

CSC 412/2506:  
Probabilistic Learning and Reasoning  
Week 4 - 2/2: Sampling

Michal Malyska

University of Toronto

# Overview

- Ancestral Sampling
- Simple Monte Carlo
- Importance Sampling
- Rejection Sampling

# Sampling

- A sample from a distribution  $p(x)$  is a single realization  $x$  whose probability distribution is  $p(x)$ . Here,  $x$  can be high-dimensional or simply real valued.
- We assume the density from which we wish to draw samples,  $p(x)$ , can be evaluated to within a multiplicative constant. That is, we can evaluate a function  $\tilde{p}(x)$  such that

$$p(x) = \frac{\tilde{p}(x)}{Z}.$$

# Ancestral Sampling

Given a DAG, and the ability to sample from each of its factors given its parents, we can sample from the joint distribution over all the nodes by **ancestral sampling**, which simply means sampling in a topological order.

- at each step, sample from any conditional distribution that you haven't visited yet, whose parents have all been sampled.

Example: In a chain you would always start with  $z_1$  and move to the right. In a tree, you would always start from the root.

# Ancestral Sampling Example

# Main objectives of sampling

We will be using Monte Carlo methods to solve one or both of the following problems.

- **Problem 1:** To generate samples  $\{x^{(r)}\}_{r=1}^R$  from a given probability distribution  $p(x)$ .
- **Problem 2:** To estimate expectations of functions,  $\phi(x)$ , under this distribution  $p(x)$

$$\Phi = \mathbb{E}_{x \sim p(x)} [\phi(x)] = \int \phi(x)p(x)dx$$

$\phi$  is called a test function.

## Example

Examples of test functions  $\phi(x)$ :

- the mean of a function  $f$  under  $p(x)$  by finding the expectation of the function  $\phi_1(x) = f(x)$ .
- the variance of  $f$  under  $p(x)$  by finding the expectations of the functions  $\phi_1(x) = f(x)$  and  $\phi_2(x) = f(x)^2$

$$\phi_1(x) = f(x) \Rightarrow \Phi_1 = \mathbb{E}_{x \sim p(x)} [\phi_1(x)]$$

$$\phi_2(x) = f(x)^2 \Rightarrow \Phi_2 = \mathbb{E}_{x \sim p(x)} [\phi_2(x)]$$

$$\Rightarrow \text{var}(f(x)) = \Phi_2 - (\Phi_1)^2$$

# Estimation problem

We start with the estimation problem using simple Monte Carlo:

- **Simple Monte Carlo:** Given  $\{x^{(r)}\}_{r=1}^R \sim p(x)$  we can estimate the expectation  $\mathbb{E}_{x \sim p(x)} [\phi(x)]$  using the estimator  $\hat{\Phi}$ :

$$\Phi = \mathbb{E}_{x \sim p(x)} [\phi(x)] \approx \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) = \hat{\Phi}$$

- The fact that  $\hat{\Phi}$  is a consistent estimator of  $\Phi$  follows from the Law of Large Numbers (LLN).



# Basic properties of Monte Carlo estimation

- **Unbiasedness:** If the vectors  $\{x^{(r)}\}_{r=1}^R$  are generated independently from  $p(x)$ , then the expectation of  $\hat{\Phi}$  is  $\Phi$ .

$$\begin{aligned}\mathbb{E}[\hat{\Phi}] &= \mathbb{E}\left[\frac{1}{R} \sum_{r=1}^R \phi(x^{(r)})\right] = \frac{1}{R} \sum_{r=1}^R \mathbb{E}[\phi(x^{(r)})] \\ &= \frac{1}{R} \sum_{r=1}^R \mathbb{E}_{x \sim p(x)}[\phi(x)] = \frac{R}{R} \mathbb{E}_{x \sim p(x)}[\phi(x)] \\ &= \Phi\end{aligned}$$

## Simple properties of Monte Carlo estimation

- **Variance:** As the number of samples of  $R$  increases, the variance of  $\hat{\Phi}$  will decrease with rate  $\frac{1}{R}$

$$\begin{aligned}\text{var}[\hat{\Phi}] &= \text{var}\left[\frac{1}{R} \sum_{r=1}^R \phi(x^{(r)})\right] \\ &= \frac{1}{R^2} \text{var}\left[\sum_{r=1}^R \phi(x^{(r)})\right] \\ &= \frac{1}{R^2} \sum_{r=1}^R \text{var}\left[\phi(x^{(r)})\right] \\ &= \frac{R}{R^2} \text{var}[\phi(x)] \\ &= \frac{1}{R} \text{var}[\phi(x)]\end{aligned}$$

Accuracy of the Monte Carlo estimate depends on the variance of  $\phi$ .

# Sampling problem

- Assume we know the density  $p(x)$  up to a multiplicative constant

$$p(x) = \frac{\tilde{p}(x)}{Z}$$

- There are two difficulties:

- ▶ We do not generally know the normalizing constant,  $Z$ . The main difficulty is computing it

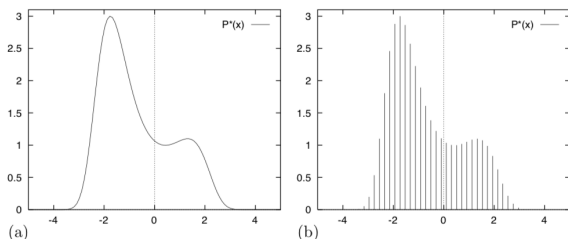
$$Z = \int \tilde{p}(x) dx$$

which requires computing a high-dimensional integral.

- ▶ Even if we did know  $Z$ , the problem of drawing samples from  $p(x)$  is still a challenging one, especially in high-dimensional spaces.

## Bad Idea: Lattice Discretization

Imagine that we wish to draw samples from the density  $p(x) = \frac{\tilde{p}(x)}{Z}$  given in figure (a).

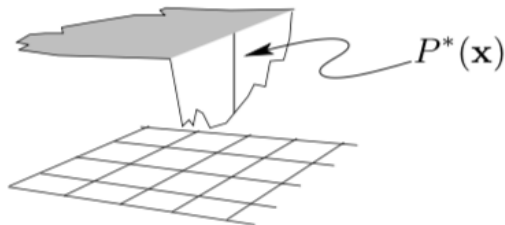


- How to compute  $Z$ ?
- We could discretize the variable  $x$  and sample from the discrete distribution (figure (b)).
- In figure (b) there are 50 uniformly spaced points in one dimension. If our system had,  $D = 1000$  dimensions say, then the corresponding number of points would be  $50^D = 50^{1000}$ . Thus, the cost is exponential in dimension!

## An analogy

Imagine the tasks of drawing random water samples from a lake and finding the average plankton concentration. Let

- $\tilde{p}(\mathbf{x})$  = the depth of the lake at  $\mathbf{x} = (x, y)$
- $\phi(\mathbf{x})$  = the plankton concentration as a function of  $\mathbf{x}$
- $Z$  = the volume of the lake =  $\int \tilde{p}(\mathbf{x}) d\mathbf{x}$



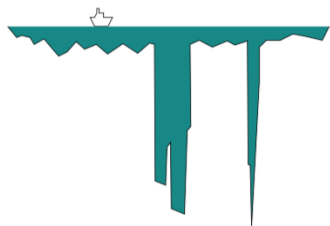
The average concentration of plankton is therefore

$$\Phi = \frac{1}{Z} \int \phi(\mathbf{x}) \tilde{p}(\mathbf{x}) d\mathbf{x}.$$

## An analogy

You can take the boat to any desired location  $\mathbf{x}$  on the lake, and can measure the depth,  $\tilde{p}(\mathbf{x})$ , and plankton concentration,  $\phi(\mathbf{x})$ , at that point. Therefore,

- **Problem 1** is to draw water samples at random such that each sample is equally likely to come from any point within the lake.
- **Problem 2** is to find the average plankton concentration.

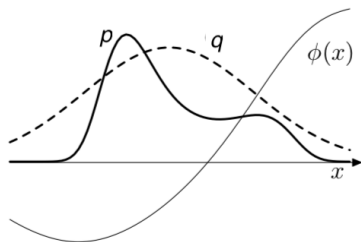


A slice through a lake that includes some canyons.

- We don't know the depth  $\tilde{p}(\mathbf{x})$ .
- To correctly estimate  $\Phi$ , our method must implicitly discover the canyons and find their volume relative to the rest of the lake.

# Estimation tool: Importance Sampling

**Importance sampling** is a method for estimating the expectation of a function  $\phi(x)$ .



- The density from which we wish to draw samples,  $p(x)$ , can be evaluated up to normalizing constant,  $\tilde{p}(x)$

$$p(x) = \frac{\tilde{p}(x)}{Z}$$

- There is a simpler density,  $q(x)$  from which it is easy to sample from and easy to evaluate up to normalizing constant (i.e.  $\tilde{q}(x)$ )

$$q(x) = \frac{\tilde{q}(x)}{Z_q}$$

## Estimation tool: Importance Sampling

- In importance sampling, we generate  $R$  samples from  $q(x)$

$$\{x^{(r)}\}_{r=1}^R \sim q(x)$$

- If these points were samples from  $p(x)$  then we could estimate  $\Phi$  by

$$\Phi = \mathbb{E}_{x \sim p(x)} [\phi(x)] \approx \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) = \hat{\Phi}$$

That is, we could use a simple Monte Carlo estimator.

- But we sampled from  $q$ . We need to correct this!
- Values of  $x$  where  $q(x)$  is greater than  $p(x)$  will be over-represented in this estimator, and points where  $q(x)$  is less than  $p(x)$  will be under-represented. Thus, we introduce weights.



- Introduce weights:  $\tilde{w}_r = \frac{\tilde{p}(x^{(r)})}{\tilde{q}(x^{(r)})}$  and notice that

$$\frac{1}{R} \sum_{r=1}^R \tilde{w}_r \approx \mathbb{E}_{x \sim q(x)} \left[ \frac{\tilde{p}(x)}{\tilde{q}(x)} \right] = \int \frac{\tilde{p}(x)}{\tilde{q}(x)} q(x) dx = \frac{Z_p}{Z_q}$$

- Finally, we rewrite our estimator under  $q$

$$\Phi = \int \phi(x) p(x) dx = \int \phi(x) \cdot \frac{p(x)}{q(x)} \cdot q(x) dx \approx \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) \frac{p(x^{(r)})}{q(x^{(r)})} = (*)$$

- However, the estimator relies on  $p$ . It can only rely on  $\tilde{p}$  and  $\tilde{q}$ .

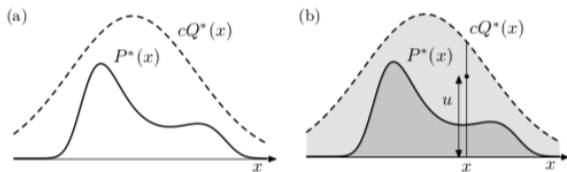
$$\begin{aligned} (*) &= \frac{Z_q}{Z_p} \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) \cdot \frac{\tilde{p}(x^{(r)})}{\tilde{q}(x^{(r)})} = \frac{Z_q}{Z_p} \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) \cdot \tilde{w}_r \\ &\approx \frac{\frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) \cdot \tilde{w}_r}{\frac{1}{R} \sum_{r=1}^R \tilde{w}_r} = \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) \cdot w_r = \hat{\Phi}_{iw} \end{aligned}$$

where  $w_r = \frac{\tilde{w}_r}{\sum_{r=1}^R \tilde{w}_r}$  and  $\hat{\Phi}_{iw}$  is our importance weighted estimator.

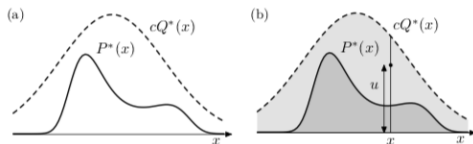
## Sampling tool: Rejection sampling

- We want expectations under  $p(x) = \tilde{p}(x)/Z$  which is a very complicated one-dimensional density.
- Assume that we have a simpler proposal density  $q(x)$  which we can evaluate (within a multiplicative factor  $Z_q$ , as before), and from which we can generate samples.
- Further assume that we know the value of a constant  $c$  such that

$$c\tilde{q}(x) > \tilde{p}(x) \quad \forall x$$



# Sampling tool: Rejection sampling



The procedure is as follows:

1. Generate two random numbers.
  - 1.1 The first,  $x$ , is generated from the proposal density  $q(x)$ .
  - 1.2 The second,  $u$  is generated uniformly from the interval  $[0, c\tilde{q}(x)]$  (see figure (b) above).
2. Evaluate  $\tilde{p}(x)$  and accept or reject the sample  $x$  by comparing the value of  $u$  with the value of  $\tilde{p}(x)$ 
  - 2.1 If  $u > \tilde{p}(x)$ , then  $x$  is rejected
  - 2.2 Otherwise  $x$  is accepted;  $x$  is added to our set of samples  $\{x^{(r)}\}$  and the value of  $u$  discarded.

# Why does rejection sampling work?

1.  $x \sim q(x)$
2.  $u|x \sim \text{Unif}[0, c\tilde{q}(x)]$
3.  $x$  is accepted if  $u \leq \tilde{p}(x)$ .

For any set  $A$

$$\mathbb{P}_{x \sim p}(x \in A) = \int_A p(x) dx = \int \mathbf{1}_{\{x \in A\}} p(x) dx = \mathbb{E}_{x \sim p}[\mathbf{1}_{\{x \in A\}}].$$

$$\begin{aligned} \mathbb{P}_{x \sim q}(x \in A | u \leq \tilde{p}(x)) &= \mathbb{P}_{x \sim q}(x \in A, u \leq \tilde{p}(x)) / \mathbb{E}_{x \sim q}[\mathbb{P}(u \leq \tilde{p}(x) | x)] \\ &= \mathbb{E}_{x \sim q}[\mathbf{1}_{\{x \in A\}} \mathbb{P}(u \leq \tilde{p}(x) | x)] / \mathbb{E}_{x \sim q}\left[\frac{\tilde{p}(x)}{c\tilde{q}(x)}\right] \\ &= \mathbb{E}_{x \sim q}\left[\mathbf{1}_{\{x \in A\}} \frac{\tilde{p}(x)}{c\tilde{q}(x)}\right] / \frac{Z_p}{cZ_q} \\ &= \mathbb{P}_{x \sim p}(x \in A) \frac{Z_p}{cZ_q} / \frac{Z_p}{cZ_q} \\ &= \mathbb{P}_{x \sim p}(x \in A) \end{aligned}$$

## Rejection sampling in many dimensions

- In high-dimensional problems, the requirement that  $c\tilde{q}(x) \geq \tilde{p}(x)$  will force  $c$  to be huge, so acceptances will be very rare.
- Finding such a value of  $c$  may be difficult too, since we don't know where the modes of  $\tilde{p}$  are located nor how high they are.
- In general  $c$  grows exponentially with the dimensionality, so the acceptance rate is expected to be exponentially small in dimension

$$\text{acceptance rate} = \frac{\text{area under } \tilde{p}}{\text{area under } c\tilde{q}} = \frac{1}{Z}$$

# Summary

- Estimating expectations is an important problem, which is in general hard. We learned 3 sampling-based tools for this task:
  - ▶ Simple Monte Carlo
  - ▶ Importance Sampling
  - ▶ Rejection Sampling
  - ▶ Ancestral Sampling
- Next lecture, we will learn to generate samples from a particular distribution.