CS 8501 Advanced Topics in Machine Learning

Lecture 09: Monte Carlo Methods

Yangfeng Ji

Information and Language Processing Lab

Department of Computer Science

University of Virginia

https://yangfengji.net/

Introduction

Posterior Distribution

Consider a latent variable model with X as the observed variable and Z as the latent variable

$$p(Z \mid X) = \frac{p(X \mid Z)p(Z)}{p(X)}$$

where

$$p(X) = \int p(X|Z)p(Z)dz$$

Problem Setup

Target density: in the rest of this lecture, we will consider a generic distribution

$$p(X)$$

as the target distribution that we would like to study

In general, we assume X is an N-dimensional random vector, aka $X \in \mathbb{R}^N$

Expectation

Consider a function $f(x)$ of x, its expectation is defined as

$$E[f(X)] = \int f(x)p(x)dx$$

With a simple example $f(x) = x$, the approximated expectation

$$E[X] \approx \frac{1}{N}\sum_{n=1}^{N} x_n$$

Expectation (II)

With a generic function $f(x)$, the approximated expectation

$$E[f(X)] = \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

Therefore, the major challenge is to get a collection of samples

$$\{x_n\}_{n=1}^{N}$$

that are drawn from $p(X)$

Two Central Questions

- **Problem 1**: How to generate samples $\{x_n\}_{n=1}^{R}$ from a given probability distribution $p(X)$
  - Sampling from simple distributions
  - Reject-Accept sampling
  - Importance sampling
- **Problem 2**: How to measure the quality of approximation

$$E[f(X)] = \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

  - Variance reduction

Monte Carlo Integration

Integration

Many problem in statistical estimation is related to integration

- marginalization:

$$p(x) = \int p(x, z) dz$$

- expectation:

$$E[f(x)] = \int f(x) p(x) dx$$

Integration via Sampling

Assume we have a set of N samples from $p(x)$, $\{x_n\}_{n=1}^{N} \sim p(x)$, the expectation can be approximated as

$$E[f(x)] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Justification

- Approximate $p(x)$ with an empirical distribution

$$\hat{p}(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x = x_n)$$

- Substitute $p(x)$ with $\hat{p}(x)$ in the expectation definition

$$E[f(x)] \approx \int f(x) \cdot \frac{1}{N} \sum_{n=1}^{N} \delta(x = x_n) dx = \frac{1}{N} \sum_{n=1}^{N} \int f(x) \delta(x = x_n) dx$$

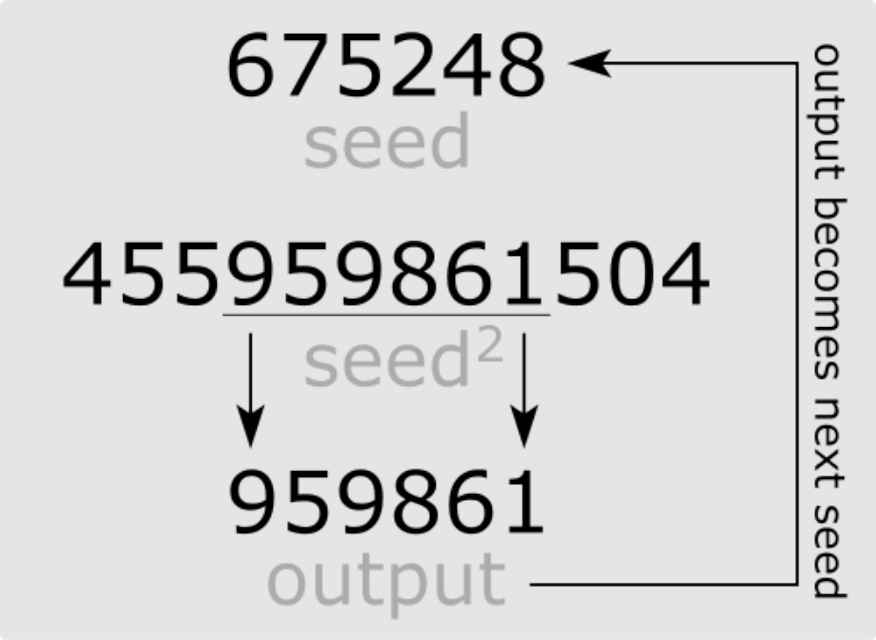
Sampling from Simple Distributions

Some Special Cases of Sampling

1. Sampling from $U(0, 1)$
2. Sampling using the inverse CDF
3. Sampling from a Gaussian

Pseudo-random Number Generator

John von Neumann's Middle-square method



- Start to repeat a number in the previous sequence quickly
- Can be used to generate **uniform distributions**

Cumulative Distribution Function (CDF)

For a (either discrete or continuous) random variable X, CDF is defined as

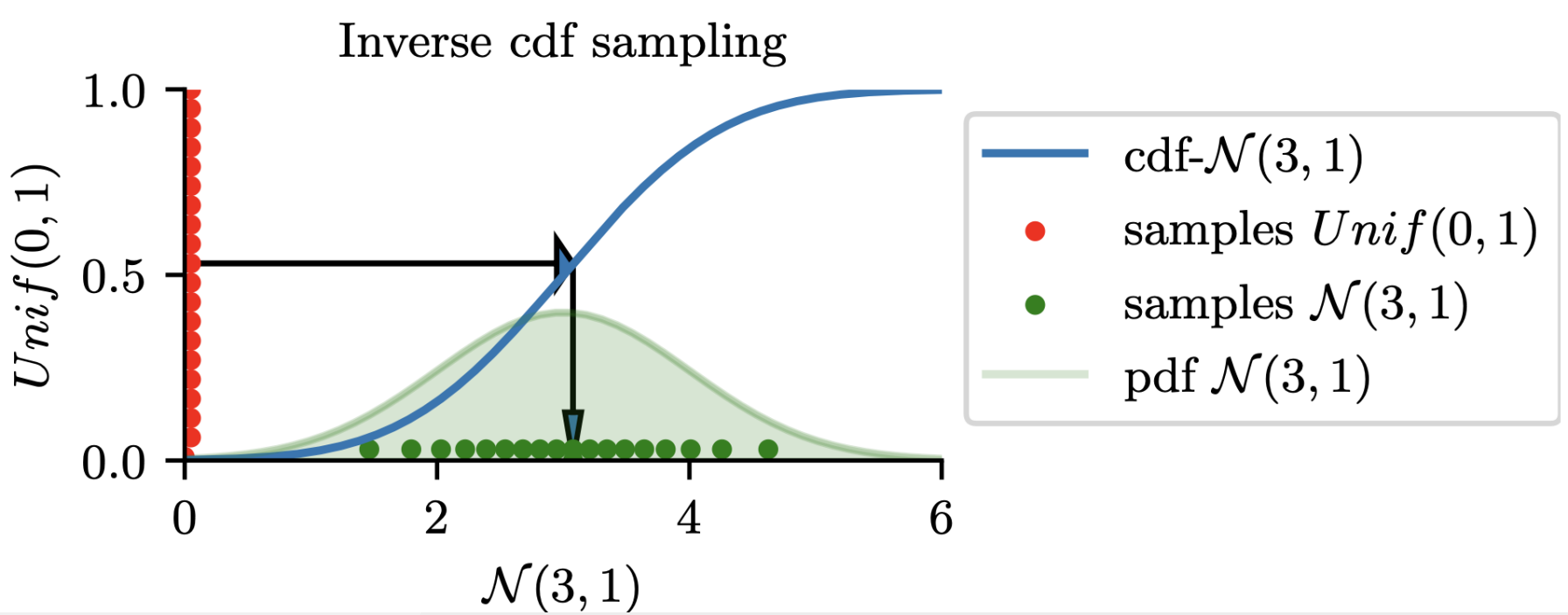
$$F(a) = \int_{x \leq a} p(X)dx = p(X \leq a)$$

A few examples of F with respect to $\mathcal{N}(x; 0, 1)$

- $F(0) = 0.5$
- $F(\infty) = 1.0$

Sampling with Inverse CDF

Theorem If $u \sim U(0,1)$ is a uniform random variable, then $F^{-1}(u)$ is a random variable following the distribution with F as its CDF



Accept-Reject Sampling

Two other names:

- Rejection sampling
- Acceptance-rejection sampling

Starting Point

Define the target distribution as

$$p(x) = \frac{1}{Z_p}\tilde{p}(x)$$

where $\tilde{p}$ is the unnormalized distribution with

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

Basic Idea

- Choose a proposal distribution $q(x)$, such that

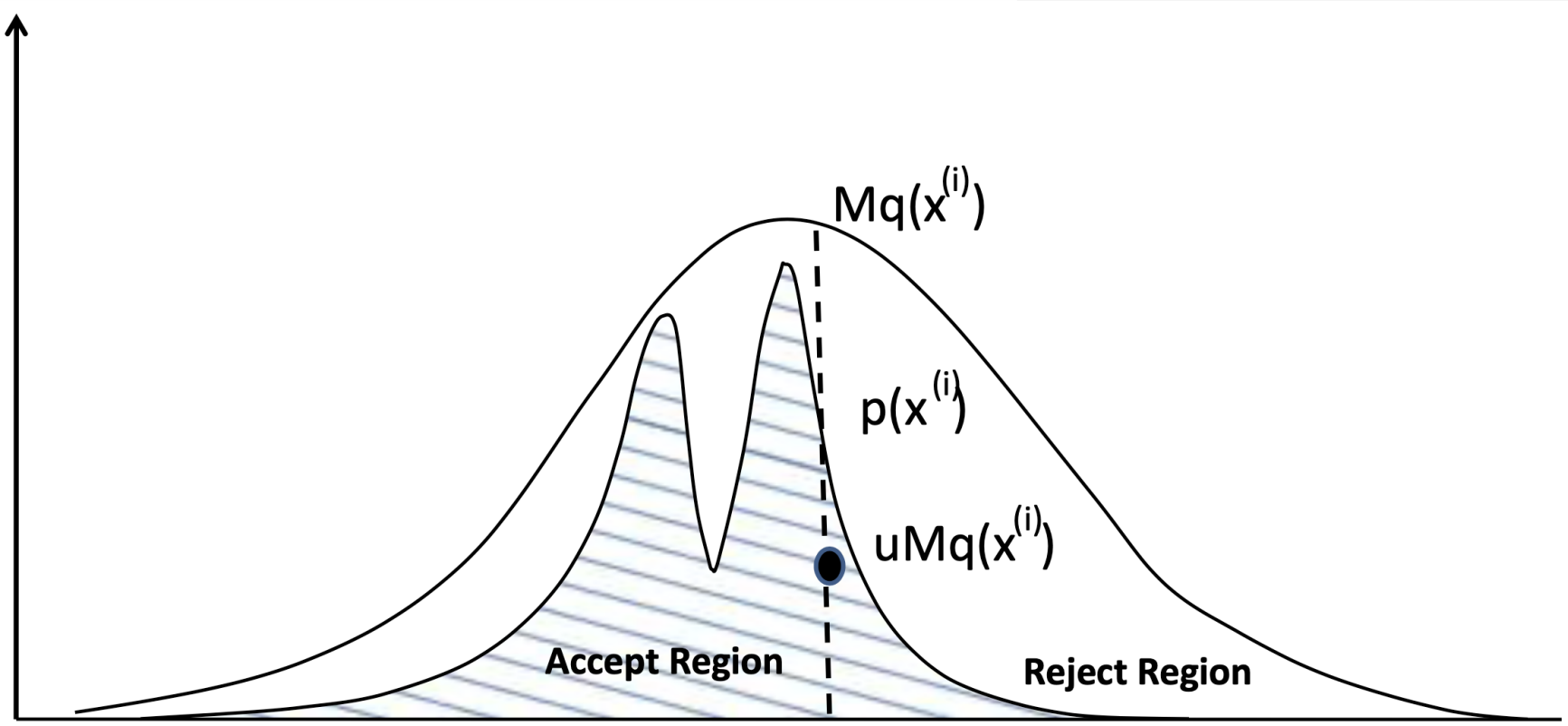
$$Cq(x) \geq \tilde{p}(x)$$

  where C is a constant, and $Cq(x)$ gives an upper envelope for $\tilde{p}$

- Sampling procedure
  - Sample $x_0 \sim q(x)$
  - Sample $u_0 \sim U(0, Cq(x_0)$
  - If $u_0 > \tilde{p}(x_0)$, accept x_0 as a sample from $p(x)$; otherwise, reject it

Why It Works?

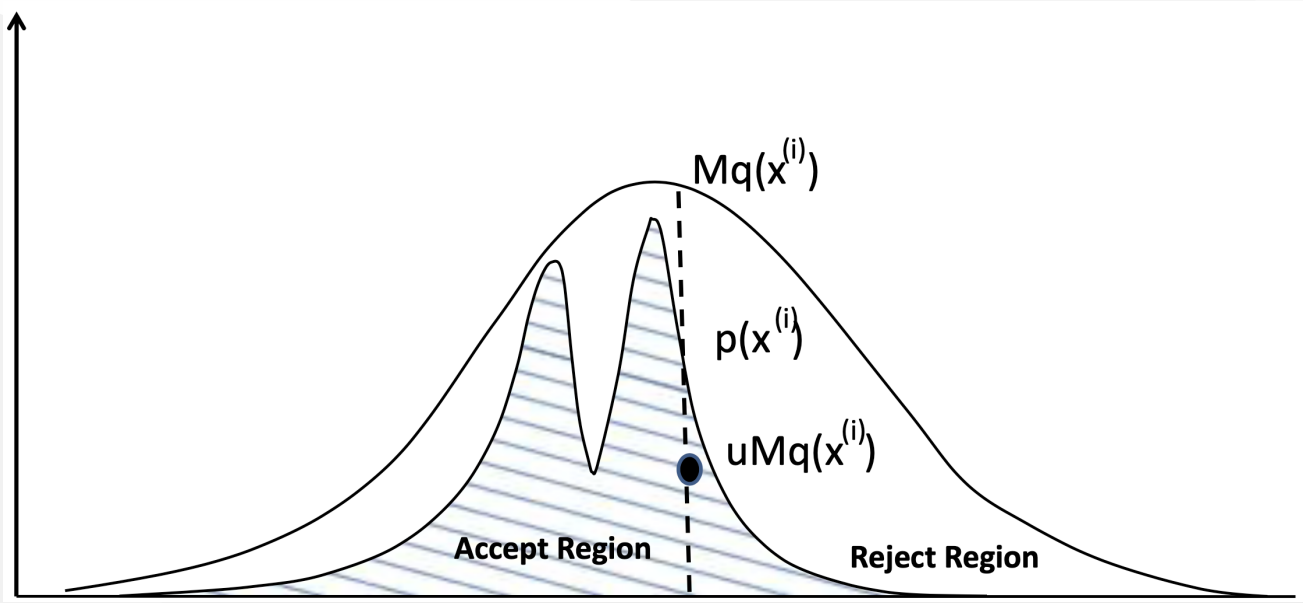
In the following plot, $M = C$



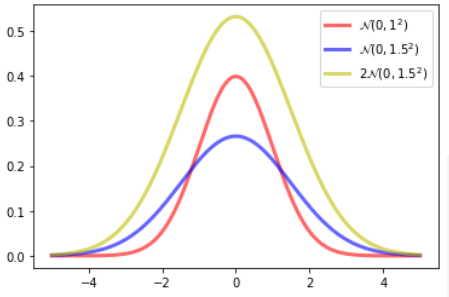
The proof can be found in section 11.4.1

Issues of Rejection Sampling

- The shape similarity between $q(x)$ and $\tilde{p}(x)$
  - Adaptive rejection sampling
- The acceptence rate
  - A fundamental weakness



Acceptance Rate: 1 dimension



- Target distribution: $p(x) = \frac{1}{\sqrt{2\pi}\sigma_p} \exp(-\frac{x^2}{2\sigma_p^2})$
- Proposal distribution: $q(x) = \frac{1}{\sqrt{2\pi}\sigma_q} \exp(-\frac{x^2}{2\sigma_q^2})$
- To make sure the proposal distribution be a good envelope, we need C to be at least

$$C = \frac{p(0)}{q(0)} = \frac{\sigma_q}{\sigma_p}$$

Acceptance Rate: D dimensions

- Target distribution: $p(x) = (2\pi)^{-\frac{D}{2}} \det(\Sigma_p)^{-\frac{1}{2}} \exp(-\frac{1}{2} x^\top \Sigma_p^{-1} x)$
- Proposal distribution: $q(x) = (2\pi)^{-\frac{D}{2}} \det(\Sigma_q)^{-\frac{1}{2}} \exp(-\frac{1}{2} x^\top \Sigma_q^{-1} x)$
- Therefore, C should at least be

$$C = \frac{p(0)}{q(0)} = \frac{\det(\Sigma_q)^{\frac{1}{2}}}{\det(\Sigma_p)^{\frac{1}{2}}} = \frac{\sigma_q^D}{\sigma_p^D} = (\frac{\sigma_q}{\sigma_p})^D$$

Because $\det(\Sigma) = \det(\sigma^2 I) = \sigma^{2D}$

Sampling from High Dimensions

This is a fundamental challenge of all sampling methods

- The methods discussed in this lecture is mostly applicable to low dimensions
- Next lecture will discuss the sampling methods for high-dimensional space

Importance Sampling

Problem Setup

Consider the following **integral** problem

$$E[f(x)] = \int f(x)p(x)dx$$

Recall the previous discussion, if we can sample from $p(x)$ directly

$$E[f(x)] = \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

Sampling from a Proposal Distribution

Now, if we sample from a proposal distribution $q(x)$ and approximate the expectation as

$$E[f(x)] = \frac{1}{N}\sum_{n=1}^{N} w_n f(x_n)$$

What w_n could be?

$$w_n = \frac{p(x_n)}{q(x_n)}$$

Note: $p(x)$ needs to be a normalized probability distribution in this case

Justification

Consider the expectation

$$E[f(x)] = \int f(x)p(x)dx = \int q(x)[\frac{p(x)}{q(x)}]f(x)dx$$

Sampling $\{x_n\} \sim q(x)$, we have

$$E[f(x)] \approx \frac{1}{N}\sum_{n=1}^{N}\frac{p(x_n)}{q(x_n)}f(x_n)$$

What about unnormalized $\tilde{p}(X)$?

Consider an unnormalized target distribution $\tilde{p}(x)$, we define the unnormalized weight as

$$\tilde{w}_n = \frac{\tilde{p}(x)}{q(x)}$$

With the normalization constant is $Z_p = \int \tilde{p}(x)dx$, we have $w_n = \frac{\tilde{w}_n}{Z_p}$

Note that

$$Z_p = \int \tilde{p}(x)dx = \int [\frac{\tilde{p}(x)}{q(x)}]q(x)dx \approx \frac{1}{N}\sum_{n=1}^{N}\frac{\tilde{p}(x_n)}{q(x_n)} = \frac{1}{N}\sum_{n=1}^{N}\tilde{w}_n$$

Therefore

With unnormalized $\tilde{p}(x)$, we have

$$w_n \approx \frac{\tilde{w}_n}{\frac{1}{N}\sum_{n=1}^{N} \tilde{w}_n}$$

and

$$E[f(x)] \approx \frac{\frac{1}{N}\sum_{n=1}^{N} \tilde{w}_n f(x_n)}{\frac{1}{N}\sum_{n=1}^{N} \tilde{w}_n} = \frac{\sum_{n=1}^{N} \tilde{w}_n f(x_n)}{\sum_{n=1}^{N} \tilde{w}_n}$$

Controlling Variance

Rao-Blackwellisation

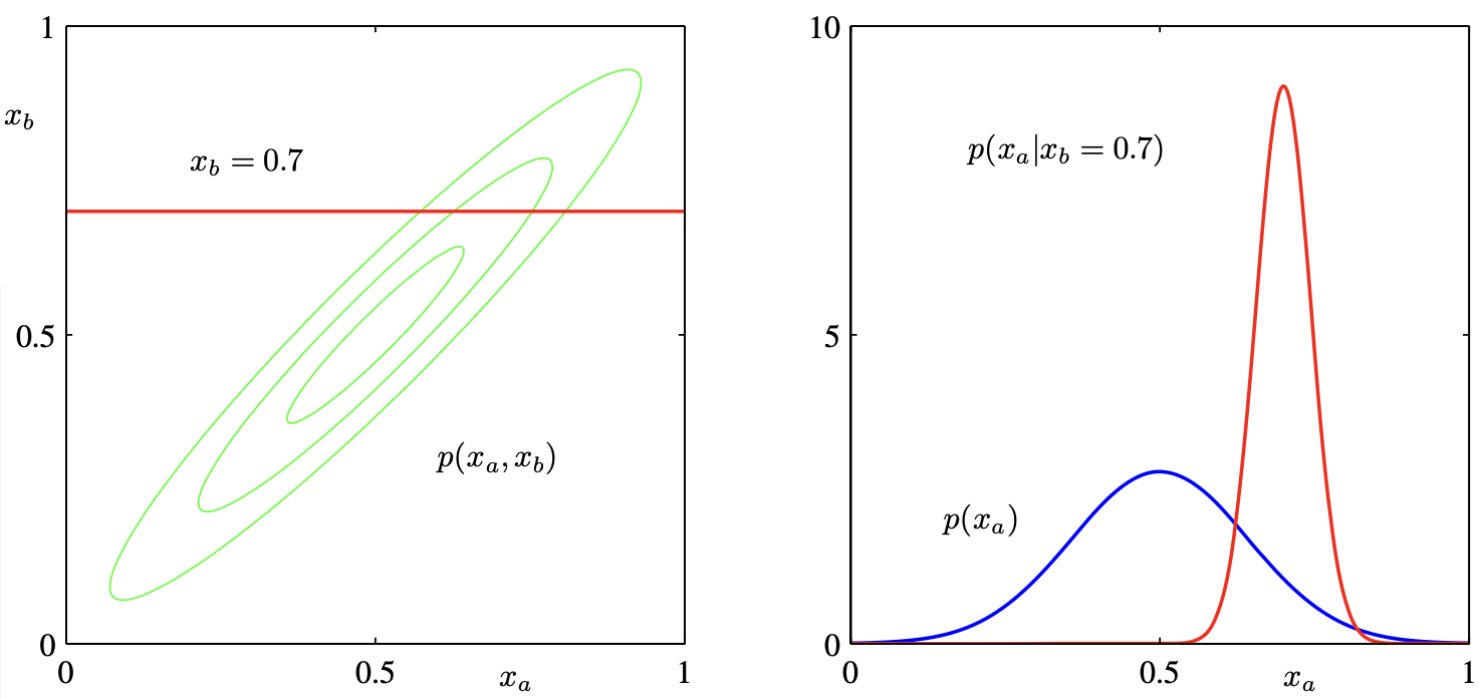
For $E[f(X,Y)]$, instead of sampling on both random variables directly, we can do

- Marginalize Y: $f(X) = \int p(Y|X) f(X,Y) dy = E[f(X,Y)|X]$
- Sampling X from $p(X)$

$$E[f(X,Y)] = \int E[f(X,Y)|X)p(X)]dx = \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

Rao-Blackwellisation (II)

Why it reduce the variance



Essentially, it reduce the variance by sampling from low-dimensional space.

Thank You!