Quantile Importance Sampling, and Inverse Probability Weighting

Jyotishka Datta [1]

June 3, 2025

Virginia Tech

Quantile Importance Sampling and Inverse Probability Weighting[2]

Outline of the Talk

- Motivation: Evidence estimation.
- Riemann sum estimators [Yakowitz et al., 1978]: fast $O(n^{-4})$ convergence.
- Nested sampling [Skilling, 2006]: transforms high-dimensional integrals to 1D.
- Key idea: Combine Riemann sums with nested sampling via quantile reordering.
- Result: **Quantile Importance Sampling (QIS)** — fast, flexible, and design-aware.
- Connection: Inverse Probability Weighting (IPW) $\leftrightarrow$ Evidence estimation.

---

[2]Joint work with Nick Polson.

Evidence Estimation

- Interest: approximation of integrals of the form:

$$\psi = \mathbb{E}_F L(\boldsymbol{\theta}) = \int_{\chi} L(\boldsymbol{\theta}) dF(\boldsymbol{\theta})$$

where $L(\boldsymbol{\theta})$ is a measurable function of interest and $F(\cdot)$ is a finite measure, such as a probability.

- Evidence estimation, i.e.

$$m(\mathbf{x}) = \int_{\Theta} L(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where $L(\mathbf{x} \mid \boldsymbol{\theta})$ is the likelihood with observed data $\mathbf{X} = \mathbf{x}$, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the parameter of interest and $\pi(\boldsymbol{\theta})$ is the prior over $\boldsymbol{\theta}$.

- Bayes factors for model comparison, numerical analysis, statistical mechanics etc [Llorente et al., 2020].

Evidence Estimation

- Llorente et al. [2020] classifies evidence estimation into four categories:
    1. deterministic approximation,
    2. density estimation,
    3. importance sampling, and
    4. vertical representation, which encompasses **nested sampling** [Skilling, 2006].
- I'll show that the Riemann sum estimator due to Yakowitz et al. [1978] can be used in the context of nested sampling [Skilling, 2006] to achieve a $O(n^{-4})$ rate of convergence, faster than the usual Ergodic Central Limit Theorem.

Merge two ideas:

Riemann sum estimators [Yakowitz et al., 1978]

Merge two ideas:

Riemann sum estimators [Yakowitz et al., 1978]
+
Nested sampling [Skilling, 2006]
⇓

Merge two ideas:

Riemann sum estimators [Yakowitz et al., 1978]
+
Nested sampling [Skilling, 2006]
⇓

Quantile Importance Sampling: QIS

Riemann sum estimators

Naive Monte Carlo

- Naïve Monte Carlo or the *usual* importance sampling method:
- The empirical average of a sample $(x_1, \ldots, x_n)$ drawn from either the target density f or a proposal density g:

$$\hat{\psi}_{IS} = \frac{1}{n} \sum_{i=1}^{n} \frac{L(x_i) f(x_i)}{g(x_i)} \to \int L(x) f(x) dx. \qquad (1)$$

- $\hat{\psi}_{IS}$ converges to the true value ψ at a $O(n^{-1})$ rate.
- The density function $g(\cdot)$ is selected such that the weights $w(x_i)$ are approximately constant.
- Many, many alternatives and strategies to increase efficiency.

Riemann sums integrals over $[0, 1]$

- Yakowitz et al. [1978] proposed a weighted Monte-Carlo scheme.
- Consider the Riemann sum estimator:

$$\hat{\psi}_Y = \sum_{i=0}^{n-1} \frac{L(u_{(i)}) + L(u_{(i+1)})}{2}(u_{(i+1)} - u_{(i)}), \text{ for } \psi = \int_0^1 L(x)dx,$$

(2)

where $\{u_{(i)}\}_{i=1}^n$ are ordered uniform draws from $f = \mathbb{I}_{[0,1]}$.

- Yakowitz et al. [1978] proposed a weighted Monte-Carlo scheme.
- Consider the Riemann sum estimator:

$$\hat{\psi}_Y = \sum_{i=0}^{n-1} \frac{L(u_{(i)}) + L(u_{(i+1)})}{2} (u_{(i+1)} - u_{(i)}), \text{ for } \psi = \int_0^1 L(x)\,dx,$$

$$(2)$$

where $\{u_{(i)}\}_{i=1}^n$ are ordered uniform draws from $f = \mathbb{I}_{[0,1]}$.

Yakowitz (1978)

For $f \equiv \mathcal{U}(0, 1)$, the Riemann sum estimator $\hat{\psi}_Y$ in (2) attains a $O(n^{-4})$ rate of convergence.

- Yakowitz's estimator (2) was extended to a general (prior) density $f(\cdot)$ by Philippe [1997], Philippe and Robert [2001].

Riemann sums for general f

- Yakowitz's estimator (2) was extended to a general (prior) density $f(\cdot)$ by Philippe [1997], Philippe and Robert [2001].
- Take ordered samples:

$$x_{(1)} \doteq F^-(u_{(1)}) \leq x_{(2)} \doteq F^-(u_{(2)}) \leq \cdots \leq x_{(n)} \doteq F^-(u_{(n)}),$$

- Construct the Riemann sum estimator

$$\hat{\psi}_R = \sum_{i=0}^{n-1} L(x_{(i)})f(x_{(i)})(x_{(i+1)} - x_{(i)}). \qquad (3)$$

- This attains a convergence rate of $O(n^{-2})$, while admitting a vanishing bias of $O(n^{-1})$ order.
- Philippe [1997], Philippe and Robert [2001] argue this convergence is 'far from formal', i.e. noticeable within a short range.

Demonstration

Two integrals:

- Beta(a, b) integral, for $a = 3$, $b = 3$,

- $\mathbb{E}(1/(1 + U))$ for $U \sim \text{Exp}(1)$.

Demonstration

Two integrals:

- Beta(a, b) integral, for $a = 3$, $b = 3$,
- True ψ: Beta normalizing constant Beta$(3, 3)$.

- $\mathbb{E}(1/(1 + U))$ for $U \sim \text{Exp}(1)$.
- True ψ: $E_1(x) = \int_x^\infty e^{-t}/t\,dt$, and is given by:

$$\psi = \int_0^\infty \frac{e^{-x}}{1 + x}\,dx = e \cdot \int_1^\infty \frac{e^{-t}}{t}\,dt = e \cdot E_1(1).$$
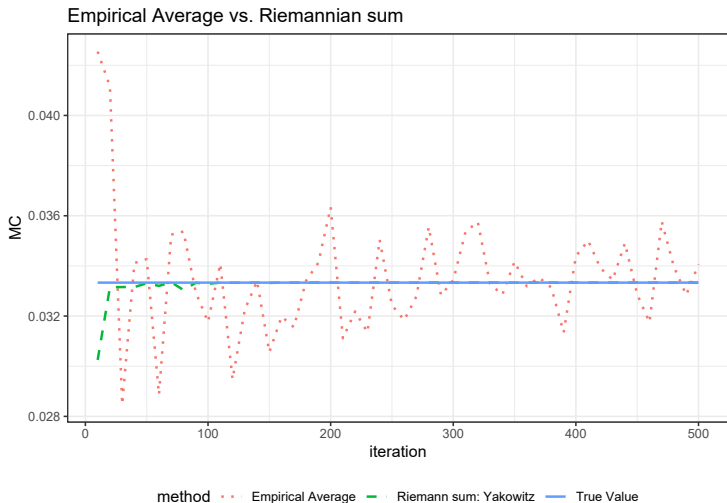
Riemann vs. Naïve: Beta
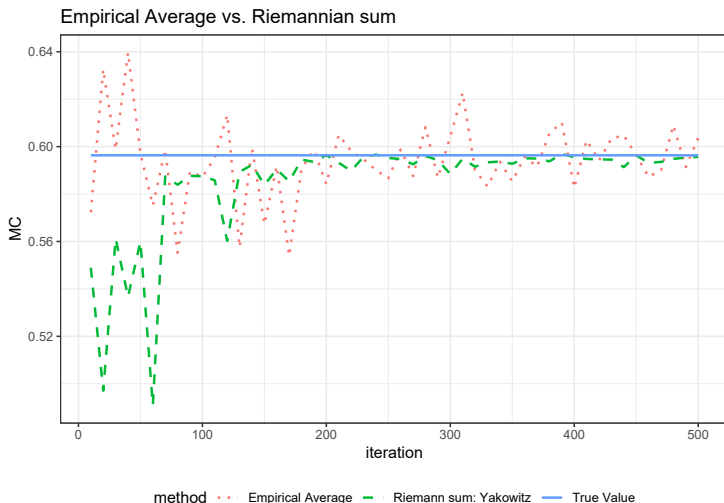

Figure 1: Estimating Beta$(3, 3)$

Figure 2: Estimating $\mathbb{E}(1/(1+U))$ for $U \sim \mathrm{Exp}(1)$

Riemann estimators

- Despite the much faster convergence rates, the Yakowitz estimator has limited use.
- Extending trapezoid rules to higher dimension would be thwarted by the 'curse of dimensionality'.
- An overlooked fact was that high-dimensional integrals can be written as uni-dimensional integrals over the positive real line or over the unit interval (Nested Sampling).
- **Nested sampling** is a perfect situation to invoke the Yakowitz estimator.

Nested Sampling

We start with the Lorenz identity that is at the heart of nested sampling and the quantile importance sampling.

Nested Sampling

- Let $\mathbf{X} \sim P(\mathbf{x})$ and $Y \equiv L(\mathbf{X})$, called the *likelihood ordinate*.
- Let $F_Y(y) = \mathbb{P}\{L(\mathbf{X}) \leq y\}$ be the CDF of Y, and the upper cumulant or the survival function for Y is:

$$Z(y) = 1 - F_Y(y) = \int_{L(\mathbf{x}) > y} \mathrm{d}P(\mathbf{x}). \tag{4}$$

- Let $\Lambda(s)$ denote the pseudo-inverse of $Z(y)$:

$$\Lambda(s) = \sup\{y : Z(y) > s\}, \tag{5}$$

Volume

We can write the evidence ψ (for any p with $\mathbf{x} \in \mathbb{R}^p$) as

$$\psi = \int_0^\infty Z(y)\mathrm{d}y = \int_0^1 \Lambda(s)\mathrm{d}s. \tag{6}$$

The function $Z(y) \in [0, 1]$ is also called the *volume variable* in Ashton et al. [2022], as it is the volume enclosed by the likelihood contour.

Lorenz identity

Lemma

Since $\mathbf{X} \sim P$, and $U \sim \mathcal{U}(0,1)$, it follows that $\Lambda(U) \stackrel{\mathrm{D}}{=} L(\mathbf{X})$, that is the likelihood ordinates are distributionally same as the $\Lambda(u)$ values at uniform grid points. It also follows from the definition of $\Lambda \equiv Z^{-1}$, that $Z\{L(\mathbf{X})\} \sim \mathcal{U}(0,1)$.

Proof

- To quickly see why (6) holds true, consider the following result:

$$\psi = \int_{\chi} L(x)\mathrm{d}P(x) = \int_{\chi} \int_0^{\infty} \mathbb{I}\{y < L(x)\}\mathrm{d}y\mathrm{d}P(x) = \int_0^{\infty} Z(y)\mathrm{d}y \tag{7}$$

Proof

- To quickly see why (6) holds true, consider the following result:

$$\psi = \int_{\mathcal{X}} L(x) \mathrm{d}P(x) = \int_{\mathcal{X}} \int_0^\infty \mathbb{I}\{y < L(x)\} \mathrm{d}y \mathrm{d}P(x) = \int_0^\infty Z(y) \mathrm{d}y \tag{7}$$

- Intuitively, $\Lambda(s)$ is the pseudo-inverse of $Z(y)$, i.e. it gives the value y such that s is the fraction of prior draws with likelihood ordinates larger than y.

Proof

- To quickly see why (6) holds true, consider the following result:

$$\psi = \int_{\mathcal{X}} L(x)\mathrm{d}P(x) = \int_{\mathcal{X}} \int_0^\infty \mathbb{I}\{y < L(x)\}\mathrm{d}y\mathrm{d}P(x) = \int_0^\infty Z(y)\mathrm{d}y \tag{7}$$

- Intuitively, $\Lambda(s)$ is the pseudo-inverse of $Z(y)$, i.e. it gives the value y such that s is the fraction of prior draws with likelihood ordinates larger than y.
- That is, $\{s < Z(y)\}$ if and only if $\{y < \Lambda(s)\}$.

$$\psi = \int_0^\infty \int_0^1 \mathbb{I}\{s < Z(y)\}\mathrm{d}s\mathrm{d}y = \int_0^\infty \int_0^1 \mathbb{I}\{y < \Lambda(s)\}\mathrm{d}s\mathrm{d}y = $$
$$\int_0^1 \Lambda(s)\mathrm{d}s. \tag{8}$$

Couple of remarks

1. Critically, we do not have to assume that either $F^{-1}(s)$ or $\Lambda(s)$ are available in closed form, as we can find an unbiased estimate of this by simulating the Lorenz curve.
2. It also follows from the Lorenz identity that the implied distribution of the likelihood ordinates are the same as the Lorenz curve $\Lambda(\cdot)$ evaluated at uniform grid points.

$$Z(L(\mathrm{X})) \stackrel{\mathbf{D}}{=} \mathcal{U}(0,1), \text{ and } \Lambda(U) \stackrel{\mathrm{D}}{=} L(\mathbf{X}).$$

Algorithm [Ashton et al., 2022]

> ### Basic idea
>
> "NS starts with an ensemble of n_{live} random locations Θ drawn from the prior, $\pi(\Theta)$, each of which has its likelihood $L(\Theta)$ which we can place in ascending order. Crudely, if we discarded the lowest half of the values, the survivors would be random samples taken within the restricted volume $L > median(L)$, which would statistically be roughly half the original volume. This allows us to make a statistical estimate of the volume variable $Z(y)$. Repeating that n_{iter} times would yield compression by a factor of about $2^{n_{iter}}$. This is the exponential behaviour required to overcome the curse of dimensionality"'.

Algorithm

Algorithm 1 Nested Sampling

Require: n_{live}, the number of live points; $\mathcal{L}$, the likelihood function; $p(\boldsymbol{\theta})$ prior.

Ensure: The evidence ψ and other quantities of interest.

1: Initialize $\psi = 0$, and prior area/volume $X_0 = 1$.
2: Generate n_{live} live points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim p(\boldsymbol{\theta})$.
3: Calculate the likelihood $\mathcal{L}_i = \mathcal{L}(\boldsymbol{\theta}_i)$, $i = 1, 2, \ldots, n_{live}$ for each live points.
4: **repeat**
5:      Identify the live point $\boldsymbol{\theta}_{i^*}$ with the lowest likelihood, call it $\mathcal{L}_{i^*}$.
6:      Remove $\boldsymbol{\theta}_{i^*}$ from the set of live points.
7:      Sample a new point $\boldsymbol{\theta}_{\text{new}}$ from $p(\boldsymbol{\theta})$ subject to the constraint $\mathcal{L} > \mathcal{L}_{i^*}$.
8:      Calculate the contraction factor ΔX_i enclosed by the likelihood contour, i.e.,
9:      $\Delta X_i = (1 - t)X_i$, where $t = \mathrm{e}^{-1/n_{live}}$ and $X_i = tX_{i-1}$.
10:      Update the evidence: $Z = Z + \mathcal{L}_{i^*}\Delta X_i$.
11:      $i \rightarrow i + 1$.
12: **until** stopping criterion is met, e.g. $\max \mathcal{L}(\theta)X_i > \mathrm{e}^{tol}Z$.
13: Adjust for the remainder of evidence $Z = Z + \frac{1}{n_{live}}\sum_{n=1}^{n_{live}} \mathcal{L}(\boldsymbol{\theta}_n)X_I$.
     return Evidence ψ and importance weights $p_i = \mathcal{L}_{i^*}\omega_i/Z$, $i = 1, \ldots, I$.

---

a | **The NS evidence identity.** The colours represent contours of a two-dimensional likelihood function. Rather than summing over little cubes (left), we combine cubes of similar likelihood together and sum over them (right).

b | **NS on a two dimensional problem.** We show the dead points and their iso-likelihood contours (left) and the corresponding contributions to the evidence integral (right). The volumes X_i are estimated statistically in NS.

c | **Compression in one iterate of NS.**

Figure 1 | **Illustrations of NS algorithm.**

Vertical Likelihood

Alternative presentation of nested sampling [Polson and Scott, 2014]:

The evidence $\psi = \int_0^1 \Lambda(s)\mathrm{d}s$ is approximated at $n_{\mathrm{iter}} \doteq n$ grid points $1 \equiv s_0 > s_1 > s_2 > \cdots > s_n > 0$) by a numerical integration rule:

$$\hat{\psi}_{NS} = \frac{1}{2} \sum_{i=1}^n w_i \Lambda(s_i) \text{ for } w_i = (s_{i-1} - s_i) \text{ or } w_i = \frac{1}{2}(s_{i-1} - s_{i+1}). \quad (9)$$

In the case of nested sampling, one can use the deterministic grid points $s_i = \exp(-i/n)$ for $i = 1, \ldots, m$ for a moderate n and a large m.

The rationale for the deterministic choice $s_i = \exp(-i/n)$ follows from two facts:

1. the 'volume' variable $Z(y)$ is uniformly distributed (see Lemma 1), and
2. the ith order statistics from $n\,\mathcal{U}(0,1)$ random sample has a $\mathcal{B}eta(i, n-i+1)$ distribution.

Quantile Importance Sampling

Yakowitz + Nested

- Now, recall the Yakowitz estimator i.e. the Riemann sum.
- Let $\{U_{(i)}\}_{i=1}^{n}$ denote n ordered draws from $\mathcal{U}(0,1)$, with augmented fixed boundary points: $U_{(0)} \equiv 0$ and $U_{(n+1)} \equiv 1$.
- We can think of these as stochastic grid points on $[0,1]$, with $s_i = U_{(n+1-i)}$, with $s_{n+1} \equiv U_{(0)}$, and $s_0 = U_{(n+1)}$.
- The Quantile Importance Sampling estimator is:

$$\hat{\psi}_{QIS} = \sum_{i=1}^{n} \frac{1}{2}(u_{(i+1)} - u_{(i-1)})\Lambda(u_{(i)})$$

To apply Yakowitz's method for nested sampling, the main trick is to **reorder the quantiles of the likelihood ordinates**.

1. Draw m samples from the prior $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim p(\mathbf{x})$,

QIS

To apply Yakowitz's method for nested sampling, the main trick is to **reorder the quantiles of the likelihood ordinates**.

1. Draw m samples from the prior $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim p(\mathbf{x})$,

2. Calculate the likelihood ordinates at the prior values: $Y_1 = L(x_1), \ldots, Y_m = L(x_m)$, and sort the sequence of Y_j's obtained thus, *i.e.* $Y_{(1)}, \ldots, Y_{(m)}$,

QIS

To apply Yakowitz's method for nested sampling, the main trick is to **reorder the quantiles of the likelihood ordinates**.

1. Draw m samples from the prior $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim p(\mathbf{x})$,

2. Calculate the likelihood ordinates at the prior values: $Y_1 = L(x_1), \ldots, Y_m = L(x_m)$, and sort the sequence of Y_j's obtained thus, *i.e.* $Y_{(1)}, \ldots, Y_{(m)}$,

3. Generate $n(< m)$ 'live' points from $\mathcal{U}(0,1)$, and sort them: $u_{(1)} < u_{(2)} < \ldots < u_{(n)}$. Augment $u_{(i)}$'s with $u_{(0)} = 0$, and $u_{(n+1)} = 1$.

QIS

To apply Yakowitz's method for nested sampling, the main trick is to **reorder the quantiles of the likelihood ordinates**.

1. Draw m samples from the prior $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim p(\mathbf{x})$,

2. Calculate the likelihood ordinates at the prior values: $Y_1 = L(x_1), \ldots, Y_m = L(x_m)$, and sort the sequence of Y_j's obtained thus, *i.e.* $Y_{(1)}, \ldots, Y_{(m)}$,

3. Generate $n(< m)$ 'live' points from $\mathcal{U}(0,1)$, and sort them: $u_{(1)} < u_{(2)} < \ldots < u_{(n)}$. Augment $u_{(i)}$'s with $u_{(0)} = 0$, and $u_{(n+1)} = 1$.

4. Evaluate the sample quantiles for Y at the points $u_{(i)}$:
$\Lambda(u_{(i)}) \stackrel{\mathrm{D}}{=} Y_{\lceil mu_{(i)} \rceil}$, $i = 1, \ldots, n$, with $\Lambda(0) = \Lambda_{\max}$ and $\Lambda(1) = \Lambda_{\min}$, and use the formula:

$$\hat{Z}_{QIS} = \left[\sum_{i=1}^{n+1} (u_{(i)} - u_{(i-1)}) \frac{\Lambda(u_{(i-1)}) + \Lambda(u_{(i)})}{2} \right]. \quad (10)$$

Rate of convergence

QIS can achieve a $O(n^{-4})$ rate of convergence for evidence using the results of [Yakowitz et al., 1978].

Lemma

Let the evidence $\psi \equiv \int_\chi L(\mathbf{x})dF(\mathbf{x}) = \int_0^1 \Lambda(s)ds$. *Assume that the cumulant* $\Lambda(s)$ *has a continuous derivative and that the second derivative* $\Lambda''(s)$ *exists and is bounded in absolute value over the unit interval. Let* $U_{(0)} \equiv 0$, $U_{(n+1)} \equiv 1$, *and* $\{U_{(i)}\}_{i=1}^n$ *be the ordered statistics from* $\mathcal{U}(0,1)$ *distribution (with* $U_{(i)} \geq U_{(i-1)}$, $i = 1, \ldots, n+1$*). Consider the QIS:*

$$\hat{\psi}_{QIS} = \frac{1}{2}\left(\sum_{i=1}^{n+1}(U_{(i)} - U_{(i-1)})\{\Lambda(U_{(i-1)}) + \Lambda(U_{(i)})\}\right).$$

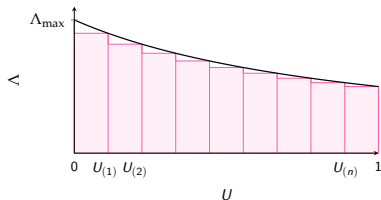*For some constant M,*

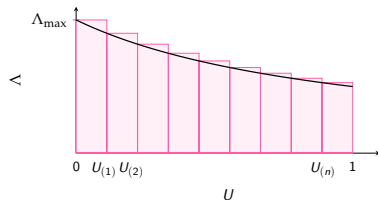$$\mathbb{E}[(\psi - \hat{\psi}_{QIS})^2] \leq M/n^4, \qquad \text{for all } n \geq 1.$$

Remarks

$\hat{\psi}_{QIS}$ admits natural upper and lower bounds:

$$\sum_{i=1}^{n}(U_{(i)} - U_{(i-1)})\Lambda(U_{(i)}) \leq \hat{\psi}_{QIS} \leq \sum_{i=1}^{n}(U_{(i+1)} - U_{(i)})\Lambda(U_{(i)}) + U_{(1)}\Lambda_{\max}.$$

(a) Lower bound for $\int_0^1 \Lambda(u)du$.

(b) Upper bound for $\int_0^1 \Lambda(u)du$.

Skilling [2006]: $\Lambda_{\max}$ is independent of the nested sampling algorithm, and there could be intervals of negligible size containing very large likelihood values (see Fig. 25) unless precluded by separate global analysis, and the error could be at most $O(n^{-1})$.

Example 1

As an illustrative example, we take Gaussian likelihood and Gaussian prior, *i.e.*,

$$L(x \mid \mu_1, \sigma_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}},$$

$$p(x \mid \mu_2, \sigma_2) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

The integral $\psi = \int L(x)p(x)dx$ is available in closed form.

Compare QIS with the traditional Monte Carlo approach, as well as the original nested sampling approach with both Riemann summation and rectangular summation.

We take $n = 20$ live points for for $m = 1000$ samples from the prior, and 100 replicates to calculate the root mean squared errors (RMSE), $\{1/r\sum_{j=1}^{r}(\hat{\psi}_j - \psi)\}^{1/2}$ and mean absolute predictive error (MAPE), $1/r\sum_{j=1}^{r}|(\hat{\psi}_j - \psi)/\psi|$.
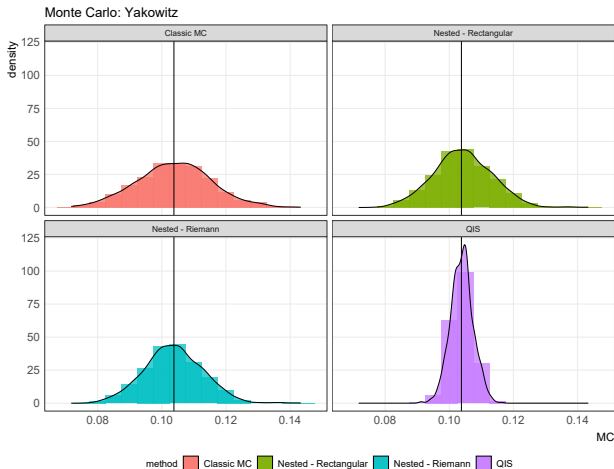
QIS vs. others

Figure 4: Comparison of three candidate schemes: Yakowitz or QIS (ordered uniform), Skilling's original scheme ($\exp(-i/N)$) and the classic Monte Carlo approach.

Multivariate t

As an example of a higher-dimensional integral, we look at a multivariate t likelihood and a multivariate Gaussian prior with dimension $d = 50$, following Polson and Scott [2014].

The target integral is:

$$\psi = \int_{\mathbb{R}^d} (1 + \frac{\mathbf{x}^T \mathbf{x}}{\nu})^{-\frac{\nu+d}{2}} (\frac{\tau}{2\pi})^{d/2} \exp\{-\tau \mathbf{x}^T \mathbf{x}/2\} d\mathbf{x}. \qquad (11)$$

This integral can be written in terms of Kummer's confluent hypergeometric function of the second kind: $\psi = s^a U(a, b, s)$, where $a = (\nu + d)/2$, $b = \nu/2 + 1$, $s = \nu\tau/2$.

For the specific values used here: $d = 50, \tau = 1$ and $\nu = 2$, we can get: $\psi = U(26, 2, 1) = 1.95 \times 10^{-29}$.
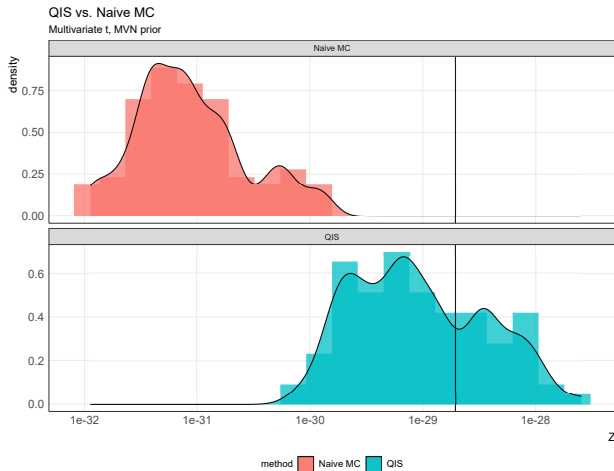
Figure 5: QIS vs. Naïve MC comparison for the multivariate *t*-multivariate Normal example.

Connections to Inverse Probability Weighting

Survey Sampling Estimators

In survey sampling, the goal is often to estimate the population mean ψ from a finite sample $(y_1, \ldots, y_n)$.

A common approach is to weight each observation y_i by a weight w_i inversely related to its probability of inclusion (e.g., probability proportional to selection, or PPS).

Examples:

- Simple estimator: $\hat{\psi} = \sum_{i \in s} w_i y_i / n$
- Ratio estimator: $\hat{\psi} = \sum_{i \in s} w_i y_i / \sum_{i \in s} w_i$

Two popular examples: **Horvitz–Thompson** estimator [Horvitz and Thompson, 1952]: unbiased using fixed weights, and **Hájek** estimator [Hájek, 1971]: normalizes weights for stability (slightly biased).

Ratio Estimator Analogy in Importance Sampling

In importance sampling, the goal is to estimate

$$\psi = \mathbb{E}_F\{I(X)\} = \int I(x)\,dF(x),$$

where F is difficult to sample from.

We instead draw samples from a candidate distribution G with density g, and estimate:

$$\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n} I(x_i)\frac{f(x_i)}{g(x_i)} \doteq \frac{1}{n}\sum_{i=1}^{n} I(x_i)w(x_i).$$

The density $g(\cdot)$ is chosen to make $w(x_i)$ nearly constant [Firth, 2011].

An analog to the ratio estimator is:

$$\hat{\psi} = \frac{\sum_{i=1}^{n} I(x_i)w(x_i)}{\sum_{i=1}^{n} w(x_i)}.$$

Introduced by Hesterberg [1988, 1995], and extended in Firth [2011]

Horvitz–Thompson and Hájek Estimators

- Population: $U = \{1, \ldots, N\}$ with inclusion probs p_k.
- HT estimator: $\hat{\psi}_{HT} = \frac{1}{N} \sum_{k \in s} Y_k / p_k$ [Horvitz and Thompson, 1952, Narain, 1951, Rao et al., 1999].
- Hájek estimator: $\hat{\psi}_H = \sum Y_k / p_k / \sum 1 / p_k$ [Hájek, 1971], normalized, doesn't require N.
- In missing data: $R_k \sim \text{Bernoulli}(p_k)$, and estimators are:
  - $\hat{S} = \sum Y_k R_k / p_k$, $\hat{n} = \sum R_k / p_k$
  - $\hat{\psi}_{HT} = \hat{S} / n$, $\quad \hat{\psi}_H = \hat{S} / \hat{n}$
- Trotter–Tukey estimator: $\hat{\psi}_{TT} = \hat{S} / \{(1 - \lambda)n + \lambda \hat{n}\}$ [Trotter and Tukey, 1956]
- Adaptive Normalization (AN): chooses λ from data to reduce variance [Khan and Ugander, 2023]

Desirable Properties and Design Sensitivity

- HT estimators are unbiased, admissible, consistent [Ramakrishnan, 1973, Delevoye and Sävje, 2020].
- Fail under heavy-tailed/skewed designs (e.g., in natural experiments).
- IPW estimators like HT are design-based: weights $\propto 1/p_k$ [Neyman, 1934, Little, 2008].
- Hájek often preferred due to lower variance, especially when:
  - y_k close to ψ (e.g., all equal).
  - Sample size varies with fixed p_k.
  - y_k negatively correlated with p_k.
- Supported in Särndal et al. [2003]; circus example in Basu [1988].

Bias-Variance Trade-offs and Importance Sampling

- HT criticized in high-dim nonresponse settings (e.g., Wasserman [2004], Robins and Ritov [1997]).
- Bayesian fixes: slightly biased but lower-variance estimators [Sims, 2010, Ghosh, 2015, Li, 2010, Harmeling and Touissant, 2007].
- Classic bias–variance trade-off applies.
- **Same problems in importance sampling:**
  - Large variability in $w(x_i) = f(x_i)/g(x_i)$ leads to unstable estimators.
  - Causes: heavy tails, poor choice of g.
  - Adaptive normalization [Khan and Ugander, 2023] helps reduce variance in both survey and MC settings.

Envoi: Revisiting Riemann Sums for Nested Sampling

- We revisited Riemann sum estimators in the context of nested sampling, a popular tool for high-dimensional evidence estimation.

- By combining quantile reordering with the Lorenz identity, we proposed **Quantile Importance Sampling (QIS)**.

- Under mild regularity, QIS achieves a faster convergence rate $O(n^{-4})$, outperforming standard MC methods.

- QIS avoids needing a closed-form $\Lambda(s)$ or $Z(L(x))$—sample quantiles suffice.

- *Caveat*: If the maximum likelihood ordinate is unknown, the final term may degrade to $O(n^{-1})$.

- **Importantly, the analogy with inverse probability weighting (IPW)** shows that connections between survey sampling and Bayesian evidence estimation remain a *fruitful and largely untapped area*.

Takeaway: QIS provides a simple yet powerful alternative to classical importance sampling—connecting ideas from numerical integration, nested sampling, and survey inference.

Thank you!

These ideas are explored in greater detail in the following papers:

- `arXiv:2305.03158` – **Adaptive Normalization for Inverse Probability Weighting**
- `arXiv:2204.14121` – **Quantile Importance Sampling: Fast Evidence Estimation via Riemann Reweighting**

Joint work with

Nick Polson

References

Greg Ashton, Noam Bernstein, Johannes Buchner, Xi Chen, Gábor Csányi, Andrew Fowlie, Farhan Feroz, Matthew Griffiths, Will Handley, Michael Habeck, et al. Nested sampling for physical scientists. *Nature Reviews Methods Primers*, 2(1):39, 2022.

Dev Basu. *Statistical information and likelihood: a collection of critical essays by Dr. D. Basu (J. K. Ghosh, Ed.)*, volume 45. Lecture Notes in Statistics, Springer Science & Business Media, 1988.

Angèle Delevoye and Fredrik Sävje. Consistency of the horvitz–thompson estimator under general sampling and experimental designs. *Journal of Statistical Planning and Inference*, 207:190–197, 2020.

David Firth. On improved estimation for importance sampling. *Brazilian Journal of Probability and Statistics*, 25(3):437–443, 2011.

Jayanta K Ghosh. Weak paradoxes and paradigms. In *Statistical Paradigms: Recent Advances and Reconciliations*, pages 3–12. World Scientific, 2015.

Stefan Harmeling and M Touissant. Bayesian estimators for robins-ritov's problem. Technical report, 2007.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

Timothy Classen Hesterberg. *Advances in importance sampling*. PhD thesis, Stanford University, 1988.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Jaroslav Hájek. Comment on "an essay on the logical foundations of survey sampling" by basu. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, pages 236–242. Holt, Rinehart and Winston, 1971.

Samir Khan and Johan Ugander. Adaptive normalization for ipw estimation. *Journal of Causal Inference*, 11(1):20220019, 2023.

Longhai Li. Are bayesian inferences weak for wasserman's example? *Communications in Statistics—Simulation and Computation®*, 39(3): 655–667, 2010.

Roderick J Little. Weighting and prediction in sample surveys. *Calcutta Statistical Association Bulletin*, 60(3-4):147–167, 2008.

Fernando Llorente, Luca Martino, David Delgado, and Javier Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *arXiv preprint arXiv:2005.08334*, 2020.

RD Narain. On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3(2):169–175, 1951.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 97(4):558–606, 1934.

Anne Philippe. Processing simulation output by riemann sums. *Journal of Statistical Computation and Simulation*, 59(4):295–314, 1997.

Anne Philippe and Christian P Robert. Riemann sums for mcmc estimation and convergence monitoring. *Statistics and Computing*, 11 (2):103–115, 2001.

Nicholas G Polson and James G Scott. Vertical-likelihood monte carlo. *arXiv preprint arXiv:1409.3601*, 2014.

MK Ramakrishnan. An alternative proof of the admissibility of the horvitz-thompson estimator. *The Annals of Statistics*, 1(3):577–579, 1973.

Jon NK Rao, Arijit Chaudhuri, JL Eltinge, Robert E Fay, JK Ghosh, Malay Ghosh, P Lahiri, and Danny Pfeffermann. Some current trends in sample survey theory and methods (with discussion). *Sankhyā: The Indian Journal of Statistics, Series B*, pages 1–57, 1999.

James M Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine*, 16(3):285–319, 1997.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

Chris Sims. Understanding non-bayesians. *Unpublished chapter, Department of Economics, Princeton University*, 2010.

John Skilling. Nested sampling for general bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.

Hale F Trotter and HF Tukey. Conditional monte carlo for normal samples. In *Proc. Symp. on Monte Carlo Methods*, pages 64–79. John Wiley and Sons, 1956.

Larry Wasserman. Bayesian inference. In *All of Statistics*, pages 175–192. Springer New York, NY, 2004.

S Yakowitz, JE Krimmel, and F Szidarovszky. Weighted monte carlo integration. *SIAM Journal on Numerical Analysis*, 15(6):1289–1300, 1978.