Computer Intensive Statistics and Applications
Chapter 2: Monte Carlo Integration

Shaobo Jin

Department of Mathematics

Intractable Integral in Statistics

Many quantities of interest in statistics can be formulated as integrals such as

$$\mu = \int_{x \in D} f(x)\, dx < \infty.$$

But no closed form expressions are available and we want to approximate it. Some easy examples are

1. We want to approximate the probability

$$\mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(x)\, dx.$$

2. We want to approximate the expectation

$$\mathrm{E}[X] = \int x f(x)\, dx.$$

More Intractable Integral in Statistics

1. In engineering, we consider a system such that the strength of the system is a random variable S and the load is a random variable L. The system fails is $S < L$. By the law of total probability, the probability of failure is

$$\mathrm{P}(S < L) = \int \mathrm{P}(s < L) f_S(s) \, ds.$$

2. In Bayesian statistics, we want to obtain the predictive distribution to predict the future value

$$p(y_{\text{new}} \mid y) = \int p(y_{\text{new}} \mid \theta) \pi(\theta \mid y) \, d\theta.$$

Convergence in Distribution

Let $X \in \mathbb{R}^d$ be a $d \times 1$ random vector of random variables. Its distribution function is $F_X(x) = P(X_1 \leq x_1, ..., X_d \leq x_d)$.

Definition (Convergence in Law/Distribution, Weak Convergence)

X_n converges in law (converges in distribution or converges weakly) to X if $F_{X_n}(x) \to F_X(x)$ as $n \to \infty$ for all points x at which $F_X(x)$ is continuous. It is denoted by $X_n \xrightarrow{d} X$ or $X_n \xrightarrow{\mathcal{L}} X$.

Convergence in Probability and Almost Surely

Definition (Convergence in probability)

X_n converges in probability to X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}\left((X_n - X)^T (X_n - X) > \epsilon^2\right) = 0.$$

It is denoted by $X_n \xrightarrow{P} X$. If X is a constant, then we also say X_n is consistent for X.

Definition (Convergence almost surely)

X_n converges almost surely to X if

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

It is denoted by $X_n \xrightarrow{a.s.} X$.

Continuous Mapping Theorem

Theorem (Continuous Mapping Theorem)

Let $X_1, X_2, \ldots$ be a sequence of random vectors. Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $P(X \in C) = 1$.

1. *If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.*
2. *If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.*
3. *If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.*

Big O and Small O Operator

Definition

Let $X_1, X_2, \ldots$ be a sequence of random vectors.

1. The sequence is bounded in probability, if for any $\epsilon > 0$ there exists $M > 0$, such that

$$\mathrm{P}\left(X_n^T X_n \leq M^2\right) \quad > \quad 1 - \epsilon, \text{ for all } n.$$

   It is denoted by $X_n = O_{\mathrm{P}}(1)$.
2. If the sequence converges in probability to zero, it is denoted by $X_n \xrightarrow{P} 0$ or $X_n = o_{\mathrm{P}}(1)$.
3. For a sequence of positive numbers a_n, we say $X_n = O_{\mathrm{P}}(a_n)$ if $X_n/a_n = O_{\mathrm{P}}(1)$.
4. For a sequence of positive numbers a_n, we say $X_n = o_{\mathrm{P}}(a_n)$ if $X_n/a_n = o_{\mathrm{P}}(1)$.

Big O: Example

Example

Let $X_1, ..., X_n$ be iid from $N\left(\mu, \sigma^2\right)$ with finite variance.

1. Show that $\sum_{i=1}^{n} X_i = O_{\mathrm{P}}(n)$.
2. Show that $\sum_{i=1}^{n} (X_i - \mu) = o_{\mathrm{P}}(n)$.

Law of Large Numbers

Theorem (Law of Large Numbers)

Let $X_1, X_2, \ldots$ be iid random vectors, and $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$.

1. *(weak law) If $E\left[\sqrt{X_i^T X_i}\right] < \infty$, then $\bar{X}_n$ converges in probability to $\mu = E[X_i]$.*
2. *(strong law) $E\left[\sqrt{X_i^T X_i}\right] < \infty$ and $\mu = E[X_i]$ if and only if $\bar{X}_n$ converges almost surely to μ.*

Central Limit Theorem

Theorem (Lindeberg-Lévy Central Limit Theorem (CLT))

Let $X_1, X_2, \ldots$ be iid random vectors, with mean μ and finite covariance matrix Σ. Then, $\sqrt{n}\left(\bar{X}_n - \mu\right)$ converges in distribution to a normal distribution $N(0, \Sigma)$, i.e., $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \Sigma)$.

Delta Method

Theorem (Delta Method)

Let $g : \mathbb{R}^k \to \mathbb{R}$ *be a mapping such that* $\frac{\partial g(x)}{\partial x}$ *is continuous in a neighborhood of* $\mu \in \mathbb{R}$. *Let* X_1, X_2, ... *be a sequence of random vectors, such that* $\sqrt{n}\left(X_n - \mu\right) \xrightarrow{d} N\left(0, \Sigma\right)$. *Then,*

$$\sqrt{n}\left(g\left(X_n\right) - g\left(\mu\right)\right) \quad \xrightarrow{d} \quad N\left(0, \left(\frac{\partial g\left(x\right)}{\partial x^T}\right) \Sigma \left(\frac{\partial g\left(x\right)}{\partial x^T}\right)^T\right).$$

Monte Carlo Approximation

Suppose that we want to evaluate

$$\mu = \int_{x \in D} f(x)\, dx,$$

where $x \in \mathbb{R}^d$, but the closed form expression is hard to obtain. We always assume $\mu < \infty$ without stating it.

Algorithm 1: Independent Monte Carlo

---

1 Find a suitable factorization $f(x) = h(x)\, p(x)$ such that $p(x)$ is a density ;
2 Treat μ as $\mu = \mathrm{E}[h(x)]$;
3 Simulate n iid random numbers from $p(x)$;
4 Approximate μ by

$$\hat{\mu}^{\mathrm{IMC}} = \frac{1}{n} \sum_{i=1}^{n} h(x_i).$$

Why Does Independent MC Work?

Suppose that $0 < \mathrm{Var}\left[h\left(X\right)\right] < \infty$. Then,

1. $\mathrm{E}\left[\hat{\mu}^{\mathrm{IMC}}\right] = \int h\left(x\right) p\left(x\right) dx = \mu$, unbiased.
2. By the Law of Large Numbers, $\hat{\mu}^{\mathrm{IMC}} - \mu = o_{\mathrm{P}}\left(1\right)$, and $\hat{\mu}^{\mathrm{IMC}} \overset{a.s.}{\to} \mu$.
3. By the Central Limit Theorem,

$$\sqrt{n}\left(\hat{\mu}^{\mathrm{IMC}} - \mu\right) \quad \overset{d}{\to} \quad N\left(0, \mathrm{Var}\left[h\left(X\right)\right]\right),$$

and

$$\hat{\mu}^{\mathrm{IMC}} - \mu \quad = \quad O_{\mathrm{P}}\left(n^{-1/2}\right),$$

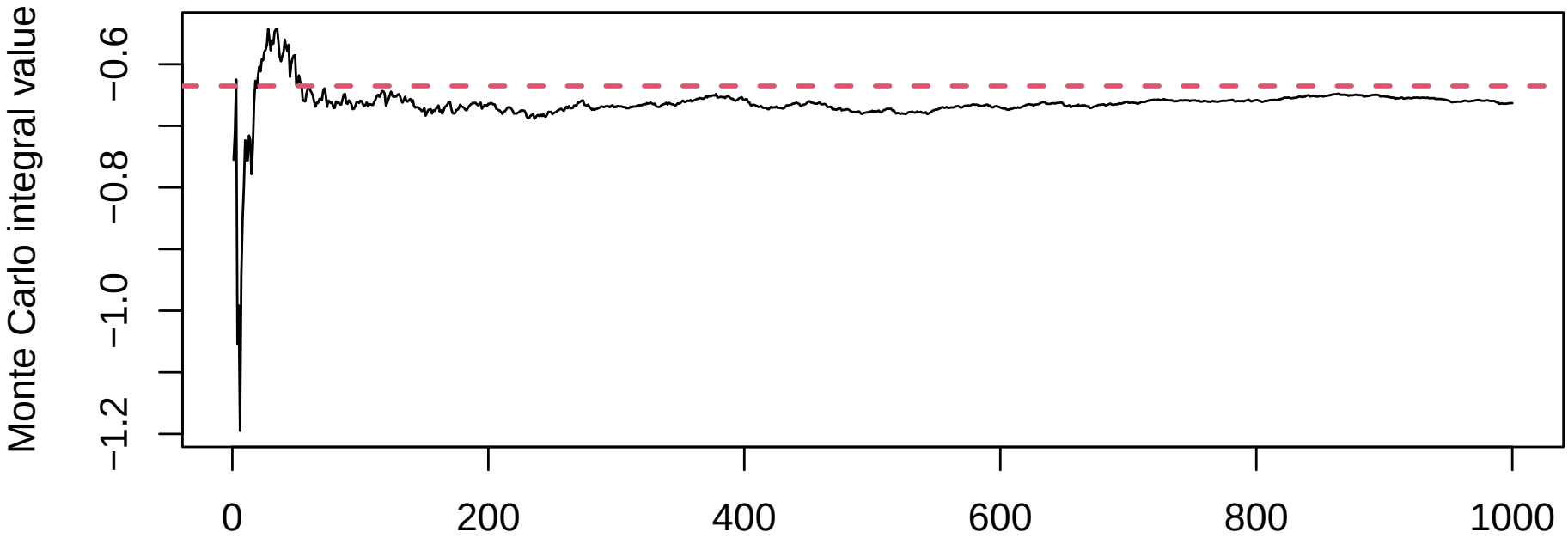
independent of dimension of integral!

Monte Carlo Integral: Example

Example

Approximate the integral

$$\mu = \int_0^\infty \log(x) \exp(-2x)\, dx.$$

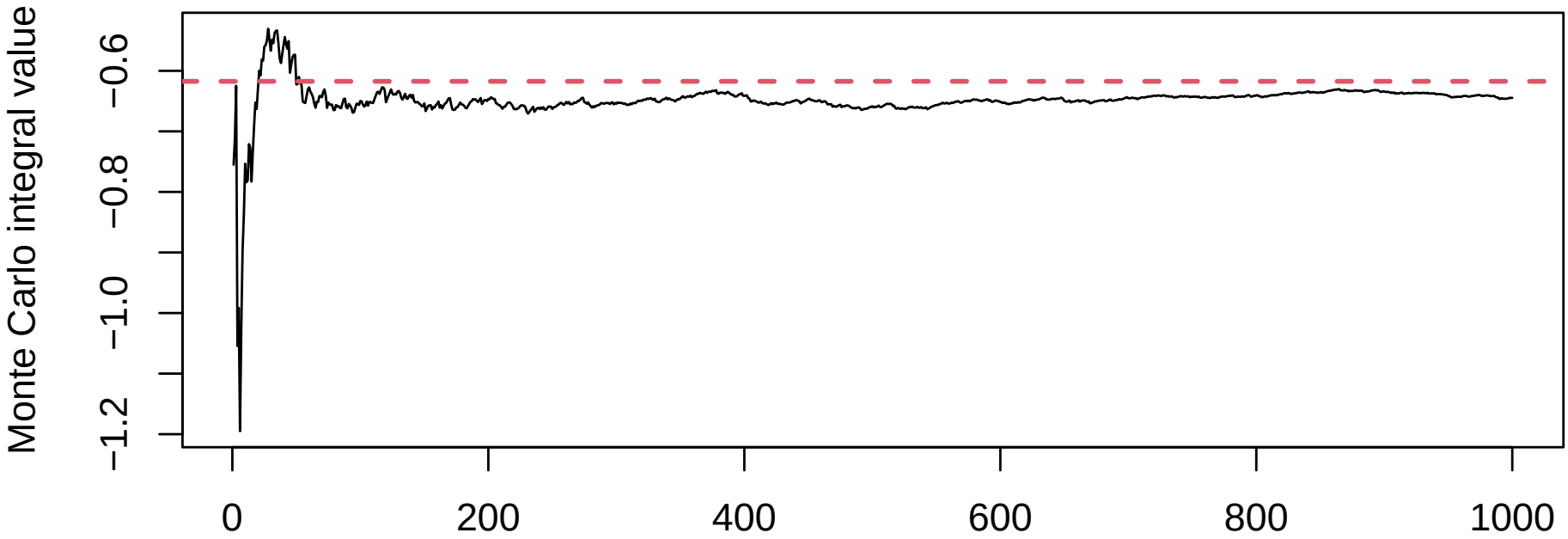


Monte Carlo Integral: Example

Example

Approximate the integral

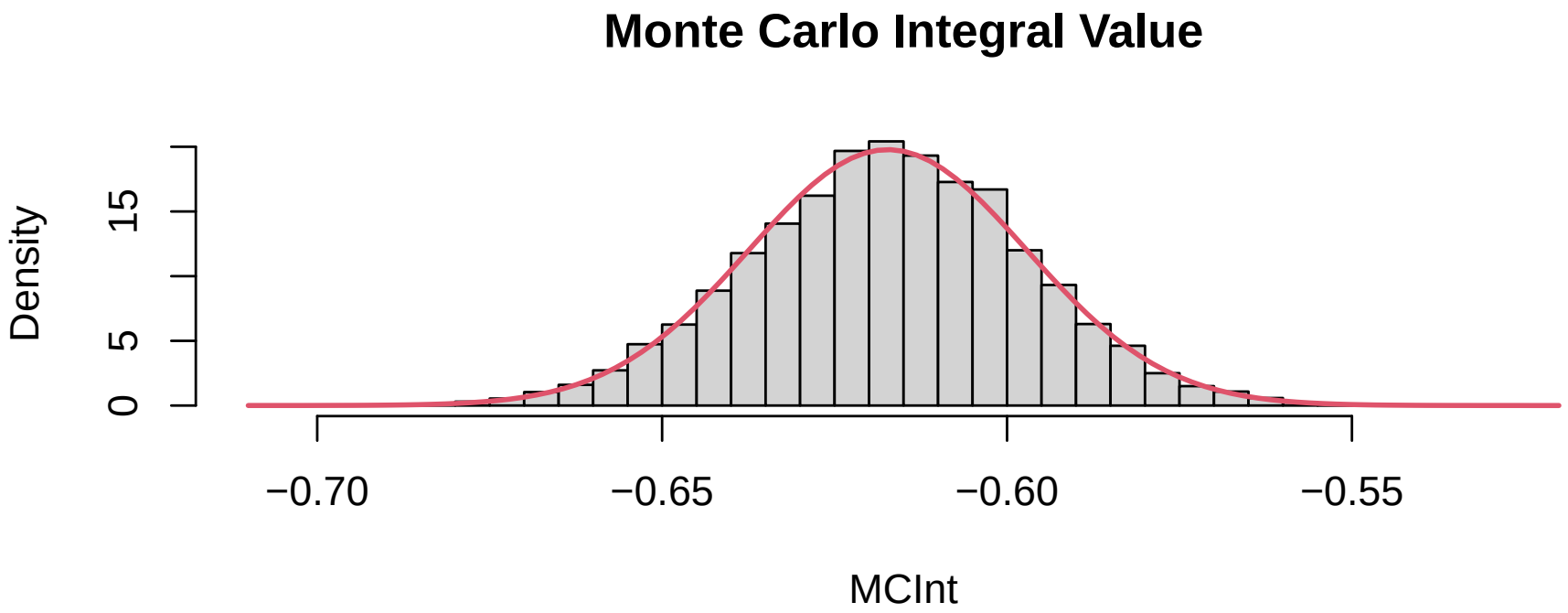
$$\mu = \int_0^{0.5} \log(x) \exp(-2x)\, dx.$$



Randomness

Since we simulate random numbers from a distribution with density $p(x)$, $\hat{\mu}^{\mathrm{IMC}}$ is a random variable.

- For a large enough n, the distribution of $\hat{\mu}^{\mathrm{IMC}}$ can be approximated by a normal distribution.
- This means that we can construct confidence interval for $\hat{\mu}^{\mathrm{IMC}}$.

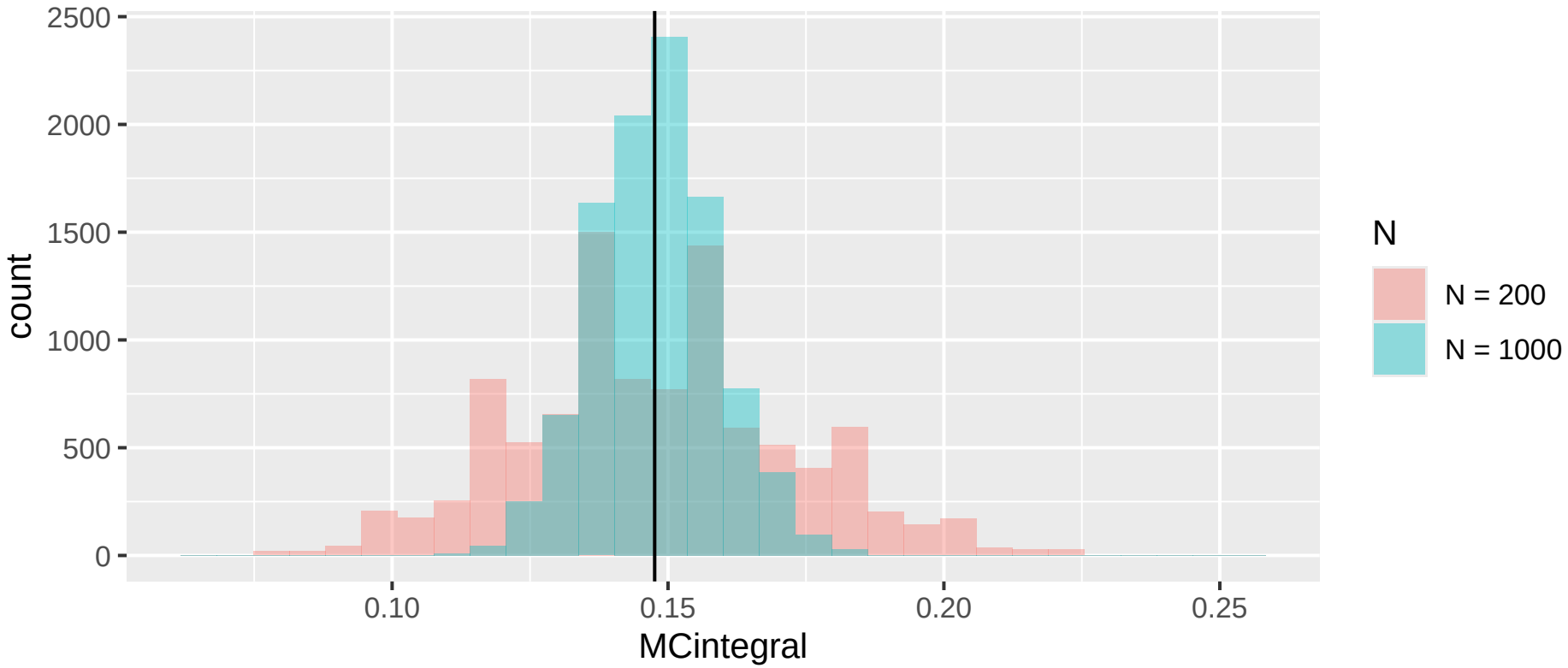


Example: Effect of n

We want to approximate $\mathrm{P}(X > 2)$, where $X \sim \text{Cauchy}(0, 1)$. That is,

$$\mu = \int_{-\infty}^{\infty} 1_{(2,\infty)}(x)\, p(x)\, dx,$$

where $p(x)$ is the density of Cauchy $(0, 1)$.



Remarks: Effects of h and p

$$\mu = \int_D f(x)\,dx < \infty,$$

1. By the Central Limit Theorem,

$$\sqrt{n}\left(\hat{\mu}^{\mathrm{IMC}} - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right).$$

   Hence, we want σ^2 to be small without the need of choosing a huge n.

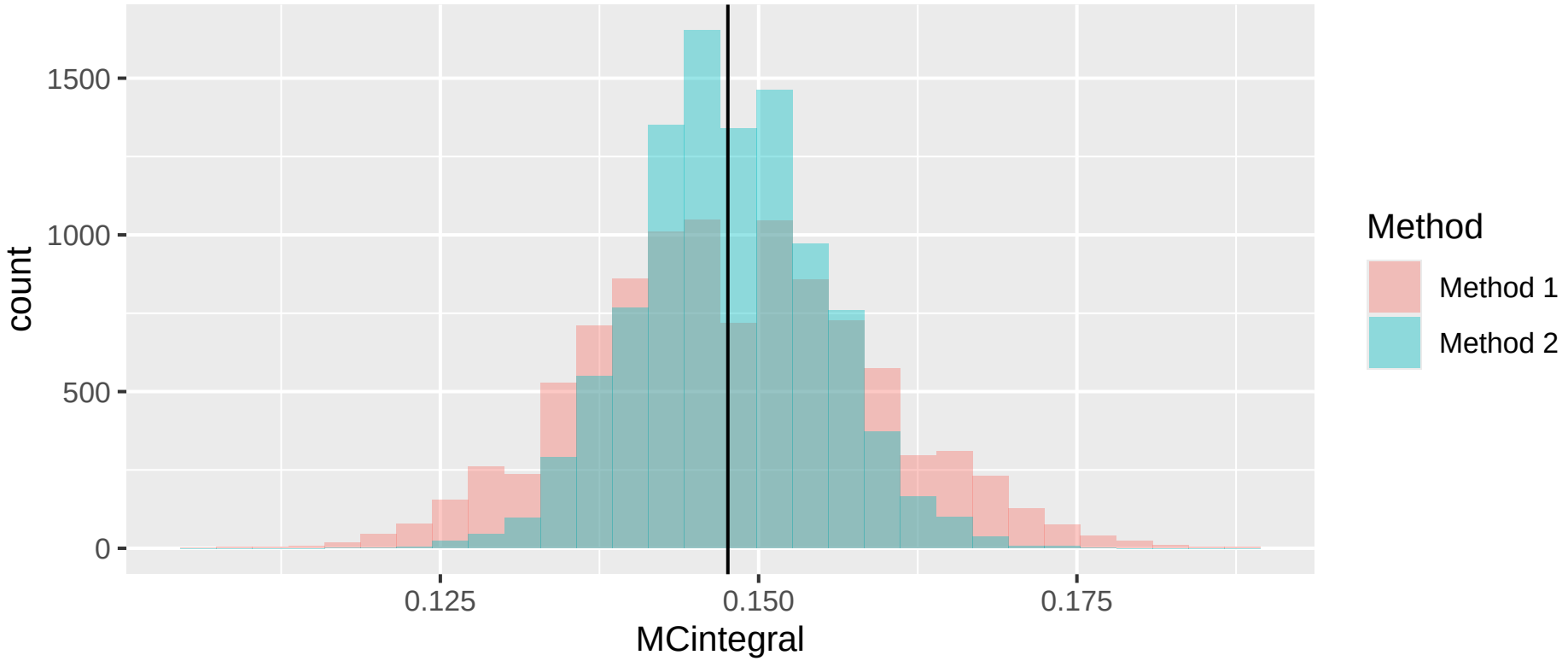
2. Since $f(x) = h(x)\,p(x)$, we want to pick a $p(x)$ such that we can easily simulate random numbers from $p(x)$.

Example: Effect of $h(x)$

To approximate the $\mathrm{P}(X > 2)$, where $X \sim \text{Cauchy}(0, 1)$,

1. Method 1: $\mu = \int_{-\infty}^{\infty} 1_{(2,\infty)}(x)\, p(x)\, dx$,
2. Method 2: $\mu = \int_{-\infty}^{\infty} 2^{-1} 1_{(2,\infty)}(|x|)\, p(x)\, dx$ using symmetry.

They have the same $p(x)$ but different $h(x)$.



Importance Sampling

It is not always the case that we can recognize a $p(x)$ as a density. But we can choose a $g(x)$ such that we can easily sample random numbers from and rewrite

$$\mu = \int f(x)\,dx = \int \frac{f(x)}{g(x)} g(x)\,dx = \mathrm{E}\left[\frac{f(X)}{g(X)}\right],$$

where the density of $X \in \mathbb{R}^d$ is $g(x)$.

Algorithm 2: Importance sampling

1 Simulate n iid random numbers from the trial distribution $g(x)$;
2 Approximate μ by

$$\hat{\mu}^{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)},$$

provided that the ratio f/g can be computed.

Importance Sampling: Example

Example

1. Suppose that we want to approximate the integral

$$\mu = \int_0^\infty \log(1+x)\, x \exp(-4x)\, dx.$$

We sample X from $\text{Exp}(1)$.

2. We want to approximate $\text{P}(X > 2)$ by importance sampling, where $X \sim \text{Cauchy}(0,1)$. That is,

$$\mu = \int_{-\infty}^\infty 1_{(2,\infty)}(x)\, p(x)\, dx,$$

where $p(x)$ is the density of $\text{Cauchy}(0,1)$.

A Property

Theorem

Let μ be approximated by importance sampling where X is sampled from a distribution with density $g(x)$. Suppose that $g(x) > 0$ whenever $f(x) \neq 0$. Then, $E\left[\hat{\mu}^{IS}\right] = \mu$ (unbiased), $\hat{\mu}^{IS} \xrightarrow{P} \mu$ (consistent), and $Var\left(\hat{\mu}^{IS}\right) = \sigma_g^2/n$, where

$$\sigma_g^2 = \int_{\{x;\, g(x)>0\}} \frac{f^2(x)}{g(x)} dx - \mu^2.$$

Confidence Interval

The central limit theorem implies that the distribution of the importance sampling estimator $\hat{\mu}^{\mathrm{IS}}$ can be approximated by a normal distribution $N\left(\mu, \sigma_g^2/n\right)$. To approximate the distribution and obtain the confidence interval, we need to estimate σ_g^2. One estimator is

$$\hat{\sigma}_g^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{f(x_i)}{g(x_i)} - \hat{\mu}^{\mathrm{IS}}\right]^2.$$

An approximated $1-\alpha$ confidence interval is

$$\hat{\mu}^{\mathrm{IS}} \pm \lambda_{1-\alpha/2}\hat{\sigma}_g/\sqrt{n},$$

where $\lambda_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of $N(0,1)$.

Effect of $g(x)$

$$n\sigma_g^2 = \int_{g(x)>0} \frac{f^2(x)}{g(x)}dx - \mu^2 = \int_{g(x)>0} \frac{[f(x)-\mu g(x)]^2}{g(x)}dx.$$

We need to choose $g(x)$ such that $g(x) > 0$ whenever $f(x) \neq 0$. But even so, the choice of $g(x)$ can still have big impacts.

1. If g is (nearly) proportional to f, then the numerator in the second integral can be small.
2. If g is not proportional to f, then, the ratio in the second integral is magnified by small values of $g(x)$.
3. It is even possible that some choices of $g(x)$ will lead to $\sigma_g^2 = \infty$.

To choose a good $g(x)$ requires some guessing and possibly numerical search.

Importance Weight

Now suppose that we know $f(x) = h(x)\, p(x)$, where $p(x)$ is a density. Then,

$$\hat{\mu}^{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{g(x_i)} = \frac{1}{n}\sum_{i=1}^{n} w(x_i)\, h(x_i),$$

where $w(x) = p(x)/g(x)$ is the importance weight.

Lemma

Suppose that $\int_{g(x)>0} h^2(x)\, p(x)\, dx < \infty$. *Then a sufficient condition for a finite* $Var\left(\hat{\mu}^{IS}\right)$ *is that* $w(x)$ *is bounded.*

Check Finite Variance

Example

Suppose that we want to approximate the integral

$$\mu = \int_0^\infty \log(1+x) \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)\, dx,$$

where $h(x) = \log(1+x)$ and $p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$. We sample X from $\text{Exp}(c)$.

Optimal $g(x)$

What is the smallest $\text{Var}_p(\hat{\mu})$ we can get?

Theorem

Consider the importance distributions that satisfy

$$\{x;\ g(x) > 0\} \quad = \quad \{x;\ p(x) > 0\}.$$

Then,

$$g(x) \quad = \quad \frac{|h(x)|}{E_p[|h(x)|]} p(x)$$

is a probability density and the resulting $Var\left(\hat{\mu}^{IS}\right)$ *is*

$$Var\left(\hat{\mu}^{IS}\right) \quad = \quad \frac{1}{n}\left(\{E_p[|h(x)|]\}^2 - \mu^2\right).$$

Scaling Creates Problems

So far we have assumed that we can evaluate both $p(x)$ and $g(x)$. In practice, it is often the case that we know $p(x)$ up to a scaling constant

$$p(x) = c\tilde{p}(x),$$

where we can easily compute $\tilde{p}(x)$, but the exact value of c is tedious to obtain.

Example

In Bayesian statistics, the posterior distribution of a parameter θ is

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{\int f(\boldsymbol{x} \mid \theta)\pi(\theta)\, d\theta},$$

where $c^{-1} = \int f(\boldsymbol{x} \mid \theta)\pi(\theta)\, d\theta$ is often difficult to obtain.

Idea of Normalized Importance Sampling

Suppose that $g(x) > 0$ whenever $h(x)\,p(x) \neq 0$. Then,

$$\begin{aligned}\mu = \int_{h(x)p(x)\neq 0} h(x)\,p(x)\,dx &= \int_{g(x)>0} h(x)\,\frac{p(x)}{g(x)}\,g(x)\,dx \\ &= c\mathrm{E}_g\left[\tilde{w}(X)\,h(X)\right],\end{aligned}$$

where we define the importance weight $\tilde{w}(x) = \tilde{p}(x)/g(x)$.

However, we still need to approximate c. If we assume $g(x) > 0$ whenever $p(x) \neq 0$, then

$$c^{-1} = \int_{p(x)\neq 0} \tilde{p}(x)\,dx = \int_{g(x)>0} \frac{\tilde{p}(x)}{g(x)}\,g(x)\,dx = \mathrm{E}_g\left[\tilde{w}(X)\right].$$

Normalized Importance Sampling

Under the assumption that $g(x) > 0$ whenever $p(x) \neq 0$, we get

$$\mu = \int h(x)\, p(x)\, dx \quad = \quad \frac{\mathrm{E}_g\left[\tilde{w}(X)\, h(X)\right]}{\mathrm{E}_g\left[\tilde{w}(X)\right]}.$$

The importance sampling estimators to the numerator and denominator are

$$\mathrm{E}_g\left[\tilde{w}(X)\, h(X)\right] \approx \frac{1}{n}\sum_{i=1}^{n} \tilde{w}(x_i)\, h(x_i),$$

$$\mathrm{E}_g\left[\tilde{w}(X)\right] \approx \frac{1}{n}\sum_{i=1}^{n} \tilde{w}(x_i).$$

The ratio is the normalized importance sampling estimator

$$\hat{\mu}^{\mathrm{NIS}} \quad = \quad \frac{\sum_{i=1}^{n} \tilde{w}(x_i)\, h(x_i)}{\sum_{i=1}^{n} \tilde{w}(x_i)}.$$

Normalized Importance Sampling: Algorithm

Algorithm 3: Normalized importance sampling

1 Simulate n iid random numbers from the proposal distribution $g(x)$;

2 Compute the weight $\tilde{w}(x) = \frac{\tilde{p}(x)}{g(x)}$;

3 Approximate μ by

$$\hat{\mu}^{\text{NIS}} = \frac{\sum_{i=1}^{n} \tilde{w}(x_i) h(x_i)}{\sum_{i=1}^{n} \tilde{w}(x_i)}.$$

Normalized Importance Sampling: Example

Example

Suppose that we want to approximate the integral

$$\mu = \int_0^\infty \log(1+x) \times cx^{a-1} \exp(-bx)\, dx,$$

where $p(x) = cx^{a-1}\exp(-bx)$ with $a = 2$, and $b = 4$, and c is the unknown normalizing constant. We sample X from $\text{Exp}(1)$.

Why Does Normalized Importance Sampling Work?

Theorem

Suppose that $g(x) > 0$ *whenever* $p(x) \neq 0$. *Let* $p(x) = c\tilde{p}(x)$. *Then,*

1. $\hat{\mu}^{NIS}$ *is a consistent estimator for* μ.
2. $\hat{\mu}^{NIS}$ *is also asymptotically normal.*

Note here that the normalized importance sampler needs $g(x) > 0$ whenever $p(x) \neq 0$, which is stronger than the ordinary importance sampler that requires $g(x) > 0$ whenever $h(x)\,p(x) \neq 0$.

General MCMC Integral

Suppose that we want to approximate

$$\mu = \int h(x)\,\pi(x)\,dx,$$

for $x \in \mathbb{R}^d$. We can also sample directly from $\pi(x)$ using MCMC.

Algorithm 4: General MCMC Integral

1. Sample a Markov chain for a given stationary distribution $\pi(x)$: $x^{(1)}$, ..., $x^{(n)}$ (after burn-in) ;
2. Approximate μ by

$$\hat{\mu}^{\text{MCMC}} = \frac{1}{n}\sum_{i=1}^{n} h(x_i).$$

Long-Run Property

Theorem

Consider a finite-state Markov chain. Suppose that the transition matrix $\boldsymbol{K}$ *is irreducible, aperiodic, and has stationary distribution* π. *Then, for all starting state* $w_0 \in \Omega$,

1. *ergodic theorem: For any initial state,*

$$n^{-1} \sum_{i=1}^{n} h\left(X_i\right) \overset{a.s.}{\to} E\left[h\left(X\right)\right].$$

2. *central limit theorem: Let* $\sigma^2 = Var\left[h\left(X\right)\right]$ *and* $\rho_j = corr\left(h\left(X^{(1)}\right), h\left(X^{(j+1)}\right)\right)$. *Then,*

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} h\left(X_i\right) - E\left[h\left(X\right)\right]\right] \overset{d}{\to} N\left(0, \sigma^2\left(1 + 2\sum_{j=1}^{\infty}\rho_j\right)\right).$$

Effective Sample Size

If we have an iid sample of size n, then

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} h\left(X_i\right) - \mathrm{E}\left[h\left(X\right)\right]\right] \quad \xrightarrow{d} \quad N\left(0, \sigma^2\right).$$

If we have a converged Markov chain of length n,

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} h\left(X_i\right) - \mathrm{E}\left[h\left(X\right)\right]\right] \quad \xrightarrow{d} \quad N\left(0, \sigma^2\left(1 + 2\sum_{j=1}^{\infty}\rho_j\right)\right).$$

The variance of $\hat{\mu}^{\mathrm{MCMC}}$ is larger than the variance of $\hat{\mu}^{\mathrm{IMC}}$. We define

$$n_{\mathrm{eff}} \quad = \quad \frac{n}{1 + 2\sum_{j=1}^{\infty}\rho_j}$$

as the effective sample size.

Estimate Effective Sample Size

We can also estimate the effective sample size, if we have m Markov chains of length n.

- Following the Gelman-Rubin $\hat{R}$ statistic, we can estimate σ^2 by

$$\hat{V} = \frac{n-1}{n} W + \frac{1}{n} B.$$

- The autocorrelations can be estimated by

$$\hat{\rho}_t = 1 - \frac{\sum_{j=1}^{m} \sum_{i=t+1}^{n} \left(y_{i,j} - y_{i-t,j}\right)^2}{2m\,(n-t)\,\hat{V}}.$$

- The effective sample size is estimated by

$$\hat{n}_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{T} \hat{\rho}_t},$$

where T is the first odd positive integer such that $\hat{\rho}_{T+1} + \hat{\rho}_{T+2}$ is negative.

Alternative Confidence Interval

Suppose that we can divide the Markov chain of length n into b batches (e.g., 20 or proportional to $n^{1/3}$) of m consecutive observations each.

- Let $\bar{y}_j$ be the average of batch j.
- We will treat $\{\bar{y}_j\}$ as iid normal random variables.

An approximate confidence interval is

$$\frac{1}{b}\sum_{j=1}^{b}\bar{y}_j \pm t_{1-\alpha/2}\,(b-1)\sqrt{\frac{1}{b\,(b-1)}\sum_{j=1}^{b}(\bar{y}_j-\bar{y})^2},$$

where $\bar{y}$ is the average of $\{\bar{y}_j\}$. This is just the usual t-confidence interval.

Rao-Blackwell Theorem in Statistical Inference

Theorem (Rao-Blackwell Theorem)

Let $\hat{\theta}$ be an unbiased estimator of θ. Suppose that $T = T(X)$ is a sufficient statistic for θ. Then, $\theta^* = E\left[\hat{\theta} \mid T\right]$ is a uniformly minimum variance unbiased estimator of θ, i.e.,

$$Var\left(\hat{\theta}\right) \quad \geq \quad Var\left(\theta^*\right).$$

A weaker version is of the theorem is based on the low of total variance:

$$\mathrm{Var}\left(X\right) \quad = \quad \mathrm{Var}\left(\mathrm{E}\left[X \mid Y\right]\right) + \mathrm{E}\left(\mathrm{Var}\left[X \mid Y\right]\right) \geq \mathrm{Var}\left(\mathrm{E}\left[X \mid Y\right]\right).$$

Rao-Blackwellization in Monte Carlo

If we are interested in $\mathrm{E}\left[f\left(X,Y\right)\right]$, then

$$\mathrm{Var}\left(f\left(X,Y\right)\right) \quad \geq \quad \mathrm{Var}\left(\mathrm{E}\left[f\left(X,Y\right)\mid Y\right]\right).$$

That is, instead of simulating (X_i, Y_i) to compute $n^{-1}\sum_{i=1}^{n} f\left(X_i, Y_i\right)$, we can simulate only Y_i and compute

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[f\left(X_i,Y_i\right)\mid Y_i\right].$$

This also suggests that we should compute as many analytical steps as possible before Monte Carlo approximation.

Rao-Blackwellization: Example

Example

Consider a Bayesian model, where $X_i \mid \mu, \lambda \sim N\left(\mu, \lambda^{-1}\right)$, $\mu \sim N\left(\mu_0, \lambda_0^{-1}\right)$, and $\lambda \sim \text{Gamma}\left(a_0, b_0\right)$. Then,

$$\begin{aligned} \mu \mid \lambda, \text{data} &\sim N\left(\frac{\lambda_0\mu_0 + \lambda\sum_{i=1}^n x_i}{\lambda_0 + n\lambda}, \frac{1}{\lambda_0 + n\lambda}\right), \\ \lambda \mid \mu, \text{data} &\sim \text{Gamma}\left(a_0 + \frac{n}{2}, b_0 + \frac{1}{2}\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i\mu + \frac{n}{2}\mu^2\right). \end{aligned}$$

We want to approximate $\text{E}\left[\lambda \mid \text{data}\right]$.

Rao-Blackwellization: Another Example

Example

Suppose that we want to estimate $\mathrm{E}\left[X1\left(X>0\right)\right]$, where $X \sim N\left(0,1\right)$.

1. Approach 1: Independent Monte Carlo

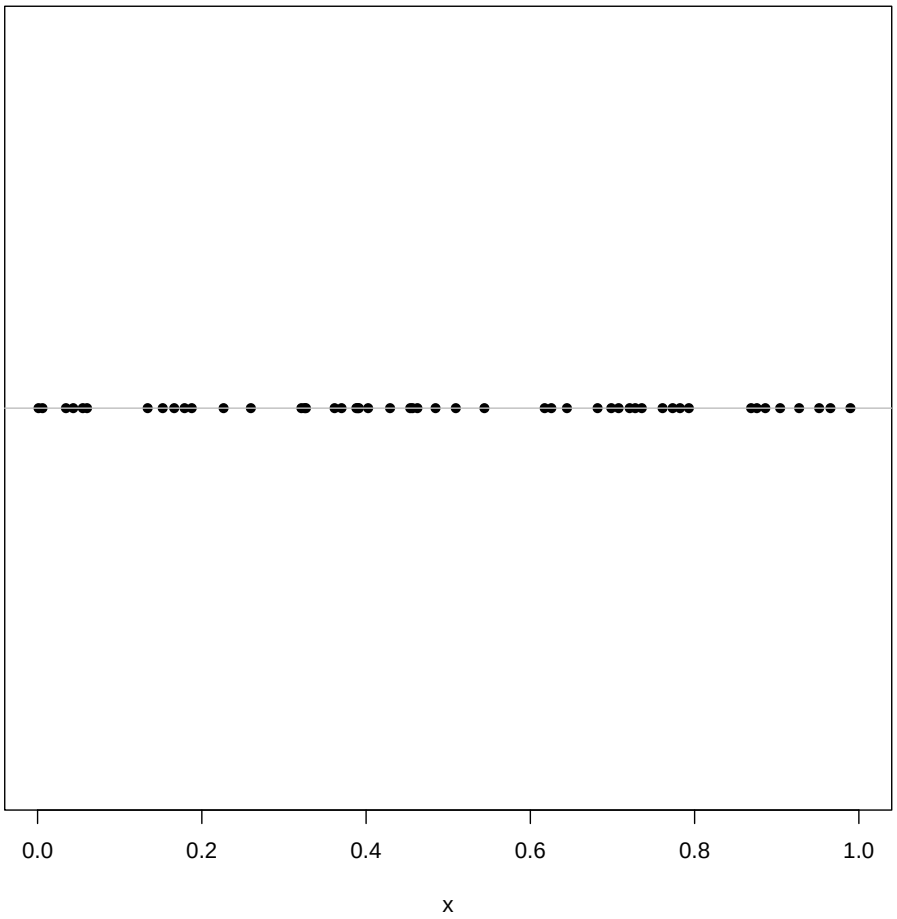
$$\mathrm{E}\left[X1\left(X>0\right)\right] = \int x1\left(x>0\right)\phi\left(x\right)dx \approx \frac{1}{n}\sum_{i=1}^{n} x_i 1\left(x_i>0\right).$$

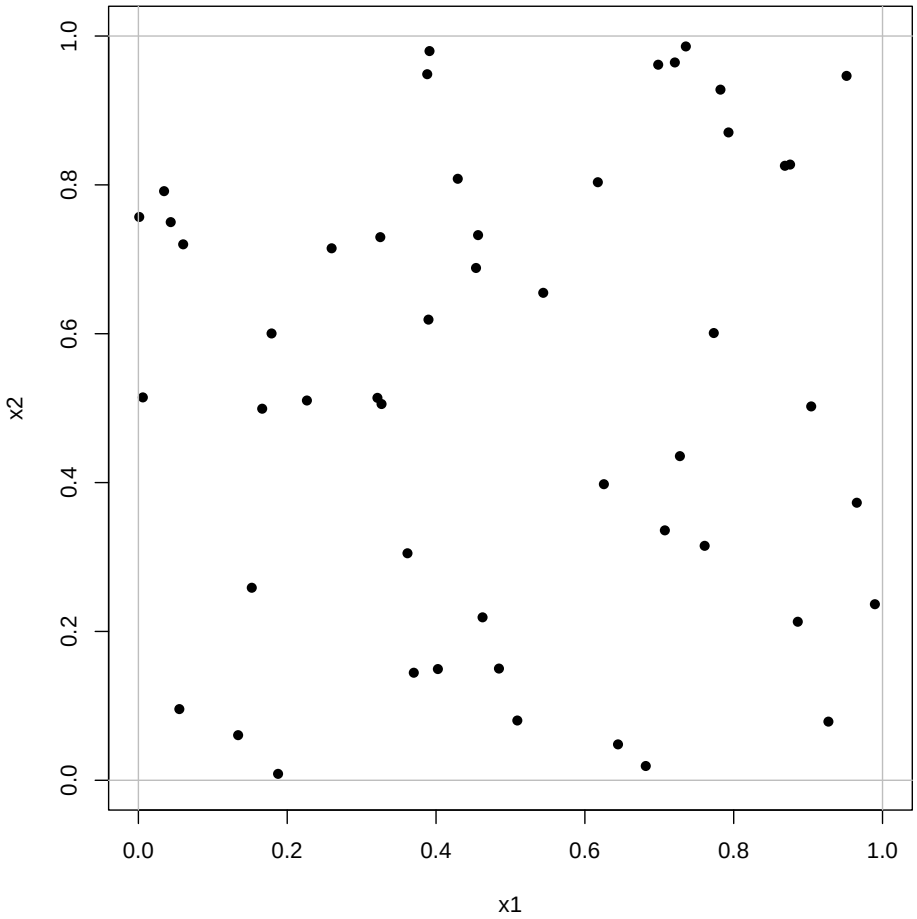
2. Approach 2: Conditioning

$$\mathrm{E}\left[X1\left(X>0\right)\right] = \mathrm{P}\left(X>0\right)\mathrm{E}\left[X \mid 1\left(X>0\right)\right].$$

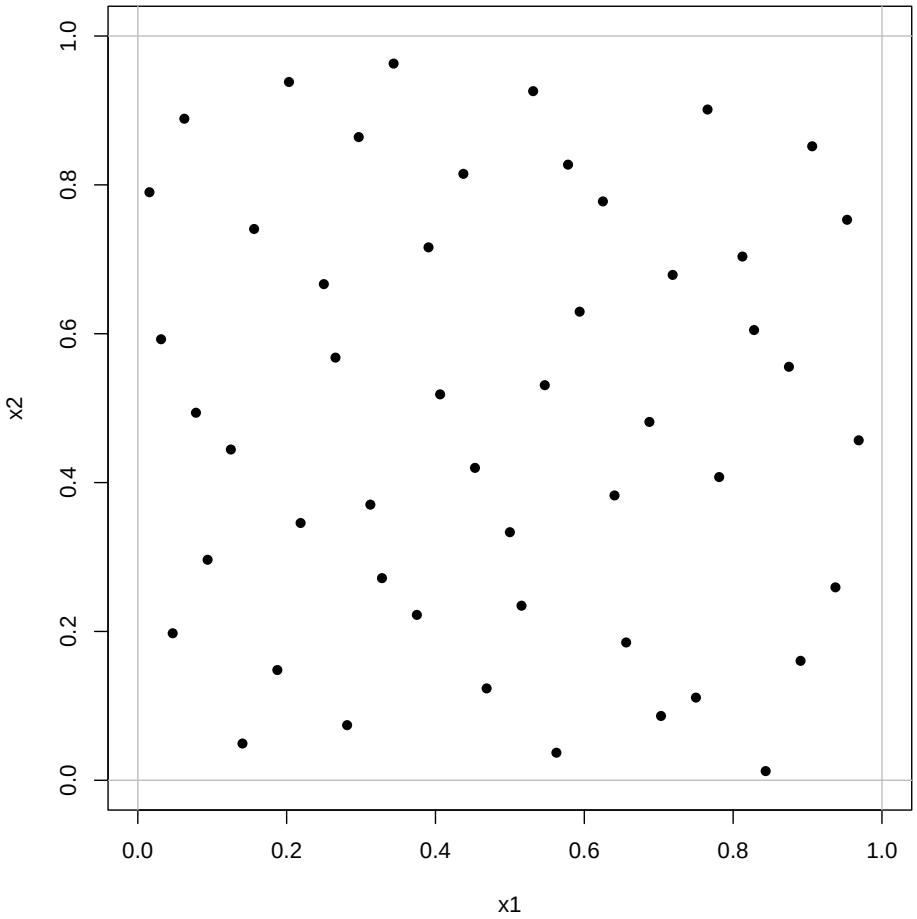
We approximate $\mathrm{E}\left[X \mid 1\left(X>0\right)\right]$.

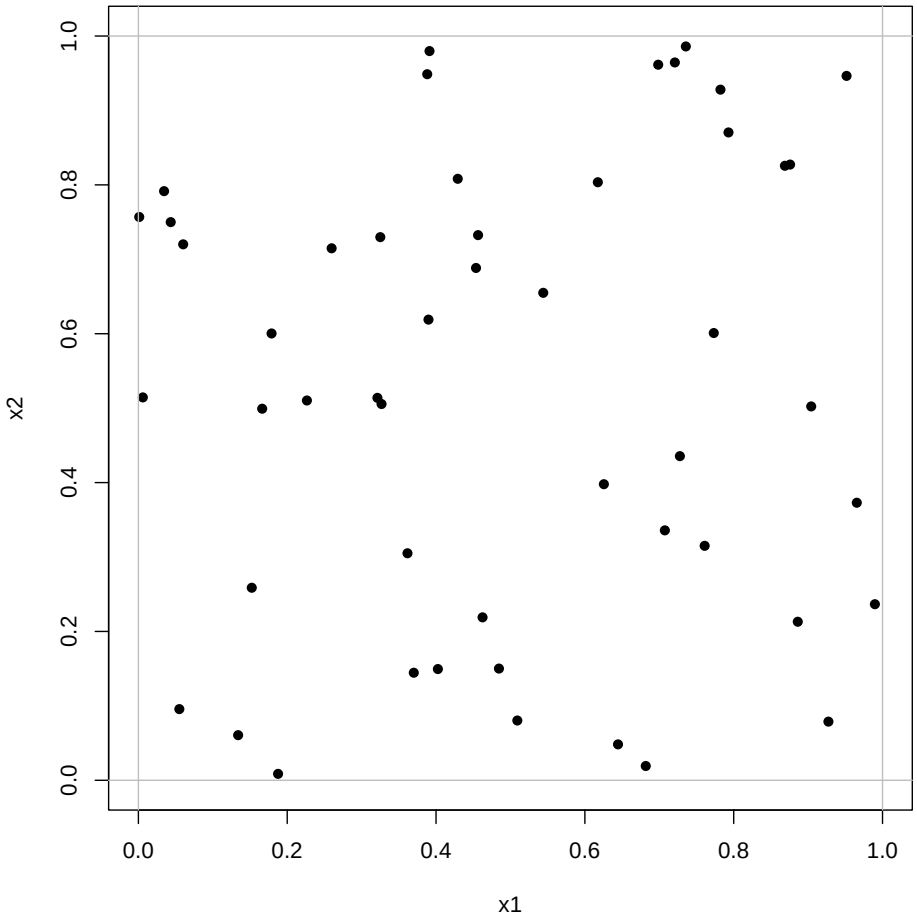
Curse of Dimensionality: Clumsy Allocation





Quasi-Random Numbers





Quasi-Monte Carlo Integral

In quasi-Monte Carlo methods, it is customary to focus on numerical integration over the unit hypercube

$$\mu = \int_{x \in [0,1]^d} f(x)\, dx.$$

- In case the domain is not a hypercube, we make transformations to obtain a hypercube, e.g., through the cumulative distribution function, we can transform $\mathbb{R}$ to $[0, 1]$.

Let X be a uniform distributed random variable/vector on $[0, 1]$. Then, $\mu = \mathrm{E}[f(X)]$. The quasi-Monte Carlo approximation is

$$\hat{\mu}^{\mathrm{QMC}} = \frac{1}{n} \sum_{i=1}^{n} f(x_i),$$

for some cleverly chosen $\{x_i\}$.

Discrepancy

Our starting point is that we want the discrepancy of the points $\{x_i\}$ to be low.

- Given a collection $\mathcal{A}$ of subsets in $[0,1]^d$, we define the discrepancy as

$$D_n\left(\{x_i\}_{i=1}^n, \mathcal{A}\right) = \sup_{A\in\mathcal{A}} \left| \frac{\#\{x_i \in A\}}{n} - \text{volume}(A) \right|.$$

- For example, we can take $\mathcal{A}$ to be the collections of all rectangles of the form

$$[0, v_1] \times [0, v_2] \times \cdots \times [0, v_d].$$

The resulting discrepancy is called the star discrepancy D_n^*.

Uniform Allocation

A natural attempt to achieve low discrepancy is to use a grid, in the spirit of Riemann sum. For example,

- consider $[0,1]$ and divide it into K equally spaced intervals, i.e., $x_i = \frac{2i-1}{2n}$ for $i = 1, ..., K$.
- consider $[0,1]^2$ and divide it into K^2 equally spaced squares, i.e., take $x_{ij} = \left(\frac{2i-1}{2n}, \frac{2j-1}{2n}\right)$ for $(i,j) \in \{1, ..., K\}$.

If we choose K points per dimension, the total number of points in the grid is $n = K^d$.

- Grows too quickly with either K or d.

Another limitation is that, if we find out n is not large enough want to increase it, we need to construct the whole grid again.

Halton Sequence

It is computationally more efficient if we have an infinite sequence and just take the first n points.

- If we want to include more points, just start with merging $(n+1)$th point, etc.

The Halton sequence is such an example.

- The Halton sequence starts with a set of coprime integers that are greater than 1, called the bases.
- For example, if the dimension is $d = 2$, then we choose $(2, 3)$.
- Another example, if the dimension is $d = 5$, then we choose $(2, 3, 5, 7, 11)$.

Create Halton Sequence

Take the dimension $d = 2$ as an example. We choose the coprime integers $(2, 3)$. The sequence is

For the first dimension

$$
\begin{array}{cccccc}
\text{initial} & 0 & & & & \\
\left(x + 2^{-1}\right): & \frac{1}{2} & & & & \\
\left(x + 2^{-2}\right): & \frac{1}{4} & \frac{3}{4} & & & \\
\left(x + 2^{-3}\right): & \frac{1}{8} & \frac{5}{8} & \frac{3}{8} & \frac{7}{8} & \\
\left(x + 2^{-4}\right): & \frac{1}{16} & \frac{9}{16} & \frac{5}{16} & \cdots & \\
\vdots & & & & &
\end{array}
$$

For the second dimension

$$
\begin{array}{cccccc}
\text{initial} & 0 & & & & \\
\left(x + 3^{-1}\right): & \frac{1}{3} & \frac{2}{3} & & & \\
\left(x + 3^{-2}\right): & \frac{1}{9} & \frac{4}{9} & \frac{7}{9} & & \\
\left(x + 2 \cdot 3^{-2}\right): & \frac{2}{9} & \frac{5}{9} & \frac{8}{9} & & \\
\left(x + 3^{-3}\right): & \frac{1}{27} & \frac{10}{27} & \frac{19}{27} & \cdots & \\
\vdots & & & & &
\end{array}
$$

We often ignore the zero point $(0, 0)$ when we use the Halton sequence in practice.

Halton as Low Discrepancy Sequence

The infinite sequence $x_1, x_2, ..., \in [0,1]^d$ is a low discrepancy sequence if

$$D_n^*(x_1, ..., x_n) = O\left(\frac{(\log n)^d}{n}\right), \quad n \to \infty.$$

- Any finite positive power of $\log n$ is asymptotically negligible compared to n.
- Hence, a low discrepancy sequence satisfies $D_n^* = O\left(n^{-1+\epsilon}\right)$ for any $\epsilon > 0$.

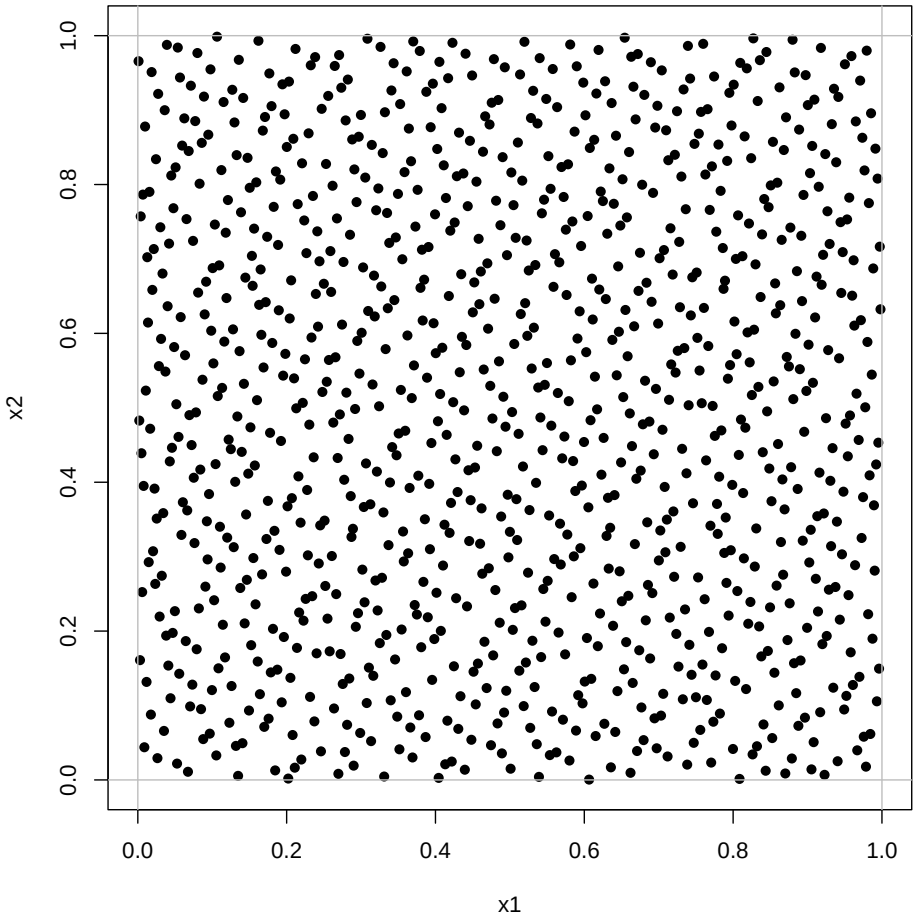
The star discrepancy of the first n Halton points has the upper bound

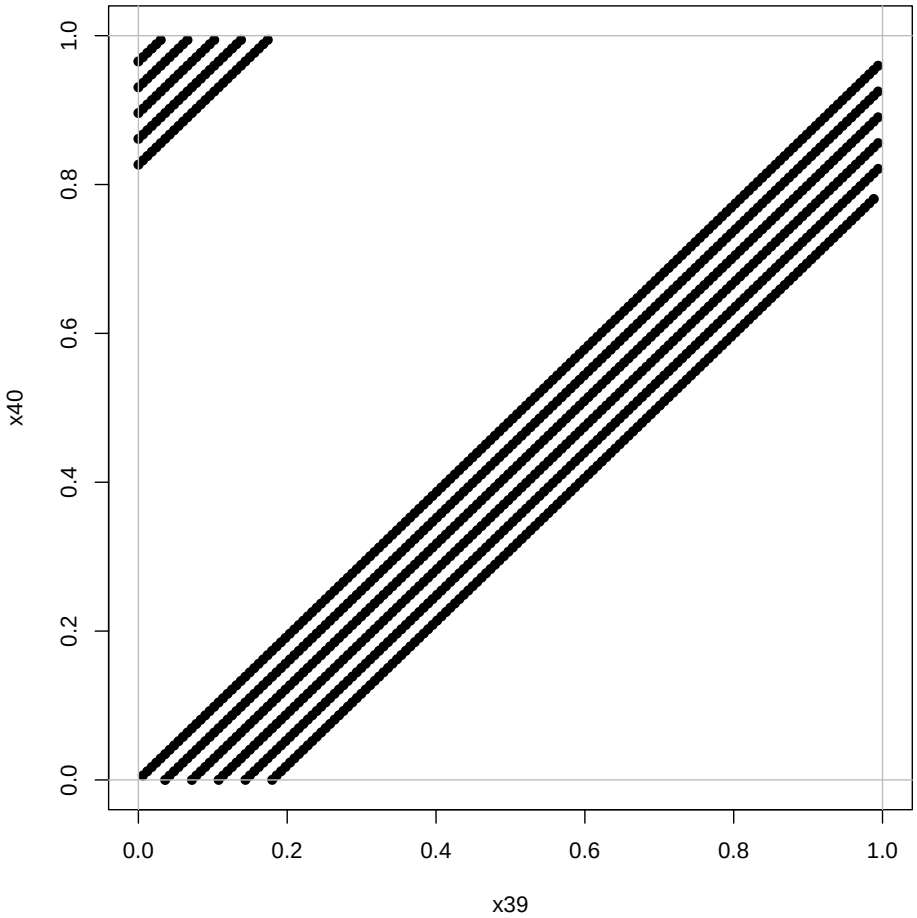
$$D_n^* \leq C_d \frac{(\log n)^d}{n} + O\left(\frac{(\log n)^{d-1}}{n}\right),$$

where C_d does not depend on n. Hence, it is a low-discrepancy sequence.

An Illustration: $n = 1000$, $d = 40$

But C_d increases in the same rate as d^d, meaning that discrepancy can be large if d is very large. This is caused by the fact that we need a large base if d is large.





Digital Nets

Let $d \geq 1$ and $b \geq 2$ be integers. An elementary interval in base b is a rectangle of the form

$$\left[\frac{c_1}{b^{k_1}}, \frac{c_1+1}{b^{k_j}}\right) \times \cdots \times \left[\frac{c_d}{b^{k_d}}, \frac{c_d+1}{b^{k_d}}\right),$$

for integers $k_j \geq 0$ and $0 \leq c_j < b^{k_j}$.

Definition

Let $m \geq t \geq 0$, $b \geq 2$, and $d \geq 1$ be integers. The sequence x_1, ..., $x_{b^m} \in [0,1)^d$ is a (t, m, d)-net in base b, if every elementary interval in base b of volume b^{t-m} contains exactly b^t points of the sequence.

- Since $m \geq t$, a smaller t indicates that the volume b^{t-m} is smaller and the points are more "equidistribution".

Sobol Sequence

- The (t, m, d)-net in base b has exactly b^m points. The infinite sequence x_1, x_2, ..., $\in [0,1)^d$ is a (t, d)-sequence in base b if for all $k \geq 0$ and $m \geq t$, the sequence x_{kb^m+1}, ..., $x_{(k+1)b^m} \in [0,1)^d$ is a (t, m, d)-net in base b, i.e., a series of (t, m, d)-nets, one after another.
- A Sobol sequence is a (t, d)-sequence in base $b = 2$, where the value of t depends on d.
- Because of the structure of a (t, m, d)-net in base b, if we use $n = b^m$ points from a (t, d)-sequence, then the next sample size that retains "equidistribution" is $n = 2b^m$.

Koksma-Hlawka Inequality

When we replace randomly sampled points by deterministic ones, we can no longer use the law of large numbers or the central limit theorem.

Theorem (Koksma-Hlawka Inequality)

For $d \geq 1$ and $x_1, ..., x_n \in [0,1]^d$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{[0,1]^d} f(x)\, dx \right| \leq D_n^*(x_1, ..., x_n)\, V_{HK}(f),$$

where $V_{HK}(f)$ denotes the total variation of f in the sense of Hardy and Krause.

- $V_{HK}(f)$ depends on the property of $f(x)$. If f' is continuous, then

$$V_{HK}(f) = \int_0^1 |f'(x)|\, dx.$$

QMC Versus MC

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{[0,1]^d} f(x)\, dx \right| \leq D_n^*(x_1, ..., x_n)\, V_{HK}(f)$$

- This result shows that as long as we have a low frequency sequence and $V_{HK} < \infty$, then we will achieve

$$\left| \hat{\mu}^{\mathrm{QMC}} - \mu \right| = O\left(n^{-1+\epsilon}\right)$$

  for any $\epsilon > 0$.
  - For large enough n, QMC is expected to be more accurate than MC with rate $O\left(n^{-1/2}\right)$.
- However, error analysis based on the central limit theorem is easy for MC.
  - For QMC, the error bound involves unknown quantities that are often even harder to obtain than μ.

Randomized QMC

It is hard to estimate of QMC integral error since QMC is deterministic.

- As a remedy, we inject some randomness into $\{x_i\}$, but still keep low discrepancy.

Definition

Random variables $X_i \in [0,1]^d$ for $i \geq 1$ comprise a randomized quasi-Monte Carlo rule, if there exist $B < \infty$ and $N > 0$ such that

$$\mathrm{P}\left(D_n^*(X_1, ..., X_n) < B(\log n)^d / n\right) = 1, \text{ for all } n \geq N,$$

and $X_i \sim \text{Uniform}\,[0,1]^d$ for all $i \geq 1$.

Properties of Randomized QMC

The randomized QMC approximation is still

$$\hat{\mu}^{\mathrm{RQMC}} = \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

- $\mathrm{E}\left[\hat{\mu}^{\mathrm{RQMC}}\right] = \mu$, i.e, $\hat{\mu}^{\mathrm{RQMC}}$ is unbiased.
- If $V_{HK} < \infty$, then

$$\mathrm{Var}\left(\hat{\mu}^{\mathrm{RQMC}}\right) < B^2 \left[V_{NK}(f)\right]^2 \frac{(\log n)^{2d}}{n^2} = O\left(n^{-2+2\epsilon}\right),$$

for large enough n.

- Randomized QMC is asymptotically more stable than Monte Carlo, since $\mathrm{Var}\left(\hat{\mu}^{\mathrm{MC}}\right) = O\left(n^{-1}\right)$.

Confidence Interval

Suppose that we can repeat randomized QMC independently R times. Then,

$$\hat{\mu}^{\text{RQMC}} = \frac{1}{R}\sum_{r=1}^{R}\hat{\mu}_r^{\text{RQMC}},$$

where $\hat{\mu}_r^{\text{RQMC}}$ is the estimate at rth replication.

- We can estimate the variance by

$$\widehat{\text{Var}}(\hat{\mu}^{\text{RQMC}}) = \frac{1}{R(R-1)}\sum_{r=1}^{R}\left(\hat{\mu}_r^{\text{RQMC}} - \hat{\mu}^{\text{RQMC}}\right)^2.$$

- A t-confidence interval is then

$$\hat{\mu}^{\text{RQMC}} \pm t_{1-\alpha/2}(R-1)\sqrt{\widehat{\text{Var}}(\hat{\mu}^{\text{RQMC}})}.$$

Cranley-Patterson Rotation

A simple way to obtain randomized QMC is the Cranley-Patterson rotation by taking

$$X_i \quad = \quad a_i + U \quad \mod 1,$$

where $U \sim \text{Uniform}\,[0,1]^d$.

- The Cranley-Patterson rotation of low discrepancy points retains low discrepancy.
- However, it is not commonly applied to a (t, m, d)-net. It is more commonly used for a lattice rule.

Scrambling

The general idea of scrambling for a (t, m, d)-net in base b is to

1. chop $[0,1]^d$ into b slices, and shuffle those slices in random order,
2. chop each of b slices into b thinner slices, and shuffle the thinner slices in random order within their respective original slices,
3. chop each of b^2 thinner slices into b even thinner slices, and shuffle the even thinner slices in random order within their respective thinner slices,
4. proceed the algorithm.

State Space Model

A state space model is a type of probabilistic models that consists of latent variables X and observed variables Y.

- For example, for $t \geq 1$,

$$\begin{aligned} Y_t \mid X_t &\sim p\left(y_t \mid X_t\right), \\ X_{t+1} \mid X_t &\sim p\left(x_{t+1} \mid X_t\right), \end{aligned}$$

  independently, where $p\left(\right)$ is used as a generic symbol for densities.

We are often interested in the posterior densities $p\left(x_{1:t} \mid y_{1:t}\right)$ and $p\left(x_t \mid y_{1:t}\right)$, or

$$\mu_t = \mathrm{E}\left[h_t\left(x_{1:t}\right) \mid y_{1:t}\right] = \int h_t\left(x_{1:t}\right) p\left(x_{1:t} \mid y_{1:t}\right) dx_{1:t},$$

for some function $h_t\left(x_{1:t}\right)$.

State Space Model: One Example

A typical example of a state space model is the location problem.

- Let X_t be the location of an object at time t. The dynamic is

$$X_{t+1} = X_t + u_t + V_t,$$

  where u_t (velocity) is known, and V_t is the unknown disturbance.

- The measurement model is

$$Y_t = h(X_t) + E_t,$$

  where E_t is the unknown disturbance, and the function $h(x)$ denotes the height of the position x.

- We want to know where we are at time t (x_t), i.e., compute the density $p(x_t \mid y_{1:t})$.

Evaluate Expectation

Suppose that we want to evaluate

$$\mu_t = \mathrm{E}\left[h_t\left(x_{1:t}\right) \mid y_{1:t}\right] = \int h_t\left(x_{1:t}\right) p\left(x_{1:t} \mid y_{1:t}\right) dx_{1:t}.$$

At any time $t \geq 1$, the posterior density satisfies

$$p\left(x_{1:t} \mid y_{1:t}\right) = \frac{p\left(y_{1:t} \mid x_{1:t}\right) p\left(x_{1:t}\right)}{p\left(y_{1:t}\right)}.$$

1. The normalizing constant $p\left(y_{1:t}\right)$ can be difficult to obtain.
2. The dimension of the integral is high as t increases, i.e., new data are collected sequentially.

Approximate Expectation

$$\mu_t = \int h_t(x_{1:t})\, p(x_{1:t} \mid y_{1:t})\, dx_{1:t}.$$

- If we can simulate $\left\{x_{1:t}^{(i)}\right\}_{i=1}^{N}$ from $p(x_{1:t} \mid y_{1:t})$, then we approximate μ_t by independent Monte Carlo.
- If simulating $\left\{x_{1:t}^{(i)}\right\}_{i=1}^{N}$ from $p(x_{1:t} \mid y_{1:t})$ is not an trivial task, we can approximate it by normalized importance sampling because

$$\mu_t = \int h_t(x_{1:t}) \frac{p(y_{1:t} \mid x_{1:t})\, p(x_{1:t})}{\int p(y_{1:t} \mid x_{1:t})\, p(x_{1:t})\, dx_{1:t}} dx_{1:t},$$

where $Z = \int p(y_{1:t} \mid x_{1:t})\, p(x_{1:t})\, dx_{1:t}$ is the intractable normalizing constant for $p(x_{1:t} \mid y_{1:t})$.

Normalized Importance Sampling

Let $g()$ be the importance distribution. Then,

$$\begin{aligned}
\mu_t &= \int h_t(x_{1:t}) \frac{p(y_{1:t} \mid x_{1:t}) p(x_{1:t})}{\int p(y_{1:t} \mid x_{1:t}) p(x_{1:t}) dx_{1:t}} dx_{1:t} \\
&= \frac{\int h_t(x_{1:t}) p(y_{1:t} \mid x_{1:t}) p(x_{1:t}) / g(x_{1:t}) \cdot g(x_{1:t}) dx_{1:t}}{\int p(y_{1:t} \mid x_{1:t}) p(x_{1:t}) / g(x_{1:t}) \cdot g(x_{1:t}) dx_{1:t}} \\
&= \frac{\int h_t(x_{1:t}) w_t g(x_{1:t}) dx_{1:t}}{\int w_t g(x_{1:t}) dx_{1:t}},
\end{aligned}$$

where the importance weight is

$$w_t(x_{1:t}) = \frac{p(y_{1:t} \mid x_{1:t}) p(x_{1:t})}{g(x_{1:t})}.$$

Here $g()$ can also depend on $y_{1:t}$, but we drop it for simplicity.

Normalized Importance Sampling: Approximation

The normalized importance sampling approximation is

$$\hat{\mu}_t = \sum_{i=1}^{N} \frac{w_t\left(x_{1:t}^{(i)}\right)}{\sum_{j=1}^{N} w_t\left(x_{1:t}^{(j)}\right)} h_t\left(x_{1:t}^{(i)}\right),$$

where $\left\{x_{1:t}^{(i)}\right\}_{i=1}^{N}$ is simulated from $g\left(x_{1:t}\right)$. Each $x_{1:t}^{(i)}$ is called a particle.

Equivalently, normalized importance sampling approximates the distribution of $x_{0:t} \mid y_{1:t}$ by an empirical distribution

$$\hat{\mathrm{P}}\left(x_{1:t} = x_{1:t}^{(i)} \mid y_{1:t}\right) = \frac{w_t\left(x_{1:t}^{(i)} \mid y_{1:t}\right)}{\sum_{j=1}^{N} w_t\left(x_{1:t}^{(j)} \mid y_{1:t}\right)}.$$

Sequential Update of Posterior $p\left(x_{1:t} \mid y_{1:t}\right)$

However, generating the whole trajectory $\left\{x_{1:t}^{(i)}\right\}$ for every t is demanding, especially when t becomes large. Luckily, many elements can be updated sequentially.

Using Bayes formula, we can show that the posterior density of our state space model becomes

$$\begin{aligned} p\left(x_{1:t} \mid y_{1:t}\right) &= \frac{p\left(y_t \mid x_{1:t}, y_{1:t-1}\right) p\left(x_{1:t} \mid y_{1:t-1}\right)}{p\left(y_t \mid y_{1:t}\right)} \\ &= \frac{p\left(y_t \mid x_t\right) p\left(x_t \mid x_{t-1}\right)}{p\left(y_t \mid y_{1:t-1}\right)} p\left(x_{1:t-1} \mid y_{1:t-1}\right), \end{aligned}$$

Hence, we can update $p\left(x_{1:t} \mid y_{1:t}\right)$ recursively.

Sequential Update of Posterior $p(x_t \mid y_{1:t})$

Likewise, the Bayes formula implies

$$p(x_t \mid y_{1:t}) = \frac{p(y_t \mid x_t, y_{1:t-1})\, p(x_t \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})} = \frac{p(y_t \mid x_t)\, p(x_t \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})}.$$

Here, $p(x_t \mid y_{1:t-1})$ can be viewed as the density used for forecasting:

$$\begin{aligned} p(x_t \mid y_{1:t-1}) &= \int p(x_t \mid x_{t-1}, y_{1:t-1})\, p(x_{t-1} \mid y_{1:t-1})\, dx_{t-1} \\ &= \int p(x_t \mid x_{t-1})\, p(x_{t-1} \mid y_{1:t-1})\, dx_{t-1}. \end{aligned}$$

However, the integrals we need (e.g., $p(y_t \mid y_{1:t-1})$ and $p(x_t \mid y_{1:t-1})$) often do not closed form expressions.

Sequential Importance Sampling (SIS): Idea

The normalized importance sampling can be used to handle integrals with unknown normalizing constant. We can further modify it so that we don't need to resimulate $\left\{x_{1:t-1}^{(i)}\right\}_{i=1}^{N}$ when we need to incorporate data at time t.

This means that the marginal $x_{1:t-1}$ from $g(x_{1:t})$ should be the same as the distribution $g(x_{1:t-1})$ we used to simulate $\left\{x_{1:t-1}^{(i)}\right\}_{i=1}^{N}$.

- The trick is to use the conditional distribution

$$g(x_{1:t}) = g(x_{1:t-1})\, g(x_t \mid x_{1:t-1}).$$

- This means that the importance distribution satisfies

$$g(x_{1:t}) = g(x_1) \prod_{k=2}^{t} g(x_k \mid x_{1:k-1}).$$

Back to State Space Model

For our state space model, we can show that

$$\begin{aligned} w_t\left(x_{1:t}\right) &= \frac{p\left(y_{1:t} \mid x_{1:t}\right) p\left(x_{1:t}\right)}{g\left(x_{1:t}\right)} \\ &= w_{t-1}\left(x_{1:t-1}\right) \times \frac{p\left(y_{1:t} \mid x_{1:t}\right) p\left(x_t \mid x_{1:t-1}\right)}{p\left(y_{1:t-1} \mid x_{1:t-1}\right) g\left(x_t \mid x_{1:t-1}\right)} \\ &= w_{t-1}\left(x_{1:t-1}\right) \times \frac{p\left(y_t \mid x_t\right) p\left(x_t \mid x_{1:t-1}\right)}{g\left(x_t \mid x_{1:t-1}\right)}. \end{aligned}$$

A special example of $g()$ that further simplifies the importance weights is

$$g\left(x_{1:t}\right) = p\left(x_{1:t}\right) = p\left(x_1\right) \prod_{k=2}^{t} p\left(x_k \mid x_{1:k-1}\right).$$

Approximate μ by Sequential Importance Sampling

Hence, the importance weights can be updated recursively and the sequential importance sampling (SIS) approximation of μ_t is

$$\hat{\mu}_t = \sum_{i=1}^{N} \frac{w_t\left(x_{1:t}^{(i)}\right)}{\sum_{j=1}^{N} w_t\left(x_{1:t}^{(j)}\right)} h_t\left(x_{1:t}^{(i)}\right).$$

Equivalently, we still let

$$\hat{\mathrm{P}}\left(x_{1:t} = x_{1:t}^{(i)} \mid y_{1:t}\right) = \frac{w_t\left(x_{1:t}^{(i)}\right)}{\sum_{j=1}^{N} w_t\left(x_{1:t}^{(j)}\right)},$$

but we update $w_t(x_{1:t})$ sequentially. This is often called a particle filter.

State Space Model: Linear Gaussian State Space Model

Let

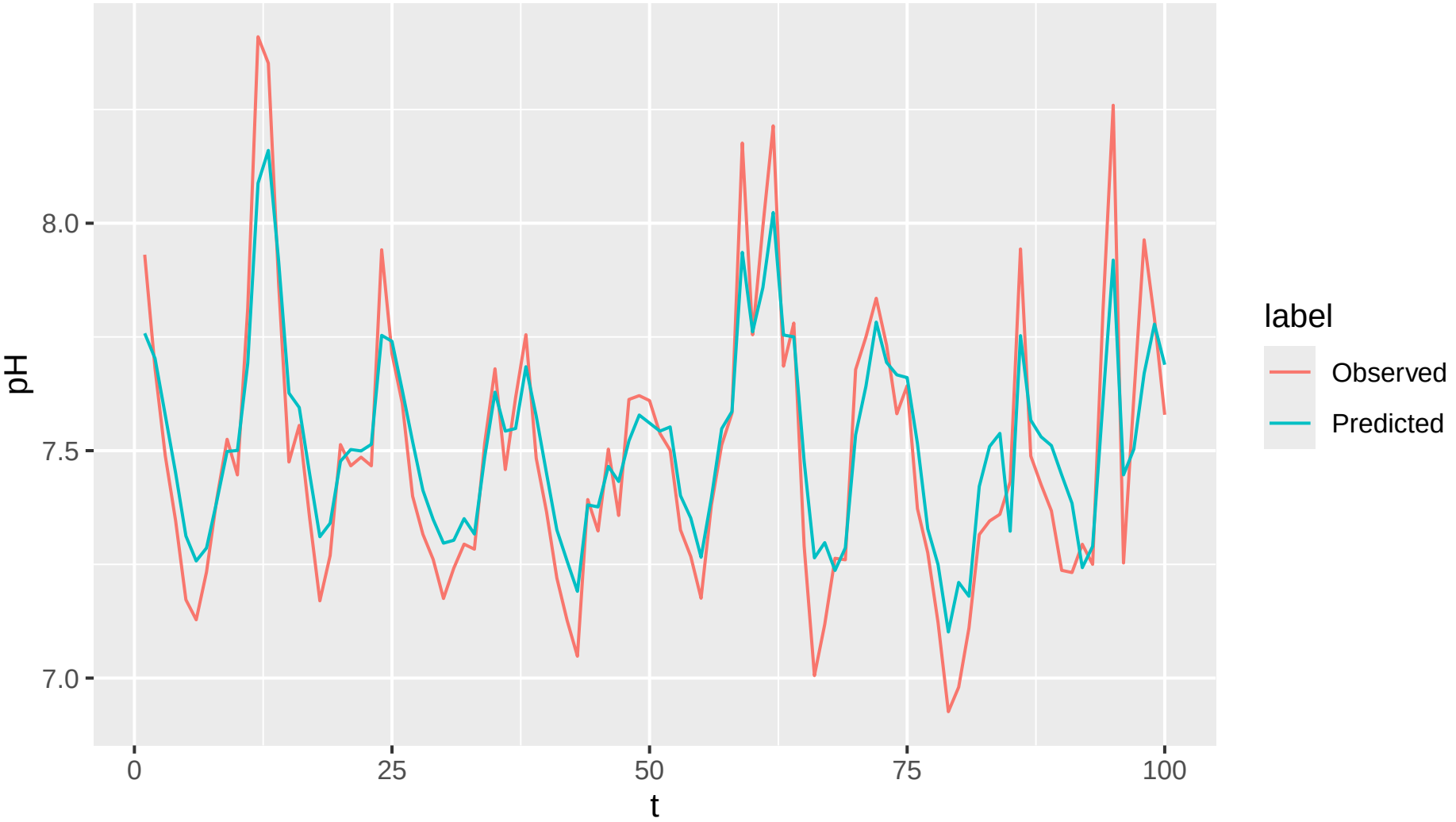
$$\begin{aligned} X_1 &\sim N\left(0, \sigma^2\right), \quad \text{initial density} \\ Y_t &= \beta X_t + W_t, \quad \text{observation density} \\ X_{t+1} &= \gamma X_t + V_{t+1}, \quad \text{transition density} \end{aligned}$$

where $W_t \sim N\left(0, \sigma_y^2\right)$, $V_t \sim N\left(0, \sigma_x^2\right)$, and errors are all mutually independent. We are interested in $\mu_t = \mathrm{E}\left[x_t \mid Y_{1:t} = y_{1:t}\right]$.

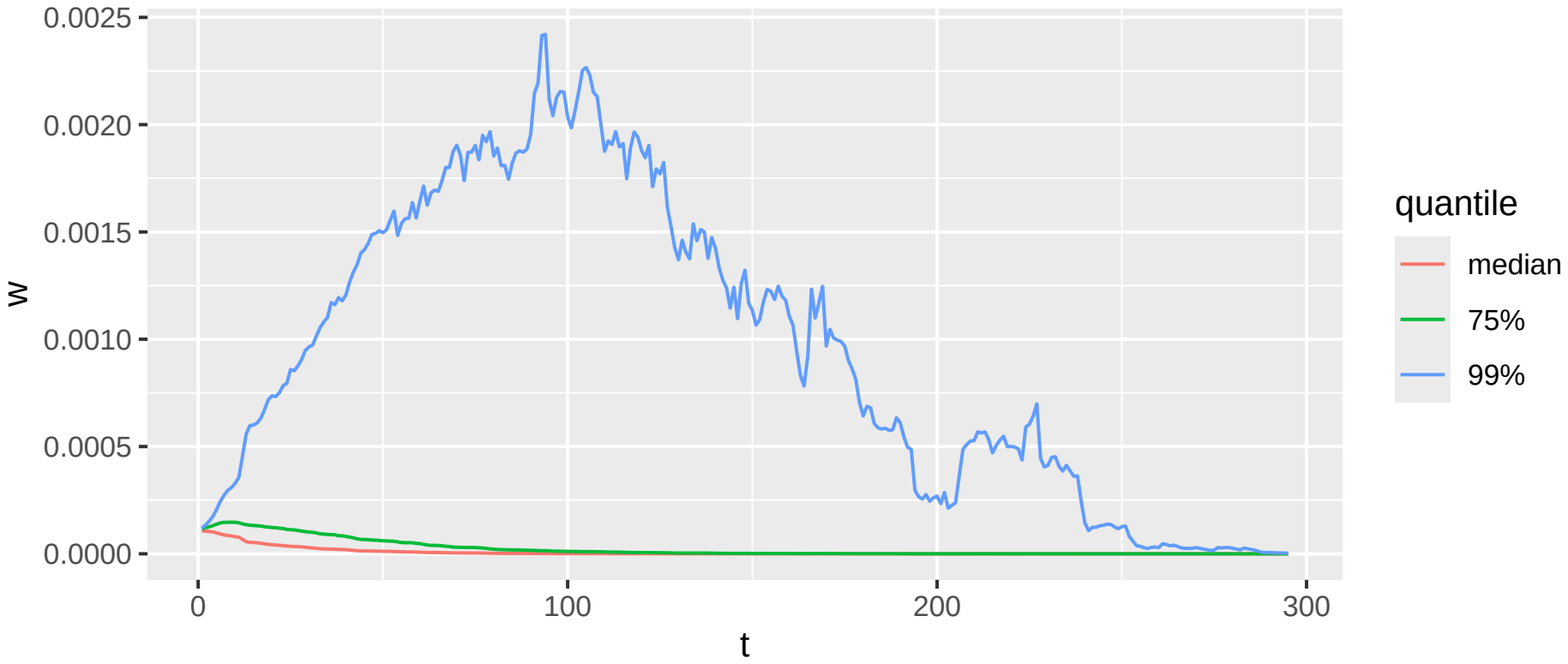
This model means that, for $t \geq 1$,

$$\begin{aligned} Y_t \mid X_t &\sim N\left(\beta X_t, \sigma_y^2\right), \\ X_{t+1} \mid X_t &\sim N\left(\gamma X_t, \sigma_x^2\right). \end{aligned}$$

Example: pH Values from Lake Washington, $N = 10000$



However, Something is Wrong



In fact, it has been proved that, as the dimension of integral increases, the maximum of weights converge in probability to 1!

Bootstrap Particle Filter

To avoid the degenerate weights, a simple but revolutionary idea is to resample the simulated trajectories (i.e., kill trajectories with low weights). This is known as SIS with resampling or bootstrap particle filter.

- We consider a multinomial distribution with possible outcomes $\left\{x_{1:t}^{(i)}\right\}_{i=1}^{N}$, with probabilities

$$\hat{\mathrm{P}}\left(x_{1:t} = x_{1:t}^{(i)} \mid y_{1:t}\right) = \frac{w_t\left(x_{1:t}^{(i)}\right)}{\sum_{j=1}^{N} w_t\left(x_{1:t}^{(j)}\right)}, \quad i = 1, \ldots N.$$

- We resample with replacement from such multinomial distribution to get N new trajectories.
- The weight of each resampled trajectory is reset to $1/N$.

After Resampling

Before resampling, we have the trajectories $\left\{x_{1:t}^{(i)}\right\}_{i=1}^{N}$, each with probability

$$\hat{\mathrm{P}}\left(x_{1:t} = x_{1:t}^{(i)} \mid y_{1:t}\right) = \frac{w_t\left(x_{1:t}^{(i)} \mid y_{1:t}\right)}{\sum_{j=1}^{N} w_t\left(x_{1:t}^{(j)} \mid y_{1:t}\right)}, \quad i = 1, \ldots N.$$

After resampling, we obtain new trajectories $\left\{\tilde{x}_{1:t}^{(i)}\right\}_{i=1}^{N}$, each has probability

$$\hat{\mathrm{P}}\left(x_{1:t} = \tilde{x}_{1:t}^{(i)} \mid y_{1:t}\right) = \frac{1}{N}, \quad i = 1, \ldots N.$$

Thus,

$$\hat{\mu}_t = \frac{1}{N}\sum_{i=1}^{N} h_t\left(\tilde{x}_{1:t}^{(i)}\right).$$

Effects of Resampling

Without resampling, only one particle receives a nonzero importance weight. Hence, it fails to represent the posterior distribution.

With resampling, the weights are more spread out.

1. Resampling does not create bias in the sense that

$$\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} h_t\left(\tilde{x}_{1:t}^{(i)}\right)\right] = \mathrm{E}\left[\sum_{i=1}^{N}\frac{w_t\left(x_{1:t}^{(i)}\right)}{\sum_{j=1}^{N} w_t\left(x_{1:t}^{(j)}\right)} h_t\left(x_{1:t}^{(i)}\right)\right].$$

2. The price is that resampling increases the variance:

$$\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} h_t\left(\tilde{x}_{1:t}^{(i)}\right)\right] \geq \mathrm{Var}\left\{\sum_{i=1}^{N}\frac{w_t\left(x_{1:t}^{(i)}\right)}{\sum_{j=1}^{N} w_t\left(x_{1:t}^{(j)}\right)} h_t\left(x_{1:t}^{(i)}\right)\right\}.$$

More General Model

Beyond our state space models, we can have more complicated models such as

$$\begin{aligned} X_t \mid X_{t-1} &\sim p\left(x_t \mid X_{1:t-1}, \theta\right), \\ Y_t \mid X_t &\sim p\left(y_t \mid X_{1;t}, \theta\right), \end{aligned}$$

where θ is the vector of unknown parameters.

In the context of Bayesian statistics, we are interested in the posterior

$$p\left(x_{1:t}, \theta \mid y_{1:t}\right)$$

and $\mathrm{E}\left[h_t\left(X_{1:t}, \theta\right) \mid y_{1:t}\right]$.