n(x)^{1-\beta_j}$

where $1 = \beta_0 > \beta_1 > \cdots \beta_{n-1} > \beta_n = 0$.

In order to define transitions between these intermediate distributions, we will just use Metropolis-Hastings. So, we can simply run MCMC for a fixed number of iterations to move from p_{j+1} to p_j. After each transition, we can recalculate the weights, and the algorithm can be run until the last MCMC allows us to sample from p_1. This final sample, arrived at using MCMC, is a weighted sample from the target distribution p_0.

As a visual example, below is an annealing betwen $\mathcal{N}(0, 1)$ and $\mathcal{N}(8, \sqrt{2})$ with 5 intermediate distributions and linear spacing of the β_j:



Application

We will attempt to estimate two expectations using AIS in this section. First, we will estimate the expectation of a simple Gaussian in order to insure that we can correctly code the AIS algorithm. We will assume that we have created a valid annealed importance sample if the estimated expectation is close to the known true expectation of the target Gaussian. The second expectation will be that of a multi-dimensional distribution which would be difficult to estimate using MCMC or importance sampling alone.

For our first application of the AIS algorithm, we want to estimate the mean of a Gaussian with mean 5 and variance 3 $\mathcal{N}(5,3)$ using the standard normal $\mathcal{N}(0,1)$ as a proposal distribution. If implemented correctly, the AIS algorithm should give a sample $(x^{(1)}, x^{(2)}, ..., x^{(n)})$ and associated weights $(w^{(1)}, w^{(2)}, ..., w^{(n)})$ whose weighted average is near the true mean of the target distribution, i.e. $\sum_{i=1}^{n}(x^{(i)}w^{(i)})/\sum_{i=1}^{n} w^{(i)} \approx 5$. There are several hyperparameters which we need to consider in order to implement this algorithm. We need to choose the number of transition probabilities t, the number of MCMC steps used with each transition kernel M, and the number of samples to draw from the proposal distribution N. As this is our first implementation of AIS, we would like to investigate how each of these hyperparameters affects the accuracy of the final estimate. In order to do this, we will vary the hyperparameters over a range of possible combinations and run the AIS algorithm 20 times for each combination, i.e. get 20 estimates of the target expectation for each hyperparameter combination. These 20 estimates will then be used to estimate the mean and standard deviation of the estimated expectation for each combination. The results are presented in the table below.

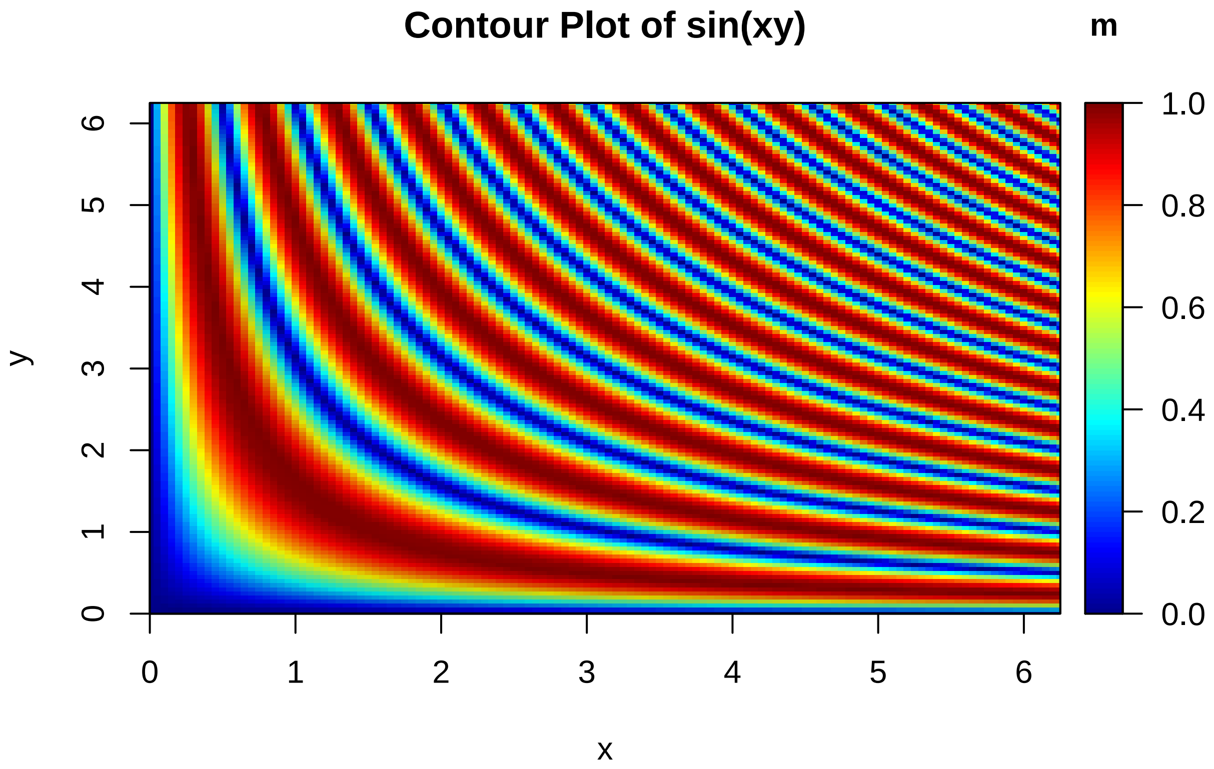
	t	M	N	Mean	SD
2	20	100	1000	5.023	0.098
3	40	100	1000	5.009	0.075
4	20	200	1000	4.974	0.106
5	20	100	2000	5.021	0.088

In all of the hyper parameter settings, the mean of the estimates is very close to the true mean (5). We might be a little surprised to see such a large standard deviation in all of the settings, however, it is helpful to note that this methodology was not intended for use in a such a nice setting. Although its error may seem large in this setting, because we can use regular Monte Carlo here and get a near perfect estimate, the settings for which AIS was designed are ones in which no other good estimators exist. In more complicated settings, a standard deviation of 0.1 on a mean might be considered very acceptable. It is important to note how the standard deviation changes with the hyperparameters here, though. It seems that doubling the number of MCMC steps at each transition didn't improve accuracy at all. This is probably because after just 100 steps we were able to attain a good sample from the next transition distribution. Doubling the sample size and number of transition densities, however, was effective in increasing accuracy.

In our second application of AIS, we will estimate a probability with respect to the probability density proportional to the function

$$f(x,y) =\mid sin(xy) \mid \cdot I(x \in (0,2\pi), y \in (0,2\pi)).$$

This function would be difficult to normalize because it is a complicated function of x and y, and the variables are not independent. We know, however, that the density must be symmetric, and thus we must have $P(X < Y) = 0.5$ with respect to this distribution. This will be the quantity we estimate because we know its true value. Now, let us look at a contour plot of the function proportional to the density from which we want to sample.



As we can see this function is multi-modal, and thus difficult for traditional MCMC sampling methods. We hope that this will be a good setting in which to demonstrate that AIS can outperform simple annealed sampling. In order to assess this, we will attempt to estimate the above probability using both simple annealed sampling and AIS. The results for both methods are listed in the following table.

Method	Mean	SD
Annealed	0.498	0.017
AIS	0.503	0.022

Conclusion

References

Kristiadi, A. (2018). *Introduction to Annealed Importance Sampling*. Retrieved from https://wiseodd.github.io/techblog/2017/12/23/annealed-importance-sampling/

Neal, R. M. (2001). Statistics and computing. *Annealed importance sampling*, 11(2), 125-139.