# Chapter 5: Monte Carlo Approximation

## Contents

## 1 Introduction

Sampling-based methods are extensively used in modern statistics, due to their ease of use and the generality with which they can be applied. The fundamental problem solved by these methods is the approximation of expectations such as

$$\mathbb{E}h(X) = \int h(x)p(x)dx$$

in the case of a continuous random variable $X$ with p.d.f. $p$, or

$$\mathbb{E}h(X) = \sum_x h(x)p(x)$$

in the case of a discrete random variable $X$ with p.m.f. $p$. The general principle at work is that such expectations can be approximated by

$$\mathbb{E}h(X) \approx \frac{1}{N}\sum_{i=1}^N h(X_i),$$

where $X_1, \ldots, X_N$ are samples from $p$. Although, at first, computing expectations may seem to be of rather limited utility, the vast majority of inferential problems can be put in this form.

Jeffrey W. Miller (2015). *Lecture Notes on Bayesian Statistics.* Duke University, Durham, NC.

**Some things that can be approximated with samples:**

- posterior probabilities
- posterior densities
- posterior expected loss
- posterior predictive distribution
- marginal likelihood
- goodness-of-fit statistics

Samples are also a great way of visualizing where a probability distribution is putting most of its mass—this is especially useful for distributions on complex and/or high-dimensional spaces, e.g., the folding of a protein or RNA strand.

**Advantages of sampling-based methods:**

- easy to implement
- general-purpose / widely applicable
- reliable
- work in complex and high-dimensional spaces

**Disadvantages of sampling-based methods:**

- slow (may require much more time to achieve the same level of accuracy)
- getting "true" samples may be difficult
- can be difficult to assess accuracy

It seems that the computationally-difficult problems in statistics always take the form of intractable integrals or sums, and sampling-based approximations are often the only known approach that works in practice.

# 2 Monte Carlo approximation

Suppose we want to know the expectation of a random variable $X$ with p.d.f. or p.m.f. $p$. To make a ***simple Monte Carlo approximation*** (or just ***Monte Carlo approximation***), we draw i.i.d. samples $X_1, \ldots, X_N \sim p$ and use

$$\frac{1}{N}\sum_{i=1}^{N} X_i$$

as an approximation to $\mathbb{E}X$. Although it might seem to be an oversimplified special case, this is in fact equivalent to seemingly more general approximations such as:

$$\mathbb{E}\big(h(Y) \mid Z = z\big) \approx \frac{1}{N}\sum_{i=1}^{N} h(Y_i)$$

where $Y_1, \ldots, Y_N$ are i.i.d. samples from the conditional distribution of $Y \mid Z = z$. (It is equivalent because we can define $X$ to have the distribution of $h(Y) \mid Z = z$.)

### 2.0.1 Basic properties

- If $\mathbb{E}|X| < \infty$, then $\frac{1}{N}\sum X_i$ is a consistent estimator of $\mathbb{E}X$, that is,

  $$\frac{1}{N}\sum_{i=1}^{N} X_i \longrightarrow \mathbb{E}X$$

  as $N \to \infty$, with probability 1, by the law of large numbers. This guarantees that the approximation will converge to the true value (if $\mathbb{E}|X| < \infty$).

- $\frac{1}{N}\sum X_i$ is an unbiased estimator of $\mathbb{E}X$, that is,

  $$\mathbb{E}\Big(\tfrac{1}{N}\sum X_i\Big) = \mathbb{E}X.$$

- The variance of $\frac{1}{N}\sum X_i$ is

  $$\mathbb{V}\Big(\tfrac{1}{N}\sum X_i\Big) = \tfrac{1}{N^2}\mathbb{V}\big(\sum X_i\big) = \tfrac{1}{N^2}\textstyle\sum_{i=1}^{N}\mathbb{V}\big(X_i\big) = \tfrac{1}{N}\mathbb{V}\big(X\big)$$

  since the variance of a sum of independent variables is the sum of the variances.

- Due to unbiasedness, the root-mean-squared-error (RMSE) equals the standard deviation (square root of the variance) of $\frac{1}{N}\sum X_i$,

  $$\begin{aligned}\text{RMSE} &= \Big[\mathbb{E}\big(|\tfrac{1}{N}\textstyle\sum X_i - \mathbb{E}X|^2\big)\Big]^{1/2}\\ &= \Big[\mathbb{V}\big(\tfrac{1}{N}\textstyle\sum X_i\big)\Big]^{1/2}\\ &= \frac{1}{\sqrt{N}}\mathbb{V}(X)^{1/2} = \sigma(X)/\sqrt{N}. \qquad (2.1)\end{aligned}$$

  The RMSE tells us how far the approximation will be from the true value, on average. Since the standard deviation $\sigma(X)$ does not depend on $N$, this tells us that the rate of convergence is of order $1/\sqrt{N} = N^{-1/2}$. It is a minor miracle that this result is so easily obtained and holds under such general conditions.

As a practical matter, we need to be able to draw the samples $X_i$ in a computationally-efficient way.

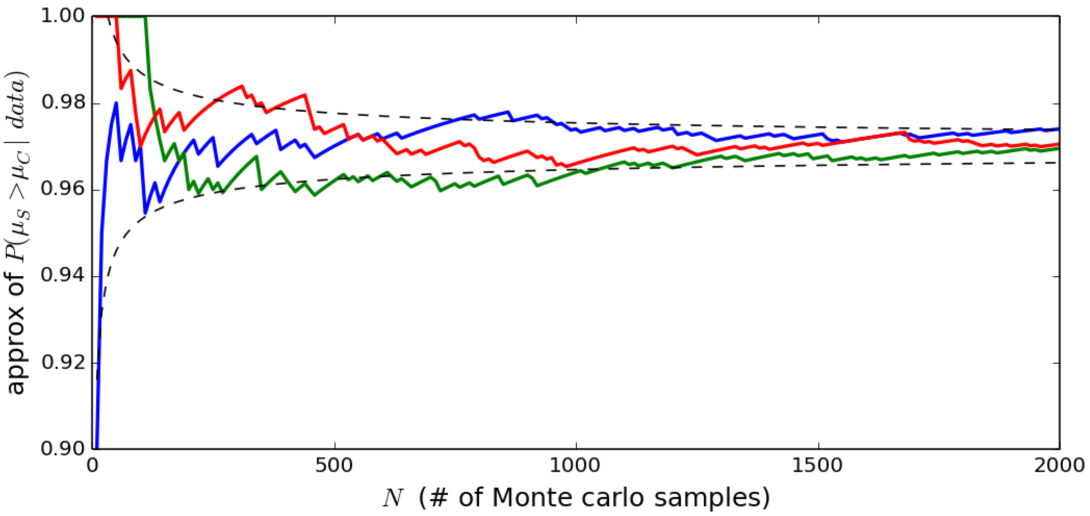

Figure 1: Monte Carlo approximations for an increasing number of samples, $N$. The red, blue, and green lines indicate three repetitions of the procedure, using different sequences of samples. The dotted lines indicate the true value ± the RMSE of the Monte Carlo estimator.

## 2.1 Examples

### 2.1.1 The Pygmalion effect

In Chapter 4, we saw an example involving the mean change in IQ score $\mu_S$ and $\mu_C$ of two groups of students (spurters and controls). To compute the posterior probability that the spurters had a larger mean change in IQ score, we drew $N = 10^6$ samples from each posterior:

$$(\boldsymbol{\mu}_S^{(1)}, \boldsymbol{\lambda}_S^{(1)}), \ldots, (\boldsymbol{\mu}_S^{(N)}, \boldsymbol{\lambda}_S^{(N)}) \sim \text{NormalGamma}(24.0, 8, 4, 855.0)$$
$$(\boldsymbol{\mu}_C^{(1)}, \boldsymbol{\lambda}_C^{(1)}), \ldots, (\boldsymbol{\mu}_C^{(N)}, \boldsymbol{\lambda}_C^{(N)}) \sim \text{NormalGamma}(11.8, 49, 24.5, 6344.0)$$

and used the Monte Carlo approximation

$$\mathbb{P}(\boldsymbol{\mu}_S > \boldsymbol{\mu}_C \mid \text{data}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\big(\boldsymbol{\mu}_S^{(i)} > \boldsymbol{\mu}_C^{(i)}\big).$$

To visualize this, consider the sequence of approximations $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big(\boldsymbol{\mu}_S^{(i)} > \boldsymbol{\mu}_C^{(i)}\big)$ for $N = 1, 2, \ldots$. Figure 1 shows this sequence of approximations for three different sets of random samples from the posterior. We can see that as the number of samples used in the approximation grows, it appears to be converging to around 0.97. To visualize the theoretical rate of convergence, the figure also shows bands indicating the true value $\alpha = \mathbb{P}(\boldsymbol{\mu}_S > \boldsymbol{\mu}_C \mid \text{data}) = 0.97$ plus or minus the RMSE of the Monte Carlo estimator, that is, from Equation 2.1:

$$\alpha \pm \sigma(X)/\sqrt{N} = \alpha \pm \sqrt{\alpha(1-\alpha)/N}$$

$$= 0.97 \pm \sqrt{0.97(1-0.97)/N}$$

where $X$ has the posterior distribution of $\mathbb{1}(\boldsymbol{\mu}_S > \boldsymbol{\mu}_C)$ given the data, in other words, $X$ is a Bernoulli($\alpha$) random variable. Recall that the variance of a Bernoulli($\alpha$) random variable is $\alpha(1-\alpha)$.

Using the same approach, we could easily approximate any number of other posterior quantities as well, for example,

$$\mathbb{P}\big(\boldsymbol{\lambda}_S > \boldsymbol{\lambda}_C \,\big|\, \text{data}\big) \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big(\boldsymbol{\lambda}_S^{(i)} > \boldsymbol{\lambda}_C^{(i)}\big)$$

$$\mathbb{E}\big(|\boldsymbol{\mu}_S - \boldsymbol{\mu}_C| \,\big|\, \text{data}\big) \approx \frac{1}{N}\sum_{i=1}^{N} |\boldsymbol{\mu}_S^{(i)} - \boldsymbol{\mu}_C^{(i)}|$$

$$\mathbb{E}\big(\boldsymbol{\mu}_S/\boldsymbol{\mu}_C \,\big|\, \text{data}\big) \approx \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\mu}_S^{(i)}/\boldsymbol{\mu}_C^{(i)}.$$

### 2.1.2 Posterior expected loss for resource allocation

In Chapter 1, we looked at a decision theory example involving allocation of resources for prevention and treatment of a disease. Rather than computing the posterior expected loss using numerical integration, as we did in that example, we could use a Monte Carlo approximation:

$$\rho(c,x) = \mathbb{E}(\ell(\boldsymbol{\theta},c)|x) = \int \ell(\theta,c)p(\theta|x)d\theta \approx \frac{1}{N}\sum_{i=1}^{N}\ell(\boldsymbol{\theta}_i,c)$$

where $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N$ are i.i.d. samples from the posterior, $p(\theta|x)$.

### 2.1.3 Distributions of various posterior quantities

Posterior samples can also be used to approximate the posterior p.d.f. or p.m.f. of a complicated quantity, when it might otherwise be difficult to derive. Here's an example.

In the 1998 General Social Survey, respondents were asked about their religious preference and (among other questions) whether they agreed with a Supreme Court ruling prohibiting state and local governments from requiring that certain religious material be read in schools. Out of the 1011 Protestants in the survey, 353 agreed with the ruling, while out of 860 non-Protestants, 441 agreed. Suppose that across the entire U.S. population, $\theta_p$ and $\theta_n$ are the proportions of Protestants and non-Protestants, respectively, that would agree with the ruling.

How many times as likely to agree are non-Protestants than Protestants? In other words, what is $\theta_n/\theta_p$? Placing independent uniform priors (that is, Beta$(1,1)$ priors) on $\theta_p$ and $\theta_n$, and using a binomial model, we find that the posteriors are $\boldsymbol{\theta}_p|\text{data} \sim \text{Beta}(354,659)$ and $\boldsymbol{\theta}_n|\text{data} \sim \text{Beta}(442,420)$. From this, we can easily construct an approximation to the posterior p.d.f. of $\theta_n/\theta_p$ by drawing independent samples

$$\boldsymbol{\theta}_p^{(1)},\ldots,\boldsymbol{\theta}_p^{(N)} \sim \text{Beta}(354,659)$$

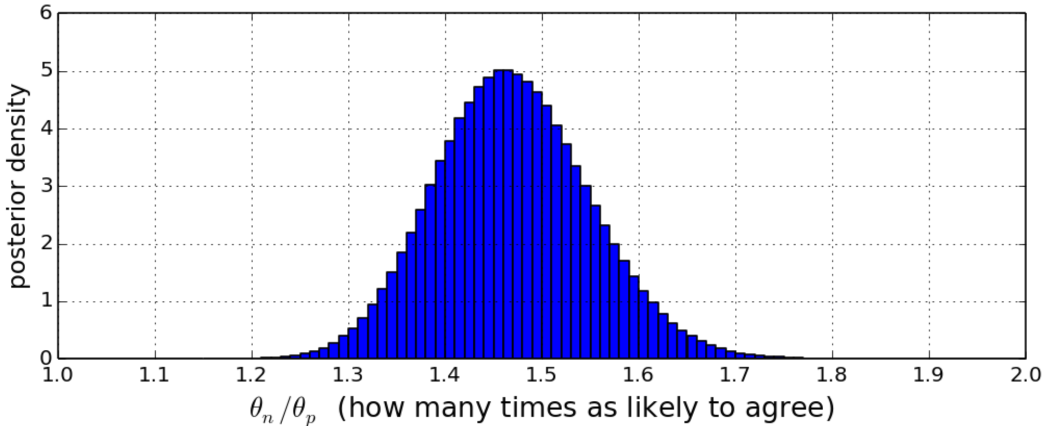

Figure 2: Estimated posterior density of $\theta_n/\theta_p$ for the Protestant / non-Protestant example.

$$\boldsymbol{\theta}_n^{(1)}, \ldots, \boldsymbol{\theta}_n^{(N)} \sim \text{Beta}(442, 420)$$

and making a histogram (or other density estimate) from the ratios of the samples,

$$\boldsymbol{\theta}_n^{(1)}/\boldsymbol{\theta}_p^{(1)}, \; \ldots, \; \boldsymbol{\theta}_n^{(N)}/\boldsymbol{\theta}_p^{(N)}.$$

See Figure 2. Note that each bin of the histogram corresponds to a Monte Carlo approximation of the probability of a sample landing in that bin.

### 2.1.4 Approximating the posterior predictive density

A Monte Carlo approximation to the posterior predictive p.d.f. or p.m.f. can be made using samples from the posterior:

$$\begin{aligned} p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1}|\theta)p(\theta|x_{1:n})d\theta \\ &= \mathbb{E}\big(p(x_{n+1}|\boldsymbol{\theta}) \mid x_{1:n}\big) \\ &\approx \frac{1}{N}\sum_{i=1}^{N} p(x_{n+1}|\boldsymbol{\theta}_i) \end{aligned}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N \overset{\text{iid}}{\sim} p(\theta|x_{1:n})$. This is useful when it is difficult or impossible to evaluate the integral analytically.

## 3 Importance sampling approximation

Importance sampling (IS) is a more powerful type of Monte Carlo approximation. The name "importance sampling" is somewhat misleading, since it is not really a method for drawing samples, but rather, a method for approximating expectations—a better name might be importance-weighted approximation.

**Advantages of importance sampling over simple Monte Carlo**

- Can *significantly* improve performance, by reducing the variance
- Can use samples from a different distribution, say $q$, to compute expectations with respect to $p$
- Can compute the normalization constant of $p$, as well as Bayes factors

**Disadvantages**

- Need to be able to evaluate the p.d.f.s/p.m.f.s $p(x)$ and $q(x)$, at least up to proportionality constants
- It might not be obvious how to choose a good $q$

## 3.1 The basic idea

Suppose $X \sim p$ is continuous (the same thing works in the discrete case), and let $q$ be the p.d.f. of a distribution we can easily sample from. Assume $q(x) > 0$ for all $x$ (or more generally, assume $q(x) > 0$ whenever $p(x) > 0$; see Section 3.3). Then

$$\begin{aligned}\mathbb{E}h(X) &= \int h(x)p(x)dx \\ &= \int h(x)\frac{p(x)}{q(x)}q(x)dx \\ &\approx \frac{1}{N}\sum_{i=1}^{N} h(Y_i)\frac{p(Y_i)}{q(Y_i)}\end{aligned}$$

where $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} q$, is called an ***importance sampling approximation***. The approximation step here is just a simple Monte Carlo approximation, as in Section 2. The distribution $q$ is sometimes called the ***proposal distribution***. The ratios $w(Y_i) = p(Y_i)/q(Y_i)$ are referred to as the ***importance weights***. The intuitive interpretation is that the importance weights correct for the fact that we are sampling from $q$ rather than $p$, since $y$'s that occur less frequently under $q$ than $p$ have large importance weight $w(y)$ (and $y$'s that occur more frequently under $q$ then $p$ have small importance weight).

For the basic IS approximation described above, it is necessary to be able to evaluate $p(x)$ and $q(x)$ in order to compute the importance weights. There is a more general version for which $p(x)$ and $q(x)$ only need to be computable up to constants of proportionality; see Section 3.3.

### 3.1.1 Basic properties

Since this is essentially just a Monte Carlo approximation, it has all the properties described in Section 2:

- consistent, as long as $\mathbb{E}|h(Y)p(Y)/q(Y)| < \infty$, where $Y \sim q$

- unbiased
- the variance of the estimator is $\frac{1}{N}\mathbb{V}\big(h(Y)p(Y)/q(Y)\big)$
- the RMSE is $\sigma\big(h(Y)p(Y)/q(Y)\big)/\sqrt{N}$.

So, the rate of convergence is still of order $1/\sqrt{N}$, however, the constant $\sigma\big(h(Y)p(Y)/q(Y)\big)$ may be smaller or larger than the constant $\sigma\big(h(X)\big)$ of the more direct estimator $\frac{1}{N}\sum h(X_i)$.

### 3.1.2 Choosing the proposal distribution

To minimize the RMSE, we want $q(x)$ to look as much like $h(x)p(x)$ as possible, up to a constant of proportionality. In fact, if we could choose $q(x)$ to be exactly proportional to $h(x)p(x)$, then we would have $h(x)p(x)/q(x) = c$ for some $c$, and this would mean that $\sigma\big(h(Y)p(Y)/q(Y)\big) = 0$, in other words, the error would be zero after only one sample! In this situation, however, there would be no need to resort to sampling at all, since in this case, $\mathbb{E}h(X) = h(x)p(x)/q(x)$ for any $x$, so if we can compute $h(x)p(x)/q(x)$ then we already know $\mathbb{E}h(X)$.

Nonetheless, this indicates that to minimize the approximation error, we want $q(x)$ to be as close as possible to being proportional to $h(x)p(x)$, and it shows that there can be a substantial reduction in the error if a good choice of $q(x)$ is made. When choosing $q$, it is usually better to err on the side of having it a little more "spread out", to make sure that it sufficiently covers the area where $h(x)p(x)$ is large, rather than not covering some of this area (because that would result in occasionally having very large importance weights, which would increase the RMSE of the estimator).

That said, in practice, we often want to estimate $\mathbb{E}h(X)$ for a variety of different functions $h$. Because of this, a common practice is to choose $q(x)$ to be as close as possible to $p(x)$ (rather than $h(x)p(x)$), so that we can reuse the same samples and the same importance weights, and still obtain reasonably good estimators for all of these $h$'s, rather than specializing for each individual $h$.

## 3.2 Example: Marginal likelihood under a non-conjugate prior

In wildlife management and conservation, animals are tagged with GPS devices in order to track their movements and study their behavior. The latitude/longitude measurements made by GPS devices are usually fairly accurate, but it is not uncommon to get extreme outliers. For instance, Figure 3 (Urbano et al., 2014) shows GPS measurements in northern Italy, with three extreme outliers visible—one of which is way down toward central Italy, and another of which is clear across Switzerland and well into France! The Normal (Gaussian) model is not robust to outliers, and if used naively in a situation like this, would give completely bogus results.

One approach to dealing with outliers is to identify and remove them, but this can be somewhat subjective, and can be difficult in high-dimensional settings where the data cannot easily be visualized. Another approach is to use a heavy-tailed distribution instead of the Normal, such as the Laplace distribution, Cauchy distribution, or $t$-distribution. A difficulty that arises, however, is that these distributions do not have nice conjugate priors,

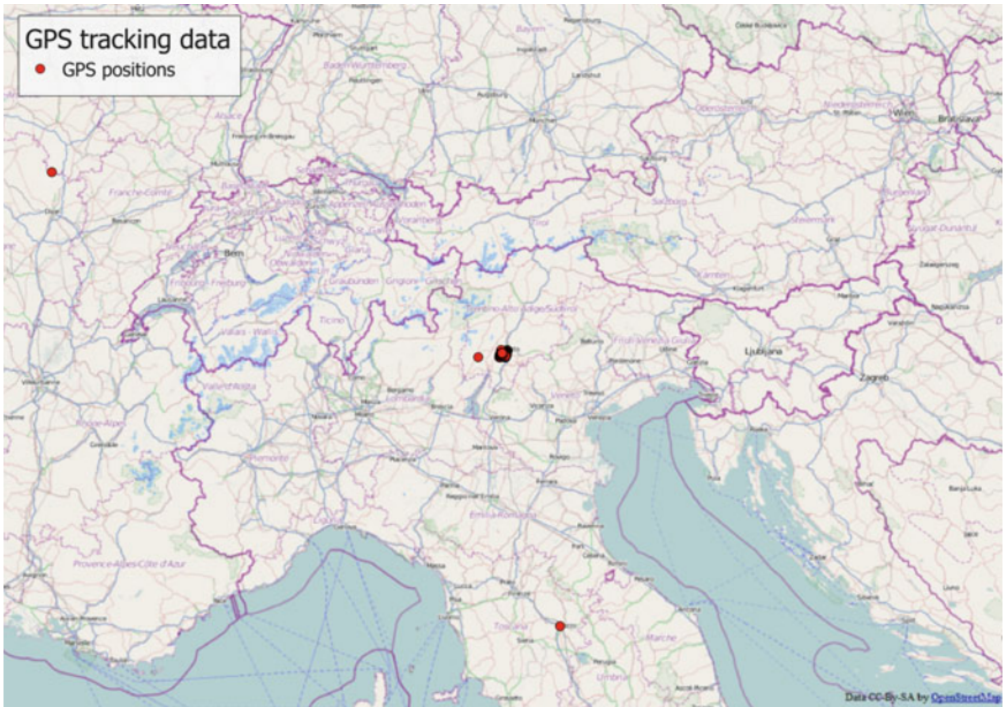

Figure 3: Animal-tracking GPS measurements, with extreme outliers. (Urbano et al., 2014)

so we cannot do inference analytically. One way to do inference in such a situation is using importance sampling.

To illustrate, consider the following 8 latitude/longitude points, one of which is an outlier:

| Latitude | Longitude |
|---|---|
| 36.077916 N | 79.009266 W |
| 36.078032 N | 79.009180 W |
| 36.078129 N | 79.009094 W |
| 36.078048 N | 79.008891 W |
| 36.077942 N | 79.008962 W |
| 36.089612 N | 79.035760 W |
| 36.077789 N | 79.008917 W |
| 36.077563 N | 79.009281 W |

To keep things simple, let's just consider the latitudes, and let's assume these points are collected in a short enough amount of time that the animal has not moved very far. See the histogram of the latitudes in Figure 4. Let's model the latitudes as

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Cauchy}(\theta, s).$$

Recall that the Cauchy distribution with location $\theta$ and scale $s$ has p.d.f.

$$\text{Cauchy}(x \mid \theta, s) = \frac{1}{\pi s\Big(1 + \big(\frac{x-\theta}{s}\big)^2\Big)}.$$

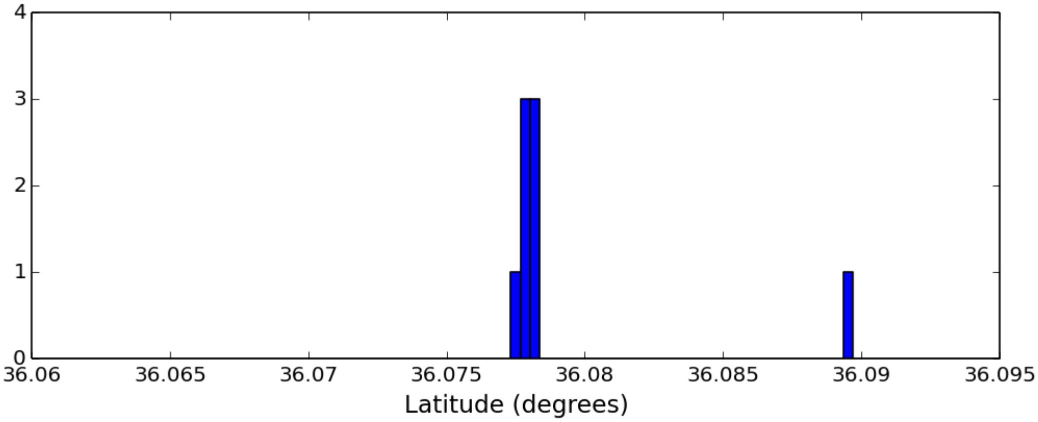

Figure 4: Histogram of latitude measurements.

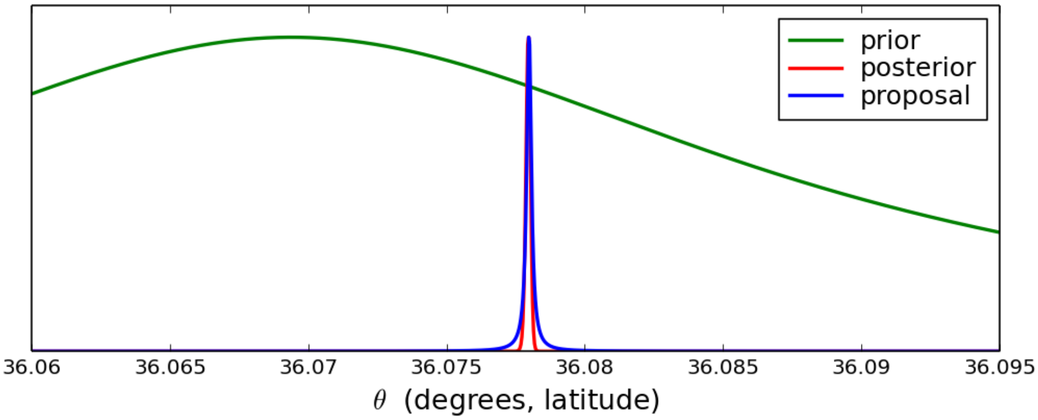

Figure 5: Prior, posterior, and proposal densities. (NOTE: To make them all visible on the same plot, each curve is scaled so that the maximum is 1.)

Unfortunately, there is not a nice conjugate prior for $\theta$. Let's put a Cauchy prior on $\theta$:

$$\boldsymbol{\theta} \sim \text{Cauchy}(\theta_0, s_0).$$

**Parameter settings**

- scale of measurement errors: $s = 0.0002$ degrees (known, say, from calibration testing or instrument specifications)
- center of prior on location: $\theta_0 = 36.07$ degrees (estimated, say, from many previous measurements for this animal)
- scale of prior on location: $s_0 = 0.02$ degrees (estimated, say, from many previous measurements for this animal)

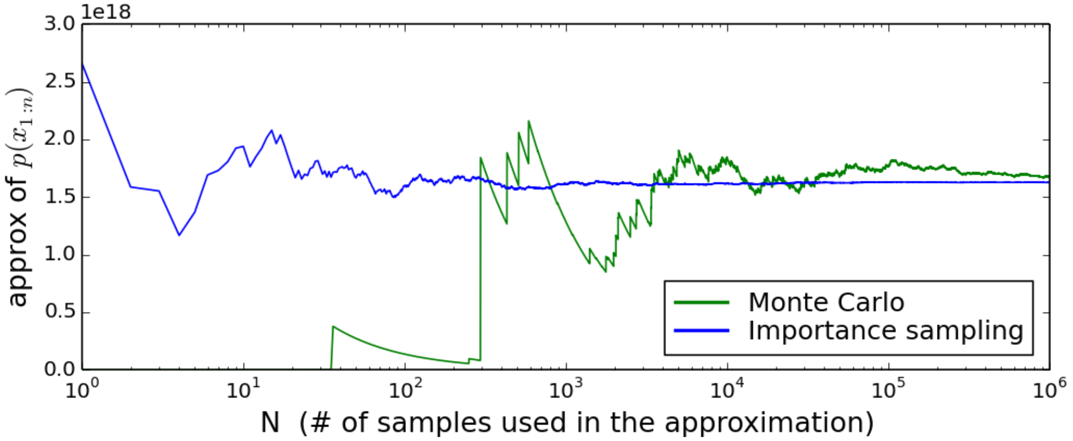

Figure 6: Sequence of approximations for $N = 1, \ldots, 10^6$, for Monte Carlo and importance sampling. Nota bene: The x-axis is on the log scale.

### Approximating the marginal likelihood

Suppose we need to know the marginal likelihood $p(x_{1:n})$. (This is needed, for example, when doing inference over multiple models, as we will see later.) Since we can't compute it analytically (as far as I know), an approximation is needed. One approach would be a simple Monte Carlo approximation:

$$p(x_{1:n}) = \int p(x_{1:n}|\theta)p(\theta)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} p(x_{1:n}|\boldsymbol{\theta}_i) \tag{3.1}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N \overset{\text{iid}}{\sim} p(\theta)$ (i.i.d. from the prior). Although this works asymptotically, it is a very poor approximation (the RMSE is large). We can do much better with importance sampling, for a good choice of $q$:

$$p(x_{1:n}) = \int p(x_{1:n}|\theta)\frac{p(\theta)}{q(\theta)}q(\theta)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} p(x_{1:n}|\boldsymbol{\theta}_i)\frac{p(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)} \tag{3.2}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N \overset{\text{iid}}{\sim} q(\theta)$ (i.i.d. from $q$). As discussed in Section 3.1.2, we want $q(\theta)$ to look as much like $p(x_{1:n}|\theta)p(\theta)$ as possible, and if necessary, to err on the side of being a little more spread out. By cheating (a little bit) and looking at a plot of the posterior, let's choose

$$q(\theta) = \text{Cauchy}(\theta \mid \text{median}(x_{1:n}), 10^{-4}).$$

(A more principled choice could be made based on the rate of convergence of the posterior, but that would be more involved.) See Figure 5 for plots of the prior, posterior, and proposal distribution $q$.

### Results

To visualize the rate of convergence, Figure 6 shows a sequence of Monte Carlo approximations (Equation 3.1) and importance sampling approximations (Equation 3.2) for

$N = 1, \ldots, 10^6$. The IS approximations appear to converge much more quickly, by several orders of magnitude.

Why does this happen? From Figure 5 we can see that the prior is so spread out, compared to the posterior (and thus, the likelihood), that samples from the prior very rarely land in the small region where the likelihood is large. So most of the terms in the Monte Carlo approximation are essentially zero, and a small number of them are enormous (making the variance, and thus the RMSE, large). This situation is typical when approximating the marginal likelihood, since the posterior becomes more and more concentrated as the size of the dataset grows. (In fact, typically the situation is much worse—here, $n$ is only eight!)

A general word of caution: Approximating the marginal likelihood can be a tricky business, and one needs to be careful when going about it.

## 3.3 Handling unknown normalization constants

As described so far, in order to implement an importance sampling approximation, it looks like we need to be able to compute the p.d.f./p.m.f. values $p(Y_i)$ and $q(Y_i)$. In many cases, it is not possible to compute $p$ and $q$ themselves, but often, we will be able to compute functions $\tilde{p}$ and $\tilde{q}$ proportional to $p$ and $q$. Fortunately, there is a neat little trick that still allows us to make an IS approximation. Suppose

$$p(x) = \tilde{p}(x)/Z_p$$
$$q(x) = \tilde{q}(x)/Z_q$$

where $\tilde{p}(x)$ and $\tilde{q}(x)$ are easy to compute (but $Z_p$ and $Z_q$ may be intractable). Also, rather than assuming $q(x) > 0$ for all $x$, let us assume only that $\tilde{q}(x) > 0$ whenever $\tilde{p}(x) > 0$. Define

$$\tilde{w}(x) = \begin{cases} \tilde{p}(x)/\tilde{q}(x) & \text{if } \tilde{q}(x) > 0 \\ 0 & \text{if } \tilde{q}(x) = 0. \end{cases}$$

The general form of an importance sampling approximation is then

$$\mathbb{E}h(X) = \int h(x)p(x)dx \approx \frac{\frac{1}{N}\sum_{i=1}^N h(Y_i)\tilde{w}(Y_i)}{\frac{1}{N}\sum_{i=1}^N \tilde{w}(Y_i)} = \sum_{i=1}^N h(Y_i)\left(\frac{\tilde{w}(Y_i)}{\sum_{j=1}^N \tilde{w}(Y_j)}\right) \tag{3.3}$$

where $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} q$. (See derivation in Section 3.3.2). This can be interpreted as a weighted average of the $h(Y_i)$'s, with weights $\tilde{w}(Y_i)/\sum_j \tilde{w}(Y_j)$.

### 3.3.1 Example: Approximating posterior expectations

Consider the animal-tracking GPS example from Section 3.2, and now suppose we would like to estimate the posterior mean. Define

$$\begin{aligned} &\pi(\theta) = p(\theta|x_{1:n}) \\ &\tilde{\pi}(\theta) = p(x_{1:n}|\theta)p(\theta) \\ &Z_\pi = p(x_{1:n}) = \int p(x_{1:n}|\theta)p(\theta)d\theta = \int \tilde{\pi}(\theta)d\theta. \end{aligned}$$

Then $\pi(\theta) = \tilde{\pi}(\theta)/Z_\pi$, and using Equation 3.3,

$$\mathbb{E}(\boldsymbol{\theta}|x_{1:n}) = \int \theta p(\theta|x_{1:n})d\theta = \int \theta\pi(\theta)d\theta \approx \frac{\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\theta}_i\tilde{w}(\boldsymbol{\theta}_i)}{\frac{1}{N}\sum_{i=1}^{N}\tilde{w}(\boldsymbol{\theta}_i)}$$

where $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N \overset{\text{iid}}{\sim} q$ and $\tilde{w}(\theta) = \tilde{\pi}(\theta)/q(\theta)$. For the data from Section 3.2, with $N = 10^6$, this yields an approximate value of $\mathbb{E}(\boldsymbol{\theta}|x_{1:n}) \approx 36.0780$. Note that, in this example, since $Z_q$ is effectively 1, the denominator in this approximation is identical to the IS approximation of $Z_\pi$ that we used in Section 3.2. More generally, however, the denominator will be an approximation of the ratio of normalization constants (see Section 3.3.2).

### 3.3.2 Derivation of Equation 3.3

Letting $S = \{x : \tilde{p}(x) > 0\}$, we have

$$\begin{aligned}
\mathbb{E}h(X) &= \int h(x)p(x)dx \\
&= \int_S h(x)\frac{\tilde{p}(x)}{Z_p}dx \\
&\overset{\text{(a)}}{=} \int_S h(x)\frac{\tilde{p}(x)}{Z_p}\frac{Z_q}{\tilde{q}(x)}q(x)dx \\
&= \frac{Z_q}{Z_p}\int_S h(x)\tilde{w}(x)q(x)dx \\
&\overset{\text{(b)}}{=} \frac{Z_q}{Z_p}\int h(x)\tilde{w}(x)q(x)dx \\
&\overset{\text{(c)}}{\approx} \frac{Z_q}{Z_p}\frac{1}{N}\sum_{i=1}^{N} h(Y_i)\tilde{w}(Y_i)
\end{aligned}$$

where in step (a), we use the assumption that $\tilde{q}(x) > 0$ whenever $\tilde{p}(x) > 0$, in step (b), we use the fact that $\tilde{w}(x) = 0$ for any $x \notin S$, and step (c) is a simple Monte Carlo approximation. Similarly, for the ratio of normalizing constants $Z_p/Z_q$,

$$\begin{aligned}
\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \tilde{p}(x)dx \\
&= \frac{1}{Z_q}\int_S \tilde{p}(x)dx \\
&= \frac{1}{Z_q}\int_S \frac{\tilde{p}(x)}{\tilde{q}(x)}\tilde{q}(x)dx \\
&= \int_S \frac{\tilde{p}(x)}{\tilde{q}(x)}q(x)dx \\
&= \int \tilde{w}(x)q(x)dx \\
&\approx \frac{1}{N}\sum_{i=1}^{N}\tilde{w}(Y_i).
\end{aligned}$$

# 4 Basic techniques for generating samples

## 4.1 Inverse c.d.f. method / Smirnov transform

The ***inverse c.d.f. method*** or ***Smirnov transform*** is a common way of generating random samples from univariate probability distributions when the inverse of the c.d.f. can be easily computed. The basic idea is that if $U \sim \text{Uniform}(0,1)$ and $G$ is the inverse (in a generalized sense) of the c.d.f. $F$, then $G(U)$ is a random variable with c.d.f. $F$.

### 4.1.1 Example: Sampling from $\text{Exp}(\theta)$

The $\text{Exp}(\theta)$ c.d.f. is $F(x) = (1 - e^{-\theta x})\mathbb{1}(x > 0)$. This is invertible on $(0, \infty)$, with inverse $G(u) = -(1/\theta)\log(1-u)$ for $u \in (0,1)$. Therefore, if $U \sim \text{Uniform}(0,1)$ then $G(U) \sim \text{Exp}(\theta)$.

### 4.1.2 A precise statement of the method

**Proposition 4.1.** *Let $F$ be a c.d.f. on $\mathbb{R}$, and define $G(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$ for $u \in (0,1)$. If $U \sim \text{Uniform}(0,1)$, then $G(U) \sim F$.*

See Appendix for proof.

**Remarks**

- If $F$ is invertible, then $G = F^{-1}$. The definition of $G$ above is a generalized inverse that allows for the possibility that $F$ has discontinuities and/or is constant on certain intervals.
- $G(u) \in \mathbb{R}$ for any $u \in (0,1)$ since $F(x) \to 1$ as $x \to \infty$, and $F(x) \to 0$ as $x \to -\infty$.
- By convention, $F$ is assumed to be continuous from the right, rather than the left.

## 4.2 Rejection sampling

Rejection sampling is a method for drawing random samples from a distribution whose p.d.f. can be evaluated up to a constant of proportionality. Compared with the inverse c.d.f. method, rejection sampling has the advantage of working on complicated multivariate distributions, however, one has to design a good proposal distribution (which can be difficult, especially in high-dimensional settings).

The method relies on two principles:

1. **The rejection principle:** Rejecting results in conditional samples. That is, if we reject any samples falling outside of a given set, the remaining samples are distributed according to the conditional distribution on that set.
2. **The projection principle:** A distribution equals the projection of the uniform distribution under its p.d.f. That is, if we sample uniformly from the region under the p.d.f. (or a function proportional to it) of a distribution, and discard the "height", we obtain a sample from that distribution.

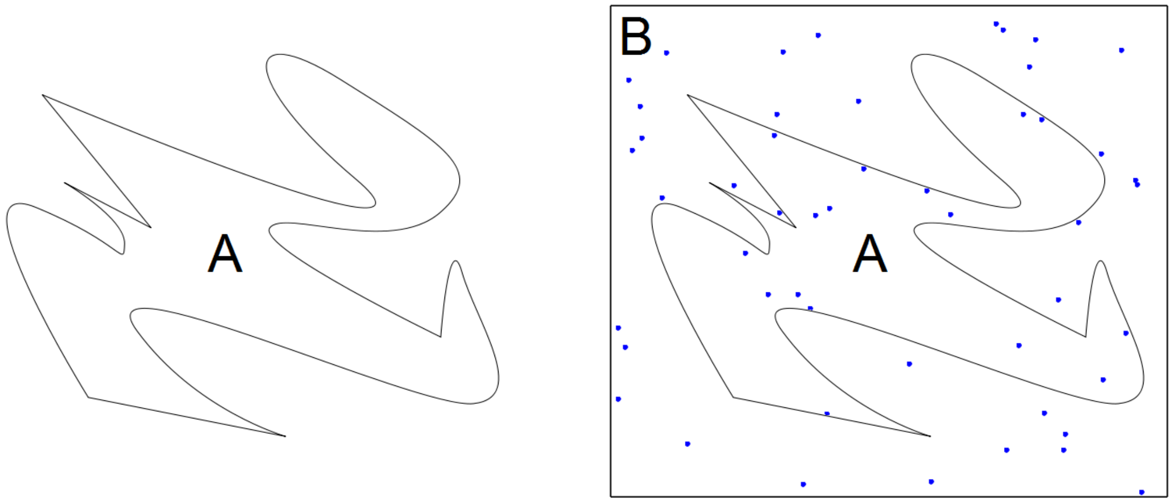

Figure 7: (Left) How to draw uniform samples from region $A$? (Right) Draw uniform samples from $B$ and keep only those that are in $A$.

### 4.2.1 The rejection principle

Consider the oddly-shaped two-dimensional region $A$ in Figure 7. How would you generate samples uniformly over $A$? In other words, how would you generate samples from $\text{Uniform}(A)$, the uniform distribution on $A$? The rejection sampling approach is to choose a simpler region $B$ containing $A$, generate samples from $\text{Uniform}(B)$, and keep only the samples that land inside $A$. The name "rejection sampling" comes from the fact that we discard, or "reject", any samples landing outside $A$. It turns out that this is guaranteed to generate samples from $\text{Uniform}(A)$. In fact, more generally, rejecting results in samples from the conditional distribution on $A$; see Proposition .1.

In order to be as efficient as possible, we want the bounding region $B$ to be as small as possible (while still containing $A$), so that the number of rejections is kept to a minimum.

### 4.2.2 The projection principle

Suppose we want to sample from a distribution on $\mathbb{R}^d$ with p.d.f. $\pi(x) = \tilde{\pi}(x)/Z_\pi$. Consider the region of $\mathbb{R}^{d+1}$ under $\tilde{\pi}$:

$$A = \Big\{(x, y) : x \in \mathbb{R}^d,\, 0 < y < \tilde{\pi}(x)\Big\}.$$

It turns out that if $(X, Y) \sim \text{Uniform}(A)$ (that is, $(X, Y)$ is uniformly distributed over $A$), then $X \sim \pi$.

To see why, first note that the volume of $A$ is

$$\text{Vol}(A) = \int \tilde{\pi}(x)dx = \int Z_\pi \pi(x)dx = Z_\pi,$$

and since the p.d.f. of the uniform distribution on $A$ is constant, we have

$$p(x, y) = \text{Uniform}(x, y \mid A) = \frac{\mathbb{1}\big((x, y) \in A\big)}{\text{Vol}(A)} = \frac{\mathbb{1}\big(0 < y < \tilde{\pi}(x)\big)}{Z_\pi}.$$

Therefore,

$$\begin{aligned}
p(x) &= \int_{-\infty}^{\infty} p(x,y)dy = \int_{-\infty}^{\infty} \frac{\mathbb{1}\big(0 < y < \tilde{\pi}(x)\big)}{Z_\pi} dy \\
&= \frac{1}{Z_\pi} \int_0^{\tilde{\pi}(x)} dy = \frac{\tilde{\pi}(x)}{Z_\pi} = \pi(x).
\end{aligned}$$

### 4.2.3 The rejection sampling procedure

Combining these two principles leads to the following procedure. Suppose we want to draw samples from a distribution on $\mathbb{R}^d$ with p.d.f. $p(x) \propto \tilde{p}(x)$.

- Choose a proposal distribution $q$ that is easy to sample from, and is as close as possible to being proportional to $\tilde{p}$.
- Choose $c > 0$ such that $cq(x) \geq \tilde{p}(x)$ for all $x$.

To draw a sample from $p$:

1. Sample $X \sim q$.
2. Sample $Y \sim \text{Uniform}(0, cq(X))$.
3. If $Y \geq \tilde{p}(X)$, then go back to step 1.
4. Otherwise, output $X$ as a sample.

Then the accepted $X$'s are distributed according to $p$. To see why, let

$$\begin{aligned}
A &= \Big\{(x,y) : x \in \mathbb{R}^d,\, 0 < y < \tilde{p}(x)\Big\} \\
B &= \Big\{(x,y) : x \in \mathbb{R}^d,\, 0 < y < cq(x)\Big\}
\end{aligned}$$

and note that

- steps 1 and 2 generate a sample $(X,Y)$ uniformly from $B$, since their joint density is

$$q(x)\,\text{Uniform}\big(y \mid (0,\, cq(x))\big) = q(x)\frac{\mathbb{1}\big(0 < y < cq(x)\big)}{cq(x)} = \frac{\mathbb{1}\big((x,y) \in B\big)}{c},$$

- step 3 rejects any pairs $(X,Y)$ that are outside of $A$, so that the distribution of accepted pairs is uniform on $A$ (by the rejection principle), and
- step 4 keeps only $X$, resulting in a sample from $p$ (by the projection principle).

# 5 Exercises

1. The Gumbel distribution with location $c \in \mathbb{R}$ and scale $\beta > 0$ has c.d.f.

   $$F(x \mid c, \beta) = \exp\big( - e^{-(x-c)/\beta}\big).$$

   This distribution has certain special properties that make it well-suited for modeling extreme values, and it is often used in hydrology as a model for measurements such as the maximum annual water level of a river, or maximum monthly rainfall in a region. Use the inverse c.d.f. method to derive a procedure for generating Gumbel$(c, \beta)$ random variables from Uniform$(0, 1)$ random variables.

2. When $X \sim$ Cauchy$(0, 1)$, we have $\mathbb{E}|X| = \infty$, and thus the Monte Carlo approximations $\frac{1}{N}\sum_{i=1}^{N} X_i$ are not guaranteed to converge as $N \to \infty$, when $X_1, X_2, \ldots \overset{\text{iid}}{\sim}$ Cauchy$(0, 1)$ (indeed, the mean $\mathbb{E}X$ does not even exist for the Cauchy distribution). Explore what happens empirically by sampling $X_1, \ldots, X_M \overset{\text{iid}}{\sim}$ Cauchy$(0, 1)$ for $M = 10^6$ and plotting the sequence of Monte Carlo approximations for $N = 1, \ldots, M$. Do this for several sets of samples $X_1, \ldots, X_M$, and pick four representative examples to display in separate plots. Discuss what you see.

3. The "harmonic mean approximation" of the marginal likelihood is

   $$p(x_{1:n}) \approx \frac{1}{\frac{1}{N}\sum_{i=1}^{N} 1/p(x_{1:n}|\boldsymbol{\theta}_i)}$$

   where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N \overset{\text{iid}}{\sim} p(\theta|x_{1:n})$ (that is, they are i.i.d. from the posterior).

   (a) Show that, in principle, this converges to the marginal likelihood $p(x_{1:n})$. Assume that $p(x_{1:n}|\theta) > 0$ for all $\theta$.

   (b) Consider the following simple example with $n = 1$: $X_1 \sim \mathcal{N}(\theta, \lambda^{-1})$ with $\lambda = 1$, and $\boldsymbol{\theta} \sim \mathcal{N}(0, \lambda_0^{-1})$ with $\lambda_0 = 1/10^2$. Compute the harmonic mean approximation for $p(x_1)$ when $x_1 = 2$, using $N = 10^6$. Report the result for 5 independent sets of samples $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ from the posterior. Compare these with the true value of the marginal likelihood, $\mathcal{N}(2 \mid 0, \lambda^{-1} + \lambda_0^{-1})$. Describe what you observe. (This example is due to Neal, 2008).

   (c) Repeat part (b) using $\lambda_0 = 1/100^2$.

   The harmonic mean approximation was fairly popular for a while, since it is so easy to compute from posterior samples (which we will often have from running MCMC, stay tuned), however, it can have extremely poor performance. Importance sampling, and related methods such as path sampling, are far better.

4. Implement both the Monte Carlo and importance sampling approximations to the marginal likelihood for the GPS example in Section 3.2. Create a plot like Figure 6, to visualize the convergence of the approximations.

5. Suppose $A \subset \mathbb{R}^d$ and $X, X_1, X_2, \ldots \in \mathbb{R}^d$ are i.i.d. with $\mathbb{P}(X \in A) > 0$. Show that if $Z = X_K$ where $K = \min\{k : X_k \in A\}$, then $Z \overset{D}{=} (X \mid X \in A)$, that is, $Z$ has the same distribution as $X \mid X \in A$. Do this by showing that $\mathbb{P}(Z \in S) = \mathbb{P}(X \in S \mid X \in A)$ for any $S \subset A$. (Hint: $\sum_{k=0}^{\infty} a^k = 1/(1-a)$ for $a \in [0, 1)$.)

## Supplementary material

- Hoff (2009), 4.1 and 4.2.
- mathematicalmonk videos, Machine Learning (ML) 17.1–17.14 https://www.youtube.com/playlist?list=PLD0F06AA0D2E8FFBA

## References

- Urbano, F., Basille, M., and Cagnacci, F. Data Quality: Detection and Management of Outliers. Spatial Database for GPS Wildlife Tracking Data. Springer International Publishing, 2014. 115-137.
- Radford Neal (2008), The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever. https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-l

## Proofs

*Proof of Proposition 4.1.* First, we show that for any $u \in (0, 1)$, $x \in \mathbb{R}$, we have

$$u \leq F(x) \text{ if and only if } G(u) \leq x. \tag{.1}$$

If $u \leq F(x)$ then

$$G(u) = \inf\{y \in \mathbb{R} : u \leq F(y)\} \leq x$$

because $x \in \{y \in \mathbb{R} : u \leq F(y)\}$. On the other hand, suppose $G(u) \leq x$. Then there exist $x_1 \geq x_2 \geq \cdots$ such that $x_n \to G(u)$ and $u \leq F(x_n)$ for all $n$. Thus,

$$u \leq \liminf F(x_n) \overset{\text{(a)}}{=} F(G(u)) \overset{\text{(b)}}{\leq} F(x)$$

where (a) is because $F$ is continuous from the right, and (b) is because $F$ is monotone increasing. This proves Equation .1.

Therefore, for any $x \in \mathbb{R}$,

$$\mathbb{P}(G(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

Hence, $G(U)$ has c.d.f. $F$. □

**Proposition .1.** *Suppose $A \subset \mathbb{R}^d$ and $X, X_1, X_2, \ldots \in \mathbb{R}^d$ are i.i.d. with $\mathbb{P}(X \in A) > 0$. If $Z = X_K$ where $K = \min\{k : X_k \in A\}$, then $Z \overset{D}{=} (X \mid X \in A)$, that is, $Z$ has the same distribution as $X \mid X \in A$.*

*Proof.* (Proof omitted temporarily since it is an exercise.) □