Statistics for Data Science

MSc Data Science WiSe 2019/20

Prof. Dr. Dirk Ostwald

Bayesian inference

(14) Numerical methods

Bibliographic remarks

The material presented in this section follows Wasserman (2004) and Held and Sabanés Bové (2014). For an introduction to Monte Carlo methods in machine learning, see Andrieu et al. (2003).

Numerical methods

- Motivation
- Quadrature
- Laplace approximation
- Monte Carlo integration
- Importance sampling
- Acceptance-rejection sampling

Numerical methods

- **Motivation**
- Quadrature
- Laplace approximation
- Monte Carlo integration
- Importance sampling
- Acceptance-rejection sampling

Motivation

- The posterior distribution is central to Bayesian inference.
- The prior and likelihood are modeling choices, but the normalization factor

$$p(x_{1:n}) = \int p(x_{1:n}|\theta)p(\theta)\, d\theta \tag{1}$$

  has to be evaluated.

- In conjugate models, the evaluation of the posterior is analytically tractable.
- In non-conjugate models, this may not be see case.

  For example, $p(\theta) = N(\theta; \mu, \sigma^2)$ and $p(x_{1:n}|\theta) = \prod_{i=1}^{n} \mathrm{C}(x_i; \theta, 1)$ yields

$$p(\theta|x_{1:n}) \propto \exp\left(\frac{-1}{2\sigma^2}(\theta - \mu)^2\right) \prod_{i=1}^{n} \left(1 + (x_i - \theta)^2\right)^{-1}. \tag{2}$$

  The right-hand side cannot be integrated analytically.

- Even is the posterior distribution is analytically tractable, evaluating Bayesian estimators

$$\hat{\theta}_B = \mathbb{E}_{p(\theta|x_{1:n})}(f(\theta)) = \int f(\theta)p(\theta|x_{1:n})\, d\theta \tag{3}$$

  may not be analytically possible.

  $\Rightarrow$ Bayesian inference often requires methods for numerical integration.

Motivation

Here, we survey the following numerical integration methods

- Quadrature approaches as classical means for numerical integration,
- the Laplace approximation as analytical integral approximation method, and
- Monte Carlo integration, i.e. numerical integration by sampling and estimation.

In addition, we consider

- Importance sampling
- Acceptance-rejection sampling

as specialized sampling schemes for Monte Carlo integration. For brevity, we omit

- discussing the pros and cons of different methods,
- considering Markov chain based sampling methods (MCMC).

Numerical methods

- Motivation
- **Quadrature**
- Laplace approximation
- Monte Carlo integration
- Importance sampling
- Acceptance-rejection sampling

Quadrature

- ... is a synonym for deterministic numerical integration.
- ... is a topic in numerical mathematics.
- ... typically works well for low-dimensional integration problems.
- ... comprises many different methods, such as
  - Riemann sums
  - Trapezoidal rule
  - Simpson's rule
  - Newton-Cotes formulas

We briefly review Riemann sums as an example.

Definition (Riemann sum and Riemann integral)

Let $f : [a, b] \to \mathbb{R}$ be a univariate real-valued function on $[a, b] \subset \mathbb{R}$.

$$P_n := \{[x_0, x_1], [x_1, x_2], ..., [x_{n-1}, x_n]\} \tag{4}$$

with

$$a =: x_0 < x_1 < x_2 < \cdots < x_n := b \tag{5}$$

be a partition of the interval $[a, b]$. Then the *Riemann sum* S_n of f over the interval $[a, b]$ based on the partition P_n is defined as

$$S_n := \sum_{i=1}^{n} f(x_i^*)\Delta x_i \text{ with } \Delta x_i := x_i - x_{i-1} \text{ and } x_i^* \in [x_{i-1}, x_i]. \tag{6}$$

The *definite Riemann integral* of f on $[a, b]$ is defined as

$$\int_a^b f(x)\, dx = \lim_{\Delta x \to 0} \sum_{i=1}^{n} f(x_i^*)\Delta x_i, \tag{7}$$

if the limit on the right-hand side exists.

Remarks

- A basic quadrature idea is to approximate integrals by Riemann sums.

Definition (Left rule, right rule, midpoint rule)

Let $f : [a, b] \to \mathbb{R}$ and assume the aim is to approximate the integral

$$I = \int_a^b f(x)\, dx \tag{8}$$

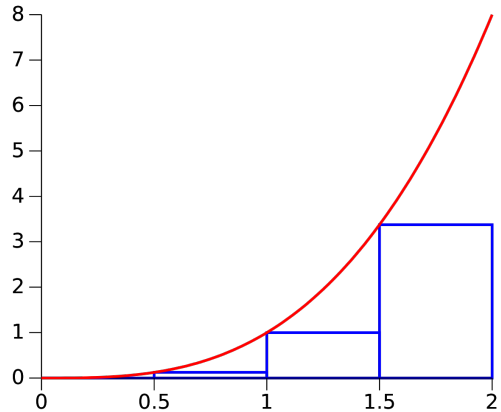
using a Riemann sum S_n. Then choosing

- $x_i^* := x_{i-1}$ and setting is called the *left rule*,
- $x_i^* := x_i$ is called the *right rule*, and
- $x_i^* := \frac{1}{2}(x_i + x_{i-1})$ is called the *midpoint rule*.

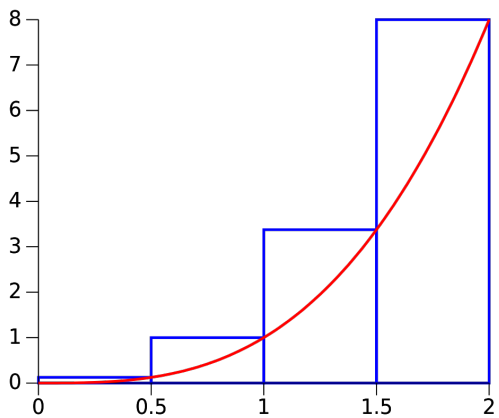
Remarks

- As $\Delta x \to 0$, the choice of x_i^* does not matter for the Riemann integral.
- Often equipartitions of the form $\Delta x = (b - a)/n$ are used.
- f is approximated by piecewise constant functions.

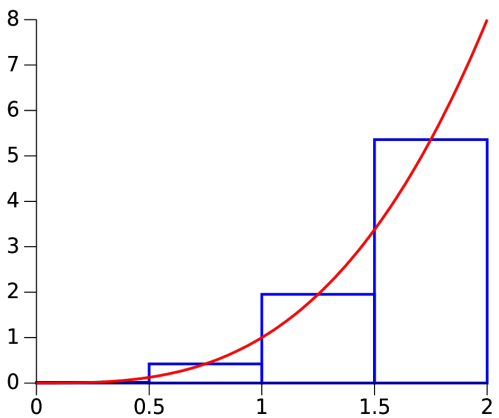
Left rule



Right rule



Midpoint rule



https://en.wikipedia.org/wiki/Riemann_sum

Numerical methods

- Motivation
- Quadrature
- **Laplace approximation**
- Monte Carlo integration
- Importance sampling
- Acceptance-rejection sampling

The Laplace approximation

- ... is a method to approximate posterior expectations of the type

$$\mathbb{E}_{p(\theta|x_{1:n})}\left(f(\theta)\right) = \int f(\theta)p(\theta|x_{1:n})\,d\theta, \tag{9}$$

  such as the posterior expected value

$$\mathbb{E}_{p(\theta|x_{1:n})} = \int \theta\, p(\theta|x_{1:n})\,d\theta. \tag{10}$$

- ... does not require the functional form of the posterior distribution.
- ... is an application of *Laplace's integral approximation method.*

Definition (Laplace's integral approximation method)

For a univariate, real-valued, convex, and twice differentiable function f with a minimum at $\tilde{x}$, a reasonable approximation of the integral

$$I_n = \int_{-\infty}^{\infty} \exp\left(-nf(x)\right)\, dx \tag{11}$$

is given by

$$I_n \approx \exp\left(-nf(\tilde{x})\right) \sqrt{\frac{2\pi}{n\kappa}}, \tag{12}$$

where

$$\kappa := f''(\tilde{x}) > 0 \tag{13}$$

denotes the second derivative of f at its minimum location $\tilde{x}$.

Remarks

- The minimum location $\tilde{x}$ and $\kappa = f''(\tilde{x})$ have to be evaluated.
- The approximation rests on a second-order Taylor approximation of f in $\tilde{x}$ and

$$\int_{-\infty}^{\infty} \exp\left(-a(x-b)^2\right)\, dx = \sqrt{\pi/a}. \tag{14}$$

- "Reasonable" means that the approximation error decreases for $n \to \infty$.

Laplace approximation

Motivation of Laplace's integral approximation method

Consider the second-order Taylor approximation of f in $\tilde{x}$,

$$f(x) \approx f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}) + \frac{1}{2} f''(\tilde{x})(x - \tilde{x})^2. \tag{15}$$

Because $\tilde{x}$ is a minimum of f, it holds that $f'(\tilde{x}) = 0$ and $\kappa := f''(\tilde{x}) > 0$. We then have

$$\begin{aligned}
I_n &= \int_{-\infty}^{\infty} \exp\left(-n f(x)\right)\, dx \\
&\approx \int_{-\infty}^{\infty} \exp\left(-n\left(f(\tilde{x}) + \frac{1}{2}\kappa(x - \tilde{x})^2\right)\right)\, dx \\
&= \int_{-\infty}^{\infty} \exp\left(-n f(\tilde{x}) + \frac{n}{2}\kappa(x - \tilde{x})^2\right)\, dx \\
&= \int_{-\infty}^{\infty} \exp\left(-n f(\tilde{x})\right) \exp\left(-\frac{n}{2}\kappa(x - \tilde{x})^2\right)\, dx \\
&= \exp\left(-n f(\tilde{x})\right) \int_{-\infty}^{\infty} \exp\left(-\frac{n\kappa}{2}(x - \tilde{x})^2\right)\, dx
\end{aligned} \tag{16}$$

With

$$\int_{-\infty}^{\infty} \exp\left(-a(x - b)^2\right)\, dx = \sqrt{\pi/a}, \tag{17}$$

it then follows that

$$I_n = \exp\left(-n f(\tilde{x})\right) \sqrt{\frac{2\pi}{n\kappa}}. \tag{18}$$

□

Definition (Laplace approximation)

Let

$$p(\theta, x_{1:n}) = p(x_{1:n}|\theta)p(\theta) \tag{19}$$

be probabilistic model with scalar parameter θ and $x_{1:n} = (x_1, .., x_n)$ denoting a value of a random sample $X = (X_1, ..., X_n)$ with $X_i \sim p(x_i|\theta)$ for $i = 1, ..., n$. Consider the problem of evaluating a posterior distribution expectation of the form

$$\mathbb{E}_{p(\theta|x_{1:n})}\left(f(\theta)\right) \text{ for } f : \mathbb{R} \to \mathbb{R}, \theta \mapsto f(\theta). \tag{20}$$

A reasonable approximation of such a feature is given by

$$\mathbb{E}_{p(\theta|x_{1:n})}\left(f(\theta)\right) \approx \sqrt{\frac{\kappa_1}{\kappa_2}} \exp\left(-n\left(h_2(\tilde{\theta}_2) - h_1(\tilde{\theta}_1)\right)\right), \tag{21}$$

where

$$\begin{aligned} h_1 &: \mathbb{R} \to \mathbb{R}, \theta \mapsto h_1(\theta) := -\ln f(\theta) - \ln p(x_i|\theta) - \ln p(\theta) \\ h_2 &: \mathbb{R} \to \mathbb{R}, \theta \mapsto h_2(\theta) := -\ln p(x_i|\theta) - \ln p(\theta), \end{aligned} \tag{22}$$

$\tilde{\theta}_1$ and $\tilde{\theta}_2$ are minimum points of h_1 and h_2, respectively, and

$$\kappa_1 := h_1''(\tilde{\theta}_1) > 0 \text{ and } \kappa_2 := h_2''(\tilde{\theta}_2) > 0. \tag{23}$$

Remarks

- "Reasonable" means that the approximation error decreases for $n \to \infty$.
- If $\kappa_1, \kappa_2, \tilde{\theta}_1, \tilde{\theta}_2$ are not available analytically, they are evaluated numerically.
- An application of the Laplace approximation is discussed in the Übung.

Motivation of the Laplace approximation

We first note that with the definitions of h_1 and h_2, we have

$$
\begin{aligned}
\mathbb{E}_{p(\theta|x)}(\theta) &= \int f(\theta) p(\theta|x_{1:n})\, d\theta \\
&= \int f(\theta) \frac{p(x_{1:n}|\theta)p(\theta)}{\int p(x_{1:n}|\theta)p(\theta)\, d\theta}\, d\theta \\
&= \frac{\int f(\theta)p(x_{1:n}|\theta)p(\theta)\, d\theta}{\int p(x_{1:n}|\theta)p(\theta)\, d\theta} \\
&= \frac{\int \exp\left(\ln\left(f(\theta)p(x_{1:n}|\theta)p(\theta)\right)\right)\, d\theta}{\int \exp\left(\ln\left(p(x_{1:n}|\theta)p(\theta)\right)\right)\, d\theta} \\
&= \frac{\int \exp\left(\ln\left(f(\theta)\prod_{i=1}^n p(x_i|\theta)p(\theta)\right)\right)\, d\theta}{\int \exp\left(\ln\left(\prod_{i=1}^n p(x_i|\theta)p(\theta)\right)\right)\, d\theta} \\
&= \frac{\int \exp\left(\ln\left(\prod_{i=1}^n f(\theta)p(x_i|\theta)p(\theta)\right)\right)\, d\theta}{\int \exp\left(\ln\left(\prod_{i=1}^n p(x_i|\theta)p(\theta)\right)\right)\, d\theta} \\
&= \frac{\int \exp\left(n\ln f(\theta) + n\ln p(x_i|\theta) + n\ln p(\theta)\right)\, d\theta}{\int \exp\left(n\ln p(x_i|\theta) + n\ln p(\theta)\right)\, d\theta} \\
&= \frac{\int \exp(-nh_2(\theta))}{\int \exp(-nh_1(\theta))}.
\end{aligned}
\tag{24}
$$

Motivation of the Laplace approximation (cont.)

Application of Laplace's integral approximation method to the numerator and denominator of

$$\mathbb{E}_{p(\theta|x_{1:n})}(\theta) = \frac{\int \exp(-nh_2(\theta))}{\int \exp(-nh_1(\theta))} \tag{25}$$

then yields

$$\begin{aligned}
\mathbb{E}_{p(\theta|x_{1:n})}(\theta) &\approx \frac{\exp(-nh_1(\tilde{\theta_1}))\sqrt{\frac{2\pi}{n\kappa_1}}}{\exp(-nh_2(\tilde{\theta_2}))\sqrt{\frac{2\pi}{n\kappa_2}}} \\
&= \sqrt{\frac{2\pi}{n\kappa_1}\frac{n\kappa_2}{2\pi}}\exp\left(-nh_1(\tilde{\theta}_1 + nh_2(\tilde{\theta}_2)\right) \\
&= \sqrt{\frac{\kappa_1}{\kappa_2}}\exp\left(-n\left(h_1(\tilde{\theta}_1) - h_2(\tilde{\theta}_2)\right)\right)
\end{aligned} \tag{26}$$

□

Numerical methods

- Motivation
- Quadrature
- Laplace approximation
- **Monte Carlo integration**
- Importance sampling
- Acceptance-rejection sampling

Monte Carlo integration

- ... is a means to approximate integrals using random sampling.
- ... was allegedly a code word in the Manhattan project.
- ... are used for high-dimensional integration with strong localization.
- ... replaces deterministic support points with random support points.

Definition (Monte Carlo estimator, Monte Carlo algorithm)

For a univariate continuous random variable X with PDF p and a univariate real-valued function f, let

$$I := \mathbb{E}_{p(x)}(f(X)) := \int_{\mathcal{X}} f(x)p(x)\,dx \tag{27}$$

Furthermore, let $X_1, ..., X_n$ be a sample of independent copies of X. Then

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \tag{28}$$

is called the *Monte Carlo estimator* of the integral I. A *Monte Carlo algorithm* to obtain a Monte Carlo estimate of I is given by

(1) Sample $X_1, ..., X_n \sim p$

(2) Evaluate and return $\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.

Theorem (Unbiasedness and consistency of the Monte Carlo estimator)

The Monte Carlo estimator is unbiased and consistent.

Proof

Unbiasedness

$$\mathbb{E}\left(\hat{I}_n\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(f(X_i)\right) = \frac{1}{n} n\mathbb{E}\left(f(X)\right) = \mathbb{E}(f(X)). \tag{29}$$

Consistency

The weak law of large numbers states that

$$\lim_{n\to\infty}\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}(f(X))| \geq \epsilon\right) = 0 \tag{30}$$

which implies the consistency of the estimator $\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$

□

Example (Monte Carlo estimator)

Consider evaluating the integral

$$I = \int_0^1 \frac{1}{1-x^2}\, dx. \tag{31}$$

Then a Monte Carlo estimator of I is

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1-x_i^2}, \tag{32}$$

where $x_1, ..., x_n$ denote n independent realizations of uniformly distributed random variables $X_1, ..., X_n \sim U(0, 1)$.

Numerical methods

- Motivation
- Quadrature
- Laplace approximation
- Monte Carlo integration
- **Importance sampling**
- Acceptance-rejection sampling

Importance sampling

- ... is a method to approximate expected values $\mathbb{E}_{p(x)}\left(f(X)\right)$
- ... works by sampling $X \sim q$ instead of $X \sim p$.
- ... can also be applied, if the normalizing constants of q and p are unknown.
- ... can be used to reduce the variance of Monte Carlo estimators.

Theorem (Importance sampling identity)

Let p and q be probability density functions on $\mathcal{X} \subseteq \mathbb{R}$ such that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x) > 0$. Then for any $f : \mathcal{X} \to \mathbb{R}$, it holds that

$$\mathbb{E}_{p(x)}\left(f(X)\right) = \mathbb{E}_{q(x)}\left(f(X)w(X)\right), \tag{33}$$

where

$$w : \mathcal{X} \to \mathbb{R}_{>0}, x \mapsto w(x) := \frac{p(x)}{q(x)} \tag{34}$$

denotes the *importance weight function*.

Proof

$$\mathbb{E}_{p(x)}\left(f(X)\right) = \int_{\mathcal{X}} f(x)p(x)\,dx = \int_{\mathcal{X}} f(x)\frac{p(x)}{q(x)}q(x)\,dx = \int_{\mathcal{X}} f(x)w(x)q(x)\,dx = \mathbb{E}_{q(x)}\left(f(X)w(x)\right) \tag{35}$$

□

Remarks

- Note that the expectation of f under p is equal to the expectation of fw under q.
- p and q are referred to as *nominal* and *importance distributions*, respectively.

Theorem (Importance sampling estimator)

Let p and q be probability density functions on $\mathcal{X} \subseteq \mathbb{R}$ and let $f : \mathcal{X} \to \mathbb{R}$ be a function such that

$$I := \mathbb{E}_{p(x)}(f(X)) \tag{36}$$

exists. Assume that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x)f(x) \neq 0$. Finally, assume that $X_1, ..., X_n \sim q$. Then

$$\hat{I}_n := \frac{1}{n} \sum_{i=1}^{n} f(X_i)w(X_i) \tag{37}$$

is an unbiased and consistent estimator of I.

Proof

We have

$$\mathbb{E}_{q(x)}\left(\hat{I}_n\right) = \mathbb{E}_{q(x)}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)w(X_i)\right) = \frac{1}{n} n\, \mathbb{E}_{q(x)}\left(f(X_i)w(X_i)\right) = \mathbb{E}_{p(x)}\left(f(X_i)\right) \tag{38}$$

and hence the estimator is unbiased. Consistency follows with the law of large numbers.

□

Remarks

- $X_1, ..., X_n \sim q$ and $\hat{I}_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i)\frac{p(X_i)}{q(X_i)}$ are used to estimate $\mathbb{E}_{p(x)}(f(X))$.

Theorem (Normalized importance sampling identity)

For normalization constants c_p, c_q let $p := \tilde{p}/c_p$ and $q = \tilde{q}/c_q$ be probability density functions on $\mathcal{X} \subseteq \mathbb{R}$ such that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x) > 0$. Then for any $f : \mathcal{X} \to \mathbb{R}$, it holds that

$$\mathbb{E}_{p(x)}\left(f(X)\right) = \frac{\mathbb{E}_{q(x)}\left(f(X)\tilde{w}(X)\right)}{\mathbb{E}_{q(x)}\left(\tilde{w}(X)\right)}, \tag{39}$$

where $\tilde{w}$ denotes the importance weight function

$$\tilde{w} : \mathcal{X} \to \mathbb{R}_{>0}, x \mapsto \tilde{w}(x) := \frac{\tilde{p}(x)}{\tilde{q}(x)}. \tag{40}$$

Remark

- $\mathbb{E}_{p(x)}\left(f(X)\right)$ is here estimated based on an "unnormalized PDF" $\tilde{p} = c_p p$.

Proof

$$
\begin{aligned}
\frac{\mathbb{E}_{q(x)}\left(f(X)\tilde{w}(X)\right)}{\mathbb{E}_{q(X)}\left(\tilde{w}(X)\right)} &= \frac{\mathbb{E}_{q(x)}\left(f(X)\frac{\tilde{p}(X)}{\tilde{q}(X)}\right)}{\mathbb{E}_{q(x)}\left(\frac{\tilde{p}(X)}{\tilde{q}(X)}\right)} \\
&= \frac{\mathbb{E}_{q(x)}\left(f(X)\frac{c_p c_q \tilde{p}(X)}{c_p c_q \tilde{q}(X)}\right)}{\mathbb{E}_{q(x)}\left(\frac{c_p c_q \tilde{p}(X)}{c_p c_q \tilde{q}(X)}\right)} \\
&= \frac{\mathbb{E}_{q(x)}\left(f(X)\frac{c_q p(X)}{c_p q(X)}\right)}{\mathbb{E}_{q(x)}\left(\frac{c_q p(X)}{c_p q(X)}\right)} \\
&= \frac{\int_{\mathcal{X}} f(x)\frac{c_q p(x)}{c_p q(x)} q(x)\,dx}{\int_{\mathcal{X}} \frac{c_q p(x)}{c_p q(x)} q(x)\,dx} \qquad (41) \\
&= \frac{\int_{\mathcal{X}} f(x)\frac{c_q p(x)}{c_p}\,dx}{\int_{\mathcal{X}} \frac{c_q p(x)}{c_p}\,dx} \\
&= \frac{\frac{c_q}{c_p}\int_{\mathcal{X}} f(x)p(x)\,dx}{\frac{c_q}{c_p}\int_{\mathcal{X}} p(x)\,dx} \\
&= \int_{\mathcal{X}} f(x)p(x)\,dx = \mathbb{E}_{p(x)}(f(X))
\end{aligned}
$$

□

Definition (Normalized importance sampling estimator)

For an unknown normalization constant c_p and $c_q := 1$ let $p := \tilde{p}/c_p$ and $q := \tilde{q}/c_q$ be probability density functions on $\mathcal{X} \subseteq \mathbb{R}$ such that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x) > 0$. Let $f : \mathcal{X} \to \mathbb{R}$ be a function such that

$$I := \mathbb{E}_{p(x)}(f(X)) \tag{42}$$

exists. Assume that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x)f(x) \neq 0$. Finally, assume that $X_1, ..., X_n \sim q$. Then

$$\hat{I}_n := \frac{\sum_{i=1}^n f(X_i)\tilde{w}(X_i)}{\sum_{i=1}^n \tilde{w}(X_i)} \tag{43}$$

is called the *normalized importance sampling estimator* of I

Remarks

- $X_1, ..., X_n \sim q$ and $\hat{I}_n$ are used to estimate I for the "unnormalized PDF" $\tilde{p}$.
- The sample mean factor n^{-1} cancels out in the $\hat{I}_n$ fraction.
- It can be shown that $\hat{I}_n$ is biased, but asymptotically unbiased and consistent.

Numerical methods

- Motivation
- Quadrature
- Laplace approximation
- Monte Carlo integration
- Importance sampling
- **Acceptance-rejection sampling**

Acceptance-rejection sampling

- ... is a method to obtain samples from $Y \sim p_Y$ by sampling $X \sim p_X$.
- ... an algorithmic alternative to the probability integral transform.
- ... works by strategically rejecting samples from X.
- ... works by inducing a sampling bias to the samples from X.
- ... can be used in MC integration, if sampling from the posterior is difficult.

Theorem (Acceptance-rejection sampling)

Let p_X and p_Y be two probability density functions, referred to as *proposal density* and *target density*, respectively, such that

- random samples of $X \sim p_X$ can be obtained,
- the function p_Y/p_X can be evaluated, and
- $p_Y \leq cp_X$ for a constant $c \in \mathbb{R}$.

Consider the following *acceptance-rejection algorithm*

1. Draw a realization $X \sim p_X$ of the proposal density.
2. Draw a realization $U \sim \mathrm{U}(0,1)$ of the uniform distribution on $[0,1]$.
3. If
$$U \leq \frac{p_Y(X)}{cp_X(X)} \tag{44}$$
return $Y = X$. Otherwise, return to step 1.

Then Y is distributed according to the target density, $Y \sim p_Y$.

Remarks

- Note that the algorithm returns $X \sim p_X$ conditional on $U \leq \frac{p_Y(X)}{cp_X(X)}$.
- The claim of the theorem can thus equivalently be expressed as

$$\mathbb{P}\left(X \leq x \middle| U \leq \frac{p_Y(X)}{cp_X(X)}\right) = \int_{-\infty}^{x} p_Y(\xi)\, d\xi. \tag{45}$$

Proof

We want to show that the conditional distribution of X given the event $U \leq p_Y(X)/cp_X(X)$ conforms to the target density p_Y. For ease of notation, let

$$f(X) := p_Y(X)/cp_X(X). \tag{46}$$

The conditional probability of the event $X \leq x$ given the event $U \leq f(X)$ can then be written in as

$$\mathbb{P}\left(X \leq x | U \leq f(X)\right) = \frac{\mathbb{P}\left(X \leq x, U \leq f(X)\right)}{\mathbb{P}\left(U \leq f(X)\right)} \tag{47}$$

To evaluate the numerator of the right-hand side of the above, we first note that

$$\begin{aligned} \mathbb{P}(X \leq x, U \leq f(X)) &= \mathbb{P}(U \leq f(X) | X \leq x)\mathbb{P}(X \leq x) \\ &= \mathbb{P}(U \leq f(x))\mathbb{P}(X \leq x) \\ &= \int_{-\infty}^{x} \mathbb{P}\left(U \leq f(\xi)\right) p_X(\xi)\, d\xi \end{aligned} \tag{48}$$

With the CDF of the continuous uniform distribution on $[0, 1]$ and the definition of $f(X)$, we then obtain

Proof (cont.)

$$\begin{aligned}
\mathbb{P}(X \le x, U \le f(X)) &= \int_{-\infty}^{x} \mathbb{P}\left(U \le f(\xi)\right) p_X(\xi)\, d\xi \\
&= \int_{-\infty}^{x} f(\xi) p_X(\xi)\, d\xi \\
&= \int_{-\infty}^{x} \frac{p_Y(\xi)}{c p_X(\xi)} p_X(\xi)\, d\xi \\
&= \frac{1}{c} \int_{-\infty}^{x} p_Y(\xi)\, d\xi
\end{aligned} \tag{49}$$

We next evaluate the denominator of the above and obtain

$$\begin{aligned}
\mathbb{P}(U \le f(X)) &= \int_{-\infty}^{\infty} \mathbb{P}(U \le f(X) | X = x) \mathbb{P}(X = x)\, dx \\
&= \int_{-\infty}^{\infty} \mathbb{P}(U \le f(x)) \mathbb{P}(X = x)\, dx \\
&= \int_{-\infty}^{\infty} \mathbb{P}\left(U \le f(\xi)\right) p_X(\xi)\, d\xi \\
&= \int_{-\infty}^{\infty} \frac{p_Y(\xi)}{c p_X(\xi)} p_X(\xi)\, d\xi \\
&= \frac{1}{c}.
\end{aligned} \tag{50}$$

Proof (cont.)

In summary, we obtain

$$\mathbb{P}\left(X \leq x | U \leq f(X)\right) = \frac{\frac{1}{c}\int_{-\infty}^{x} p_Y(\xi)\, d\xi}{\frac{1}{c}} = \int_{-\infty}^{x} p_Y(\xi) \tag{51}$$

The conditional distribution of X given $U \leq \frac{p_Y(X)}{c p_X(X)}$ thus has PDF p_Y. Because we call this random variable distributed according to this distribution Y, we have obtained $Y \sim p_Y$.

□

References

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43.

Held, L. and Sabanés Bové, D. (2014). *Applied statistical inference*, volume 10. Springer.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer.