# MONTE CARLO METHODS

**He Jiang**

University of Southern California
jian567@usc.edu

September 2, 2019

## 1 Introduction

In [2], the Monte Carlo method is described as "representing the solution of a problem as a parameter of a hypothetical population, and using a **random sequence of numbers** to construct a sample of the population, from which statistical estimates of the parameter can be obtained."

In the book, this example problem is integration or summation:

$$\mathbf{s} = \sum_{\mathbf{x}} p(\mathbf{x}) f(\mathbf{x}) = \mathbb{E}_p[f(\mathbf{x})] \tag{1}$$

or

$$\mathbf{s} = \int p(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \mathbb{E}_p[f(\mathbf{x})] \tag{2}$$

As these problems are difficult to compute, we draw $n$ samples $i.i.d.$ from $p$ and replace the expectation with the emprical average:

$$\hat{s}_n = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}^{(i)}) \tag{3}$$

We justify our method by the following properties:

- Unbiased estimator

$$\mathbb{E}[\hat{s}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f(\mathbf{x}^{(i)})] = \sum_{i=1}^{n} s = s. \tag{4}$$

- Consistent estimator. The law of large numbers ensures this property: as the number of sample increases indefinitely, the sequence of estimate converge in probability to its expectation:

$$\lim_{n \to \infty} \hat{s}_n = s. \tag{5}$$

- Central limit theorem. This property gives us an idea that how close our empirical estimate will be as the number of samples increases. We have

$$\sqrt{n}(\hat{s}_n - s) \xrightarrow{p} \mathcal{N}(0, \text{Var}[f(\mathbf{x})]) \tag{6}$$

  This means that the variance of the empirical mean is in the order of $1/n$.

However, in many cases it's hard to directly draw sample from $p$. In the following sections we present two ways to approximately sample from $p(\mathbf{x})$.

## 2 Importance Sampling

### 2.1 Formulation

Instead of directly sampling from $p$, an alternative is to directly sample from a surrogate distribution $q$. We can do that because there is no single "correct" way to regard which part in the summation problem belongs to the distribution $p(x)$. We have the following transformation:

$$p(\mathbf{x})f(\mathbf{x}) = q(\mathbf{x})\frac{p(\mathbf{x})f(\mathbf{x})}{q(\mathbf{x})} \tag{7}$$

Using this formulation, we can sample from $q$ and transform our estimator as an "importance sampling estimator":

$$\hat{s}_q = \frac{1}{n}\sum_{i=1,\mathbf{x}^{(i)}\sim q}^{n}\frac{p(\mathbf{x}^{(i)})f(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})} \tag{8}$$

We can quickly check that this new estimator is still unbiased. However, its variance is given in the following form:

$$\text{Var}[\hat{s}_q] = \text{Var}[\frac{p(\mathbf{x})f(\mathbf{x})}{q(\mathbf{x})}] \tag{9}$$

This variance can be greatly sensitive to the choice of $q$. The minimum variance occurs when $q$ is:

$$q^*(\mathbf{x}) = \frac{p(\mathbf{x})|f(\mathbf{x})|}{Z} \tag{10}$$

where $Z$ is the normalization factor. When $f(\mathbf{x})$ does not change sign, $\text{Var}[\hat{s}_{q^*}] = 0$ and we only need one example to get the exact answer.

Ideally, the choice of $q$ should be as close as possible to the original $q^*$. There could be some distribution that are both feasible and close to $q^*$.

And there are some failure cases when we choose the wrong $q(\mathbf{x})$. When $p(\mathbf{x}) \ll q(\mathbf{x})$, we are drawing a lot of useless samples. When $p(\mathbf{x}) \gg q(\mathbf{x})$, then the case contribute much to the variance.

### 2.2 Biased Importance Sampling

Sometimes we could also use arbitrary "distribution" without knowing the normalization factor. The unbiased importance sampling estimator is given by

$$\hat{s}_{BIS} = \frac{\sum_{i=1}^{n}\frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}f(\mathbf{x}^{(i)})}{\sum_{i=1}^{n}\frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}} \tag{11}$$

$$= \frac{\sum_{i=1}^{n}\frac{p(\mathbf{x}^{(i)})}{\tilde{q}(\mathbf{x}^{(i)})}f(\mathbf{x}^{(i)})}{\sum_{i=1}^{n}\frac{p(\mathbf{x}^{(i)})}{\tilde{q}(\mathbf{x}^{(i)})}} \tag{12}$$

$$= \frac{\sum_{i=1}^{n}\frac{\tilde{p}(\mathbf{x}^{(i)})}{\tilde{q}(\mathbf{x}^{(i)})}f(\mathbf{x}^{(i)})}{\sum_{i=1}^{n}\frac{\tilde{p}(\mathbf{x}^{(i)})}{\tilde{q}(\mathbf{x}^{(i)})}} \tag{13}$$

$$\tag{14}$$

That is, we can scale $p$ and $q$ respectively. However, this new estimator is biased. The issue will be solved when $n \to \infty$ as $\sum_{i=1}^{n}\frac{p(\mathbf{x})}{q(\mathbf{x})}$ will converge to 1.

### 2.3 Application

The application of importance sampling includes the following:

- accelerating training by sampling gradient instead of plain SGD [1, 3].
- facilitate training of language models by sampling negative entries.
- train classifiers where most of the total value of the cost function comes from a small number of misclassified examples. (Sample difficult examples)

## 3 Markov Chain Monte Carlo

In this section, we will first bring on an example when directly sampling is difficult. Take energy-based method as an example:

$$p(\mathbf{x}) \propto \exp(-E(\mathbf{x})) = \exp(-\sum_{C} E_C(\mathbf{x}_c)) \tag{15}$$

Since this is an undirected graphical model, any exact sampling involves calculating the partition function. And different than a directed graphical model, we can not use **ancestral sampling**. Consider the following example : we have a joint distribution $p(a, b)$. In order to sample $a$ we have to sample from $p(a|b)$, which requires having a $b$. And in order to sample $b$ we have to sample from $p(b|a)$. The case is, we have to either sample from the marginal distribution $p(a)$ or $p(b)$, otherwise we got into a chicken-and-egg problem.

As an alternative, we can define a random process (the **Markov chain**) that converges to the target distribution $p$ after some time steps. After the distribution is closed to $p$, we can sample from this random process as the approximated samples from $p$. The term **Markov chain** refers to a discrete time stochastic process on a general state space that has the Markov property: the future is independent of the past given the present state.

Consider the sampling problem whose states are positive integers. We can describe the distribution $q$ using a vector $\mathbf{v}$, with

$$q(x = i) = v_i \tag{16}$$

Using this formulation, we know the distribution of next step given the current distribution:

$$q^{(t+1)}(x') = \sum_{x} q^{(t)}(x) T(x'|x). \tag{17}$$

We can represent the transition operator $T$ using a matrix $\mathbf{A}$:

$$A_{i,j} = T(\mathbf{x}' = i|\mathbf{x} = j) \tag{18}$$

The matrix $\mathbf{A}$ is called **stochastic matrix**. Using this formulation, we can directly use matrix multiplication to describe the change of distribution over one step of time:

$$\mathbf{v}^{(t+1)} = \mathbf{v}^{(t)}\mathbf{A} \tag{19}$$

Therefore we know that the distribution at time $t$ is the the following:

$$\mathbf{v}^{(t)} = \mathbf{v}^{(0)}\mathbf{A}^t \tag{20}$$

Here, it is interesting to take a dive into the property of $\mathbf{A}$. First, the largest eigenvalue of a stochastic matrix is 1. The row vector $\boldsymbol{\pi}$ associated with eigenvalue 1 is called **stationary distribution** or **equilibrium distribution**:

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{A}. \tag{21}$$

Each stochastic matrix have at least one stationary distribution, and under some mild condition (finite, ergodic, irreducable), the eigenvector is unique. Then we know all other eigenvalues are smaller than 1. We call the difference of the first and the second eigenvector, $1 - \lambda_2$, to be the **spectral gap**. The spectral gap explicitly gives us the estimate of the **mixing time**. For some problems like random walk on the graph, the spectral gap can be efficiently bounded by some structure of the problem.

### 3.1 Gibbs Sampling

In order to build the Markov chain, we have to find a transition distribution. **Gibbs Sampling** is the most straightforward way to update the sample by selecting one variable $x_i$ and sampling from $p_{model}(\mathbf{x})$ conditioned on all of its neighbors. It is also possible to simultaneously update many variables in this way, and the method is called **block Gibbs sampling**.

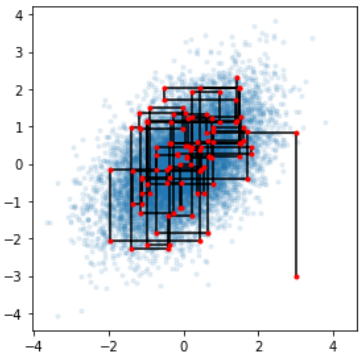


Figure 1: Gibbs sampling of a 2-d Gaussian distribution

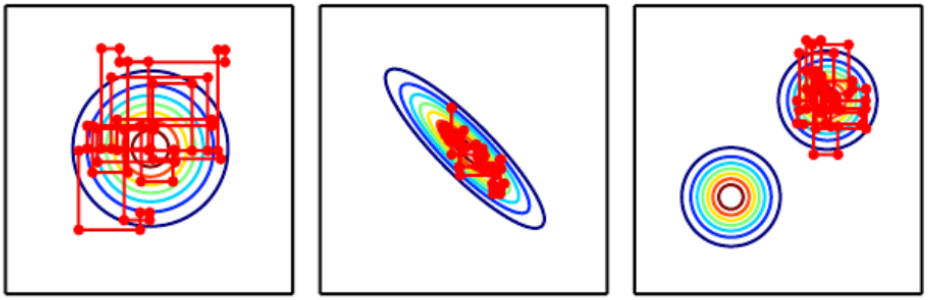

Figure 2: Paths followed by Gibbs sampling for three distributions, with the Markov chain initialized at the mode in both cases. (Left) A multivariate normal distribution with two independent variables. Gibbs sampling mixes well because the variables are independent. (Center) A multivariate normal distribution with highly correlated variables. The correlation between variables makes it difficult for the Markov chain to mix. Because the update for each variable must be conditioned on the other variable, the correlation reduces the rate at which the Markov chain can move away from the starting point. (Right) A mixture of Gaussians with widely separated modes that are not axis-aligned. Gibbs sampling mixes very slowly because it is difficult to change modes while altering only one variable at a time.

### 3.2 Problems of MCMC and the Remedy

One common problem of MCMC is that it converges poorly under tough conditions. Take Gibbs sampling as an example. When some dimensions $x_i$ and $x_j$ are highly correlated, the conditional probability $p(x_i|x_j)$ make it hard for $x_j$ to change when $x_i$ is fixed. In the extreme case, they are bounded in a deterministic way: $x_i = x_j$, and the Gibbs sampling to each of them will fail to update it's value, regardless of what their joint distribution is.

And even when the distribution converges, it is hard to tell from the statistics whether the distribution has already converged or "mixed". There have been extensive study on how to heuristically "test" whether a markov chain has mixed.

Another problem with a lot of iterative sampling method is the multimodality. For example, an unlabeled distribution of data $p(\mathbf{x})$ is generated by several class labels $y$:

$$p(\mathbf{x}) = \sum_y p(y)p(\mathbf{x}|y) \tag{22}$$

Suppose each class-conditional distribution $p(\mathbf{x}|y)$ is highly separate, then we will get big "probability gap" in the space of distribution. This will avoid algorithms like Gibbs sampling to transfer from one mode to the other. In this case, the Markov chain is very slow to mix and we can not get representative samples.

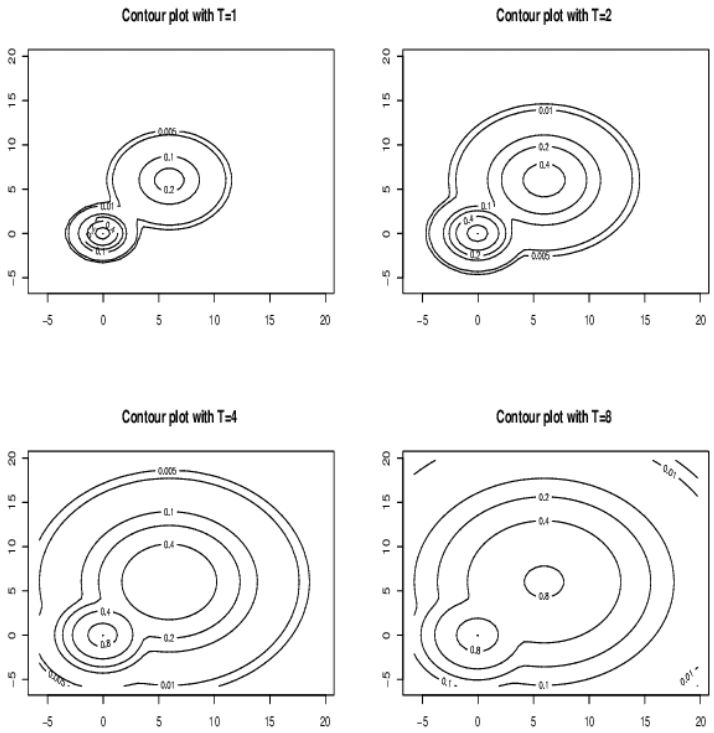


Figure 3: Tempering a multimodal distribution. As the temprature increases, the gap between different modes become less tough. The MCMC sampling in the last figure is much easier to come across the nodes than the first figure.

One technique to ease the multimodality of distribution for energy based model is to temper the distribution. Suppose we have a distribution with widely distributed modes:

$$p(\mathbf{x}) = \frac{1}{Z} exp(-E(\mathbf{x})) \tag{23}$$

We can add a temperature factor $\beta$ to the distribution:

$$p(\mathbf{x}) = \frac{1}{Z} exp(-\frac{1}{\beta} E(\mathbf{x})) \tag{24}$$

When the temperature increases, the distribution is "smoothed" so that the gap between modes are smaller, and our MCMC sampler will be easier to converge to the new distribution. However, the increased temperature will bring bias to the distribution. In the extreme case, when the temperature goes to infinity, the distribution will become a uniform distribution. Therefore, in order to get more accurate samples, we have to tune down the temperature as the Markov chain mixes.

## References


[1] BENGIO, Y., AND SENÉCAL, J.-S. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks 19*, 4 (2008), 713–722.

[2] HALTON, J. H. A retrospective and prospective survey of the monte carlo method. *Siam review 12*, 1 (1970), 1–63.

[3] KATHAROPOULOS, A., AND FLEURET, F. Biased importance sampling for deep neural network training. *arXiv preprint arXiv:1706.00043* (2017).