Sampling & Monte Carlo

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

February 20, 2023

Previously on Probabilistic Inference

We looked at **logistic regression**

- Different kind of data (binary classification)
- Different assumptions in model

Previously on Probabilistic Inference

We looked at **logistic regression**

- Different kind of data (binary classification)
- Different assumptions in model

We wanted to

- Make predictions
- Find posterior

Previously on Probabilistic Inference

We looked at **logistic regression**

- Different kind of data (binary classification)
- Different assumptions in model

We wanted to

- Make predictions
- Find posterior

Both computations were **intractable**.

Overview

Monte Carlo Estimation

Monte Carlo with Exact Sampling

Numerical Quadrature

Intractable computation are caused by **integrals**.

$$p(y^* \mid x^*, \mathbf{y}, X) = \int p(y^* \mid x^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}, X) \mathrm{d}\boldsymbol{\theta} \tag{1}$$

$$p(\boldsymbol{\theta} \mid \mathbf{y}, X) = \frac{p(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y} \mid X)} = \frac{p(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}} \tag{2}$$

Numerical Quadrature

Intractable computation are caused by **integrals**.

$$p(y^* \mid x^*, \mathbf{y}, X) = \int p(y^* \mid x^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}, X) \mathrm{d}\boldsymbol{\theta} \tag{1}$$

$$p(\boldsymbol{\theta} \mid \mathbf{y}, X) = \frac{p(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y} \mid X)} = \frac{p(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}} \tag{2}$$

Can we approximate numerically?

Numerical Quadrature

Intractable computation are caused by **integrals**.

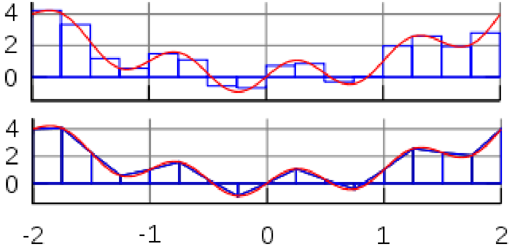
$$p(y^* \mid x^*, \mathbf{y}, X) = \int p(y^* \mid x^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}, X) \mathrm{d}\boldsymbol{\theta} \tag{1}$$

$$p(\boldsymbol{\theta} \mid \mathbf{y}, X) = \frac{p(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y} \mid X)} = \frac{p(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}} \tag{2}$$

Can we approximate numerically? Evaluate on a **grid**.

Rectangle rule

Trapezoidal rule



Numerical Quadrature in High Dimensions

We may have many parameters! For linear / logistic regression:

- $\boldsymbol{\theta} \in \mathbb{R}^D$
- Even more if we use basis functions!
- It is very common to have > 100 parameters

Numerical Quadrature in High Dimensions

We may have many parameters! For linear / logistic regression:

- $\boldsymbol{\theta} \in \mathbb{R}^D$
- Even more if we use basis functions!
- It is very common to have > 100 parameters

For D dimensions, there are P^D total points in the grid. For $P = 10$, $D = 100$, that is more than the number of atoms in the universe.

Numerical Quadrature in High Dimensions

We may have many parameters! For linear / logistic regression:

- $\boldsymbol{\theta} \in \mathbb{R}^D$
- Even more if we use basis functions!
- It is very common to have > 100 parameters

For D dimensions, there are P^D total points in the grid. For $P = 10$, $D = 100$, that is more than the number of atoms in the universe.

- **Rate** of convergence depends on dimension (e.g. $O(P_{\text{total}}^{-\frac{1}{D}})$ for rectangle rule)
- Need exponential number of points with dimension

Numerical Quadrature in High Dimensions

We may have many parameters! For linear / logistic regression:

- $\boldsymbol{\theta} \in \mathbb{R}^D$
- Even more if we use basis functions!
- It is very common to have > 100 parameters

For D dimensions, there are P^D total points in the grid. For $P = 10$, $D = 100$, that is more than the number of atoms in the universe.

- **Rate** of convergence depends on dimension (e.g. $O(P_{\text{total}}^{-\frac{1}{D}})$ for rectangle rule)
- Need exponential number of points with dimension
  to reduce error by a factor of 2, you need $P_2/P_1 = 2^D$

Numerical Quadrature in High Dimensions

We may have many parameters! For linear / logistic regression:

- $\boldsymbol{\theta} \in \mathbb{R}^D$
- Even more if we use basis functions!
- It is very common to have > 100 parameters

For D dimensions, there are P^D total points in the grid. For $P = 10$, $D = 100$, that is more than the number of atoms in the universe.

- **Rate** of convergence depends on dimension (e.g. $O(P_{\text{total}}^{-\frac{1}{D}})$ for rectangle rule)
- Need exponential number of points with dimension to reduce error by a factor of 2, you need $P_2/P_1 = 2^D$

Curse of Dimensionality

Monte Carlo Approximation

Most Bayesian computations are in fact **expectations**
E.g. prediction for logistic regression

$$p(y^* \mid x^*, \mathbf{y}, X) = \int p(y^* \mid x^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}, X) \mathrm{d}\boldsymbol{\theta} \tag{3}$$

$$= \mathbb{E}_{p(\boldsymbol{\theta} \mid \mathbf{y}, X)}[p(y^* \mid x^*, \boldsymbol{\theta})]\,. \tag{4}$$

Monte Carlo Approximation

Most Bayesian computations are in fact **expectations**
E.g. prediction for logistic regression

$$p(y^* \mid x^*, \mathbf{y}, X) = \int p(y^* \mid x^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}, X) \mathrm{d}\boldsymbol{\theta} \tag{3}$$

$$= \mathbb{E}_{p(\boldsymbol{\theta} \mid \mathbf{y}, X)}[p(y^* \mid x^*, \boldsymbol{\theta})] \,. \tag{4}$$

In general,

$$I = \mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] \tag{5}$$

$$\implies I \approx \hat{I} = \frac{1}{S} \sum_{s=1}^{S} g(\mathbf{x}^{(s)}) \,, \qquad \text{with } \mathbf{x}^{(s)} \overset{\text{iid}}{\sim} p(\mathbf{x}) \,. \tag{6}$$

Monte Carlo Properties

Monte Carlo estimator

- mean is equal to the quantity we want to estimate (**unbiased**)

$$\mathbb{E}_{p(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\dots)}[\hat{I}] = \int \prod_{t=1}^{S} p(\mathbf{x}^{(t)}) \frac{1}{S} \sum_{s=1}^{S} g(\mathbf{x}^{(s)}) \mathrm{d}\{\mathbf{x}^{(u)}\}_{u=1}^{S} = I \qquad (7)$$

(Bring sum outside, distributions for $s \neq t$ integrate to 1)

Monte Carlo Properties

Monte Carlo estimator

- mean is equal to the quantity we want to estimate (**unbiased**)

$$\mathbb{E}_{p(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\dots)}[\hat{I}] = \int \prod_{t=1}^{S} p(\mathbf{x}^{(t)}) \frac{1}{S} \sum_{s=1}^{S} g(\mathbf{x}^{(s)}) \mathrm{d}\{\mathbf{x}^{(u)}\}_{u=1}^{S} = I \tag{7}$$

(Bring sum outside, distributions for $s \neq t$ integrate to 1)

- variance decreases **independent of dimension**

$$\mathbb{V}_{p(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\dots)}[\hat{I}] = \frac{1}{S^2} \sum_{s=1}^{S} \mathbb{V}_{p(\mathbf{x})}[g(\mathbf{x})] = \frac{C}{S} \tag{8}$$

i.e. error decreases as $O(\frac{1}{\sqrt{S}})$.

Overview

Monte Carlo Estimation

Monte Carlo with Exact Sampling

How to generate samples

When specifying a Monte Carlo approximation, you need a procedure for **generating samples** from your distribution of interest $p(\mathbf{x})$.

- Some distributions are easy to sample from (e.g. Uniform, Standard Gaussian). You may assume that such samples are available in the exam.
- Often though, no direct procedure for sampling $p(\mathbf{x})$

How to generate samples

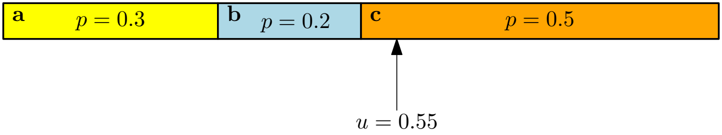
When specifying a Monte Carlo approximation, you need a procedure for **generating samples** from your distribution of interest $p(\mathbf{x})$.

- Some distributions are easy to sample from (e.g. Uniform, Standard Gaussian). You may assume that such samples are available in the exam.
- Often though, no direct procedure for sampling $p(\mathbf{x})$

Different procedures are have different sampling properties. Distributions can be

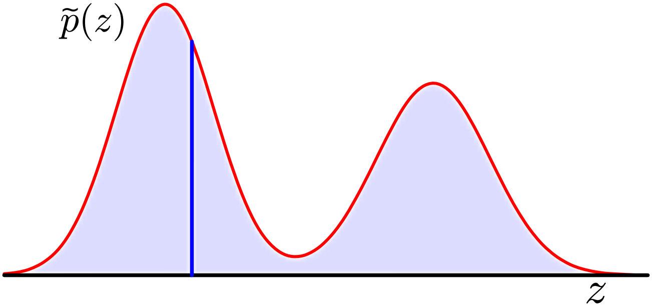
- easy to sample, hard to evaluate (GANs, VAEs),
- easy to evaluate, hard to sample.

Sampling Discrete Variables



- $u \sim \mathcal{U}[0, 1]$, where $\mathcal{U}$ is the uniform distribution
- $u = 0.55 \Rightarrow x = c$

Continuous Variables

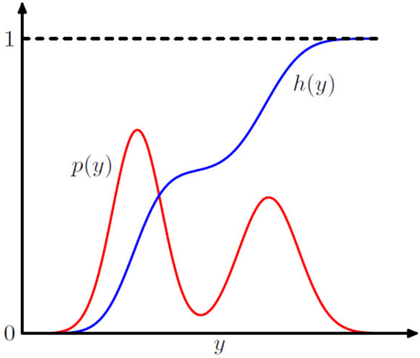


$$P(z_1 < Z < z_2) = \int_{z_1}^{z_2} p(z) \mathrm{d}z \tag{9}$$

Geometric intuition: sample uniformly from the area under the curve

Sampling Continuous Values

Let's convert samples from $\mathcal{U}[0, 1]$ to samples from densities



From Bishop: PRML (2006)

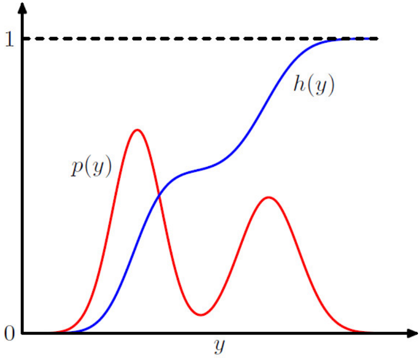
Objective: Sample from $p(y)$.

- $h(y) = \int_{-\infty}^{y} p(z)dz$ (CDF)
- Draw $u \sim \mathcal{U}[0, 1]$
- Obtain sample from $p(y)$: $y(u) = h^{-1}(u)$

▶▶ Inverse Transform Sampling

Sampling Continuous Values

Let's convert samples from $\mathcal{U}[0, 1]$ to samples from densities



From Bishop: PRML (2006)

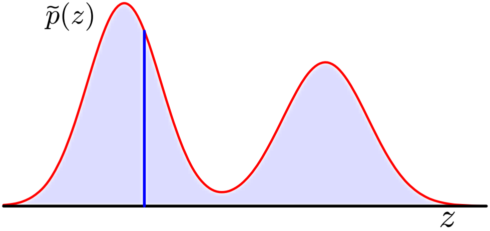
Objective: Sample from $p(y)$.

- $h(y) = \int_{-\infty}^{y} p(z)dz$ (CDF)
- Draw $u \sim \mathcal{U}[0, 1]$
- Obtain sample from $p(y)$: $y(u) = h^{-1}(u)$

▶▶ Inverse Transform Sampling

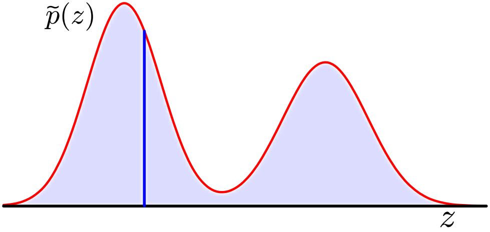
- We cannot always invert the CDF $h(y)$
- Difficult for high-dimensional distributions

Rejection Sampling: Setting



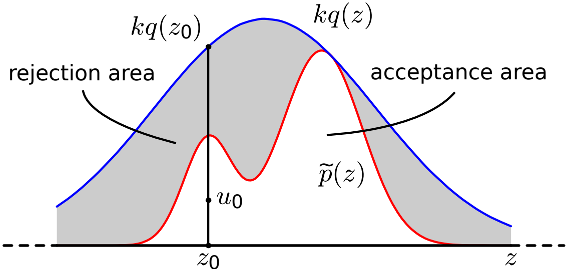
- Assume:
  - Sampling from $p(z)$ is difficult
  - Evaluating $\tilde{p}(z) = Zp(z)$ is easy (and Z may be unknown)

Rejection Sampling: Setting



- Assume:
  - Sampling from $p(z)$ is difficult
  - Evaluating $\tilde{p}(z) = Zp(z)$ is easy (and Z may be unknown)
- Find a simpler distribution (proposal distribution) $q(z)$ from which we can easily draw samples (e.g., Gaussian, Uniform)
- Find an upper bound $kq(z) \geqslant \tilde{p}(z)$

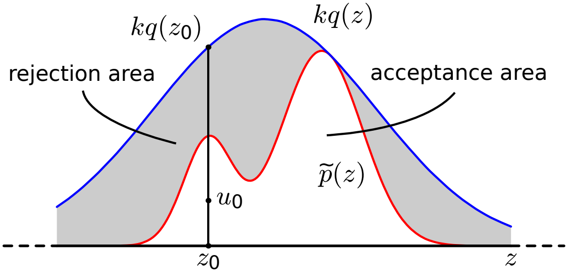
Rejection Sampling: Algorithm



Adapted from PRML (Bishop, 2006)

1. Generate $z_0 \sim q(z)$
2. Generate $u_0 \sim \mathcal{U}[0, kq(z_0)]$
3. If $u_0 > \tilde{p}(z_0)$, reject the sample. Otherwise, retain z_0

Properties



Adapted from PRML (Bishop, 2006)

- Accepted pairs (z, u) are uniformly distributed under the curve of $\tilde{p}(z)$
- Marginal probability density of the z-coordiantes of accepted points must be proportional to $\tilde{p}(z)$
- Samples are independent samples from $p(z)$

Sampling in High Dimensions

Example:

- $p(\boldsymbol{x}) = \mathcal{N}\left(\mathbf{0}, \sigma_p^2 \boldsymbol{I}\right), \quad q(\boldsymbol{x}) = \mathcal{N}\left(\mathbf{0}, \sigma_q^2 \boldsymbol{I}\right)$ where $\sigma_q = 1.01\sigma_p$
- What is the value of k if $\boldsymbol{x} \in \mathbb{R}^{1000}$?

Sampling in High Dimensions

Example:

- $p(\boldsymbol{x}) = \mathcal{N}\left(\mathbf{0}, \sigma_p^2 \boldsymbol{I}\right), \quad q(\boldsymbol{x}) = \mathcal{N}\left(\mathbf{0}, \sigma_q^2 \boldsymbol{I}\right)$ where $\sigma_q = 1.01\sigma_p$
- What is the value of k if $\boldsymbol{x} \in \mathbb{R}^{1000}$?
- $q(0) = 1/(2\pi\sigma_q^2)^{500}$ ⏩ For $kq \geqslant p$ we need to set

$$k \geqslant \frac{p(0)}{q(0)} = \frac{(\sigma_q^2)^{500}}{(\sigma_p^2)^{500}} = \exp\left(1000 \ln \frac{\sigma_q}{\sigma_p}\right) = \exp(1000 \ln 1.01) \approx 20,000$$

Sampling in High Dimensions

Example:

- $p(\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{0}, \sigma_p^2 \boldsymbol{I}\big), \quad q(\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{0}, \sigma_q^2 \boldsymbol{I}\big)$ where $\sigma_q = 1.01\sigma_p$
- What is the value of k if $\boldsymbol{x} \in \mathbb{R}^{1000}$?
- $q(0) = 1/(2\pi\sigma_q^2)^{500}$ ⏩ For $kq \geqslant p$ we need to set

$$k \geqslant \frac{p(0)}{q(0)} = \frac{(\sigma_q^2)^{500}}{(\sigma_p^2)^{500}} = \exp\left(1000 \ln \frac{\sigma_q}{\sigma_p}\right) = \exp(1000 \ln 1.01) \approx 20,000$$

- Acceptance rate is the ratio of the volume under p to the volume under kq. In our example: $1/k = 1/20,000$.

Sampling in High Dimensions

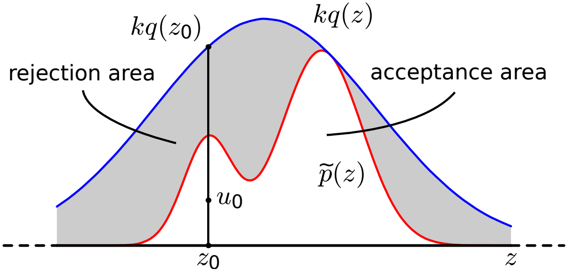
Example:

- $p(\boldsymbol{x}) = \mathcal{N}\left(\mathbf{0}, \sigma_p^2 \boldsymbol{I}\right), \quad q(\boldsymbol{x}) = \mathcal{N}\left(\mathbf{0}, \sigma_q^2 \boldsymbol{I}\right)$ where $\sigma_q = 1.01\sigma_p$
- What is the value of k if $\boldsymbol{x} \in \mathbb{R}^{1000}$?
- $q(0) = 1/(2\pi\sigma_q^2)^{500}$ ⏩ For $kq \geqslant p$ we need to set

$$k \geqslant \frac{p(0)}{q(0)} = \frac{(\sigma_q^2)^{500}}{(\sigma_p^2)^{500}} = \exp\left(1000 \ln \frac{\sigma_q}{\sigma_p}\right) = \exp(1000 \ln 1.01) \approx 20,000$$

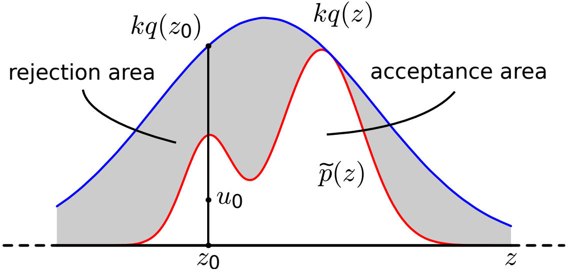
- Acceptance rate is the ratio of the volume under p to the volume under kq. In our example: $1/k = 1/20,000$.
- In high dimensions the factor k is probably huge
  ⏩ **Low acceptance rate**
- Finding k is tricky

Shortcomings



Adapted from PRML (Bishop, 2006)

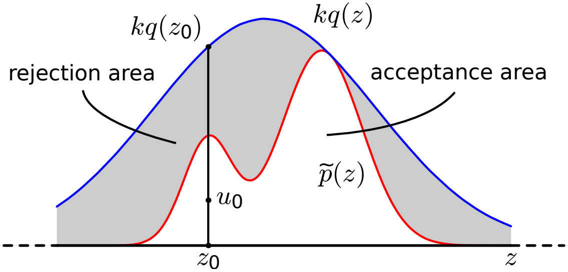
Shortcomings



Adapted from PRML (Bishop, 2006)

- Finding the upper bound k is tricky

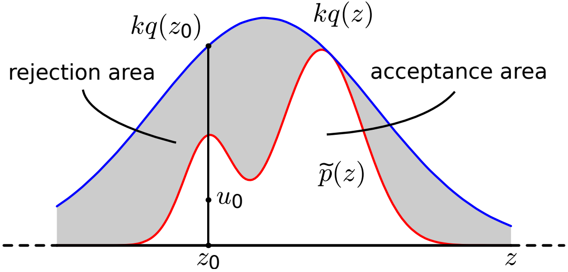
Shortcomings



Adapted from PRML (Bishop, 2006)

- Finding the upper bound k is tricky
- In high dimensions the factor k is probably huge

Shortcomings



Adapted from PRML (Bishop, 2006)

- Finding the upper bound k is tricky
- In high dimensions the factor k is probably huge
- **Low acceptance rate/high rejection rate** of samples

Importance Sampling

Key idea: Do not throw away all rejected samples, but give them lower weight by rewriting the integral as an expectation under a simpler distribution q (proposal distribution):

$$\mathbb{E}_p[f(\boldsymbol{x})] = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$$

Importance Sampling

Key idea: Do not throw away all rejected samples, but give them lower weight by rewriting the integral as an expectation under a simpler distribution q (proposal distribution):

$$\begin{aligned}\mathbb{E}_p[f(\boldsymbol{x})] &= \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\ &= \int f(\boldsymbol{x})p(\boldsymbol{x})\frac{q(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x}\end{aligned}$$

Importance Sampling

Key idea: Do not throw away all rejected samples, but give them lower weight by rewriting the integral as an expectation under a simpler distribution q (proposal distribution):

$$\begin{aligned}\mathbb{E}_p[f(\boldsymbol{x})] &= \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\ &= \int f(\boldsymbol{x})p(\boldsymbol{x})\frac{q(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} = \int f(x)\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x}\end{aligned}$$

Importance Sampling

Key idea: Do not throw away all rejected samples, but give them lower weight by rewriting the integral as an expectation under a simpler distribution q (proposal distribution):

$$\begin{aligned}
\mathbb{E}_p[f(\boldsymbol{x})] &= \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\
&= \int f(\boldsymbol{x})p(\boldsymbol{x})\frac{q(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} = \int f(x)\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x} \\
&= \mathbb{E}_q\left[f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right]
\end{aligned}$$

Importance Sampling

Key idea: Do not throw away all rejected samples, but give them lower weight by rewriting the integral as an expectation under a simpler distribution q (proposal distribution):

$$\begin{aligned}\mathbb{E}_p[f(\boldsymbol{x})] &= \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\ &= \int f(\boldsymbol{x})p(\boldsymbol{x})\frac{q(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} = \int f(x)\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x} \\ &= \mathbb{E}_q\left[f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right]\end{aligned}$$

If we choose q in a way that we can easily sample from it, we can approximate this last expectation by Monte Carlo:

$$\mathbb{E}_q\left[f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right] \approx \frac{1}{S}\sum_{s=1}^{S} f(\boldsymbol{x}^{(s)})\frac{p(\boldsymbol{x}^{(s)})}{q(\boldsymbol{x}^{(s)})} \quad , \quad \boldsymbol{x}^{(s)} \sim q(\boldsymbol{x})$$

Importance Sampling

Key idea: Do not throw away all rejected samples, but give them lower weight by rewriting the integral as an expectation under a simpler distribution q (proposal distribution):

$$\begin{aligned}\mathbb{E}_p[f(\boldsymbol{x})] &= \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\ &= \int f(\boldsymbol{x})p(\boldsymbol{x})\frac{q(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} = \int f(x)\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x} \\ &= \mathbb{E}_q\left[f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right]\end{aligned}$$

If we choose q in a way that we can easily sample from it, we can approximate this last expectation by Monte Carlo:

$$\mathbb{E}_q\left[f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right] \approx \frac{1}{S}\sum_{s=1}^{S} f(\boldsymbol{x}^{(s)})\frac{p(\boldsymbol{x}^{(s)})}{q(\boldsymbol{x}^{(s)})} = \frac{1}{S}\sum_{s=1}^{S} w_s f(\boldsymbol{x}^{(s)}), \quad \boldsymbol{x}^{(s)} \sim q(\boldsymbol{x})$$

Properties

- Unbiased if $q > 0$ where $p > 0$ and if we can evaluate p

Properties

- Unbiased if $q > 0$ where $p > 0$ and if we can evaluate p
- Breaks down if we do not have enough samples (puts nearly all weight on a single sample)
  ⏩ Degeneracy (see also Particle Filtering (Thrun et al., 2005))

Properties

- Unbiased if $q > 0$ where $p > 0$ and if we can evaluate p
- Breaks down if we do not have enough samples (puts nearly all weight on a single sample)
  ⏩ Degeneracy (see also Particle Filtering (Thrun et al., 2005))
- Many draws from proposal density q required, especially in high dimensions

Properties

- Unbiased if $q > 0$ where $p > 0$ and if we can evaluate p
- Breaks down if we do not have enough samples (puts nearly all weight on a single sample)

  ⏩ Degeneracy (see also Particle Filtering (Thrun et al., 2005))
- Many draws from proposal density q required, especially in high dimensions
- Requires to be able to evaluate true p. Generalization exists for $\tilde{p}$. This generalization is biased (but consistent).

Properties

- Unbiased if $q > 0$ where $p > 0$ and if we can evaluate p
- Breaks down if we do not have enough samples (puts nearly all weight on a single sample)
  ⏩ Degeneracy (see also Particle Filtering (Thrun et al., 2005))
- Many draws from proposal density q required, especially in high dimensions
- Requires to be able to evaluate true p. Generalization exists for $\tilde{p}$. This generalization is biased (but consistent).
- Does not scale to interesting (high-dimensional) problems

Properties

- Unbiased if $q > 0$ where $p > 0$ and if we can evaluate p
- Breaks down if we do not have enough samples (puts nearly all weight on a single sample)

  ⏩ Degeneracy (see also Particle Filtering (Thrun et al., 2005))
- Many draws from proposal density q required, especially in high dimensions
- Requires to be able to evaluate true p. Generalization exists for $\tilde{p}$. This generalization is biased (but consistent).
- Does not scale to interesting (high-dimensional) problems

⏩ Different approach to sample from complicated (high-dimensional) distributions

Conclusion

We saw:

- Why rectangle quadrature rules don't work in high dimensions
- How Monte Carlo estimators help
- How to draw samples using
  - Transformation techniques
  - Inverse Transform Sampling
  - Rejection Sampling
- How to improve over Rejection Sampling with Importance Sampling

References

[1]

References I

[1] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.