Hoff Ch.4 : Monte Carlo

- **Previously.** Conjugacy led to posterior distributions with simple, closed-form expressions for, e.g., mean and variance.
  - Beta + Binomial = Beta
  - Gamma + Poisson = Gamma

However, even for these simple conjugate models, posterior inference can be hard.

Looking back at the dental health CDC data:

- 48 40 year old men with 23 unhealthy teeth total
- 57 40 year old women w/ 33 unhealthy teeth total.
- Gamma(1,1) prior

$$\Rightarrow p(\theta_m | \vec{y}_m) = \text{Gamma}(24, 49)$$

$$p(\theta_w | \vec{y}_w) = \text{Gamma}(34, 58)$$

Probability that $\theta_w > \theta_m$:

$$\mathbb{P}(\theta_w > \theta_m \mid \vec{y}_m, \vec{y}_w) = \int_0^\infty \int_0^{\theta_w} p(\theta_m, \theta_w \mid \vec{y}_m, \vec{y}_w)\, d\theta_m d\theta_w$$

$$= \int_0^\infty \int_0^{\theta_w} p(\theta_m \mid \vec{y}_m)\, p(\theta_w \mid \vec{y}_w)\, d\theta_m d\theta_w$$

Assuming θ_m, θ_w conditionally independent given data

$$= \frac{49^{24}\ 58^{34}}{\Gamma(24)\Gamma(34)} \int_0^\infty \int_0^{\theta_w} \theta_m^{23} e^{-49\theta_m} \theta_w^{33} e^{-58\theta_w}\, d\theta_m d\theta_w.$$

Problem. Closed-form expressions often unavailable for posterior quantities of interest.

Integrals of the form

$$\mathbb{E}[f(\theta)|y] = \int_{\Theta} f(\theta)\, p(\theta|y)\, d\theta$$

often <u>extremely</u> hard to compute exactly.

- **Law of large numbers.** If $X_1, \ldots, X_S$ are iid copies of r.v. X, then

$$\lim_{S \to \infty} \frac{1}{S} \sum_{i=1}^{S} X_i = \mathbb{E}[X].$$

- More generally, for (almost) any function f,

$$\lim_{S \to \infty} \frac{1}{S} \sum_{i=1}^{S} f(X_i) = \mathbb{E}[f(X)].$$

Side note: For any $\mathcal{Z}$-valued random variable Z and any event $A \subseteq \mathcal{Z}$, we have

$$\mathbb{P}(Z \in A) = \mathbb{E}[\mathbb{1}(Z \in A)]$$

where $\mathbb{1}(Z \in A)$ is the <u>indicator function</u>

$$\mathbb{1}(Z \in A) = \begin{cases} 1 & \text{if } Z \in A \\ 0 & \text{if } Z \notin A \end{cases}.$$

Computing $\mathbb{P}(\theta_w > \theta_m \mid \vec{y}_m, \vec{y}_w)$.

- Draw S samples $\theta_{m,i} \overset{iid}{\sim} \text{Gamma}(24, 49)$
- Draw S samples $\theta_{w,i} \overset{iid}{\sim} \text{Gamma}(34, 58)$
- $\mathbb{P}(\theta_w > \theta_m \mid \vec{y}_m, \vec{y}_w) = \mathbb{E}\left[\mathbb{1}(\theta_w > \theta_m) \mid \vec{y}_m, \vec{y}_w\right]$

$$\approx \frac{1}{S} \sum_{i=1}^{S} \mathbb{1}(\theta_{w,i} > \theta_{m,i}).$$

Monte Carlo: LLN for posterior inference

- Suppose we draw S iid samples

$$\theta_i \overset{\text{iid}}{\sim} p(\theta \mid y).$$

Then for large S, by LLN,

$$\frac{1}{S}\sum_{i=1}^{S} f(\theta_i) \approx \int_{\Theta} f(\theta)\, p(\theta \mid y)\, d\theta.$$

- Posterior mean:

$$\bar{\theta}_S = \frac{1}{S}\sum_{i=1}^{S}\theta_i \longrightarrow \mathbb{E}[\theta|y]$$

- Posterior variance:

$$\frac{1}{S-1}\sum_{i=1}^{S}(\theta_i - \bar{\theta}_S)^2 \longrightarrow \mathrm{Var}[\theta|y]$$

- Probability $\theta \leq c$:

$$\frac{1}{S}\sum_{i=1}^{S} \mathbb{1}(\theta_i \leq c) \to \mathbb{E}\left[\mathbb{1}(\theta \leq c) \mid y\right]$$

$$= \mathbb{P}(\theta \leq c \mid y).$$

- Quantiles. $\alpha \in (0,1)$

The α quantile of $\theta \mid y$ is a value q_α such that

$$\mathbb{P}(\theta \leq q_\alpha \mid y) = \alpha .$$

This definition requires some assumptions about the distribution of X, but we'll skip the more general definition for simplicity.

To estimate q_α using Monte Carlo:

(i) Draw $\theta_i \overset{iid}{\sim} p(\theta \mid y)$, $i = 1, \ldots, S$

(ii) Pick $q_\alpha(S)$ so that $100\alpha\%$ of samples are $\leq q_\alpha(S)$.

$$\Rightarrow \quad q_\alpha(S) \longrightarrow q_\alpha \quad \text{as} \quad S \longrightarrow \infty .$$

Bayesian stats in a nutshell.

- Come up with model
  - Prior, likelihood
- Compute posterior, $p(\theta|y)$
- Draw samples $\theta_1, \ldots, \theta_S$ from $p(\theta|y)$
- Make inferences via Monte Carlo.

Monte Carlo error

- How much does $\bar{X}_S = \frac{1}{S}\sum_{i=1}^{S} X_i$ fluctuate around its mean, $\mathbb{E}[\bar{X}_S] = \mathbb{E}[X] := \mu$?

Set $\sigma^2 = \mathrm{Var}(X)$. Since the X_i are iid copies of r.v. X,

$$\mathbb{E}\left[(\bar{X}_S - \mu)^2\right] = \mathrm{Var}(\bar{X}_S) = \frac{1}{S^2}\sum_{i=1}^{S}\mathrm{Var}(X_i) = \frac{1}{S}\sigma^2$$

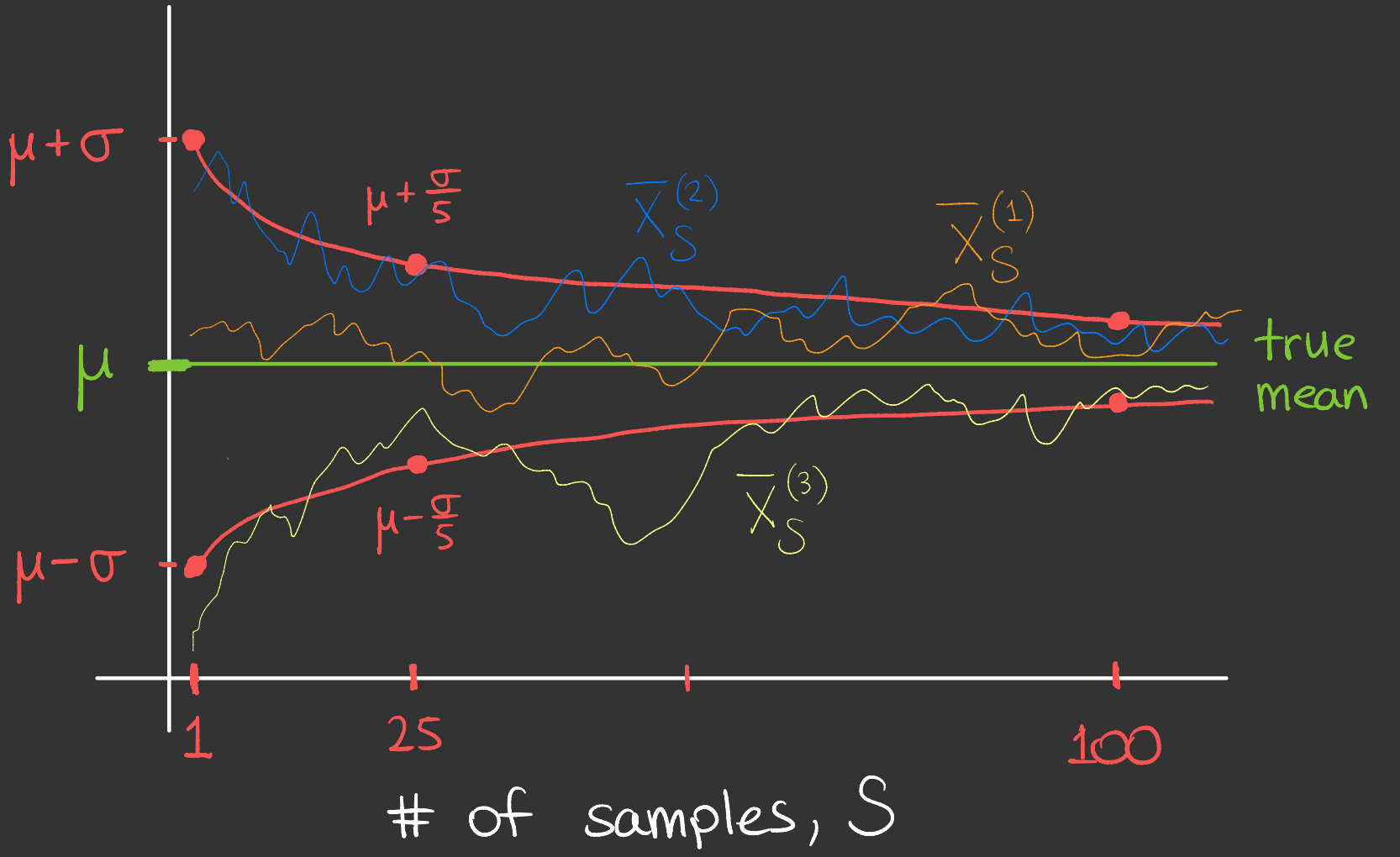
$\Longrightarrow$ Root mean squared error $= \dfrac{\sigma}{\sqrt{S}}$.

- If we approximate $\sigma^2 = \mathrm{Var}(X)$ by

$$\hat{\sigma}_S^2 = \frac{1}{S-1} \sum_{i=1}^{S} (X_i - \bar{X}_S)^2$$

then the Monte Carlo standard error is

$$\frac{\hat{\sigma}_S}{\sqrt{S}} .$$


μ+σ
μ
μ−σ
μ+σ/5
μ−σ/5
$\bar{X}_S^{(1)}$
$\bar{X}_S^{(2)}$
$\bar{X}_S^{(3)}$
true mean
1
25
100
# of samples, S

- The Central Limit Theorem (CLT) gives an even more complete characterization of MC error.

CLT. If $X_1, X_2, \ldots$ iid copies of X with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$, then

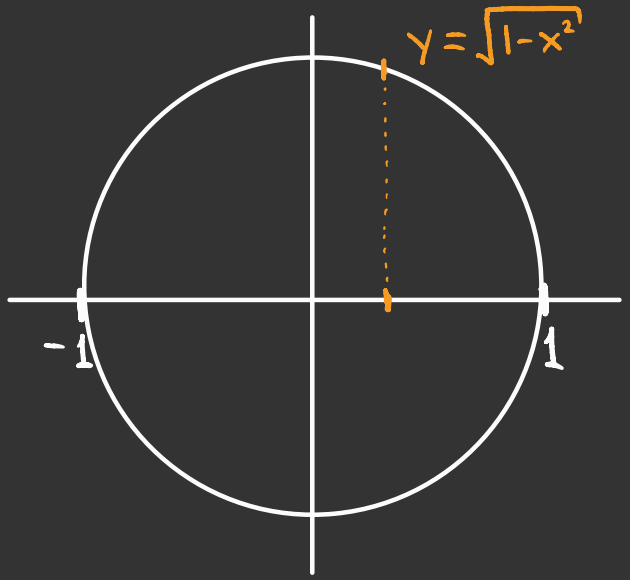
$$\sqrt{S}\left(\bar{X}_S^* - \mu\right) \xrightarrow{\text{distribution}} N(0, \sigma^2).$$

Equivalently, $$\frac{\sqrt{S}}{\sigma}\left(\bar{X}_S - \mu\right) \xrightarrow{d} N(0,1).$$

* Here, $\bar{X}_S = \frac{1}{S}\sum_{i}^{S} X_i$

A non-Bayesian example.

- Goal: Find the area of a circle of radius 1.
- One approach is to use calculus

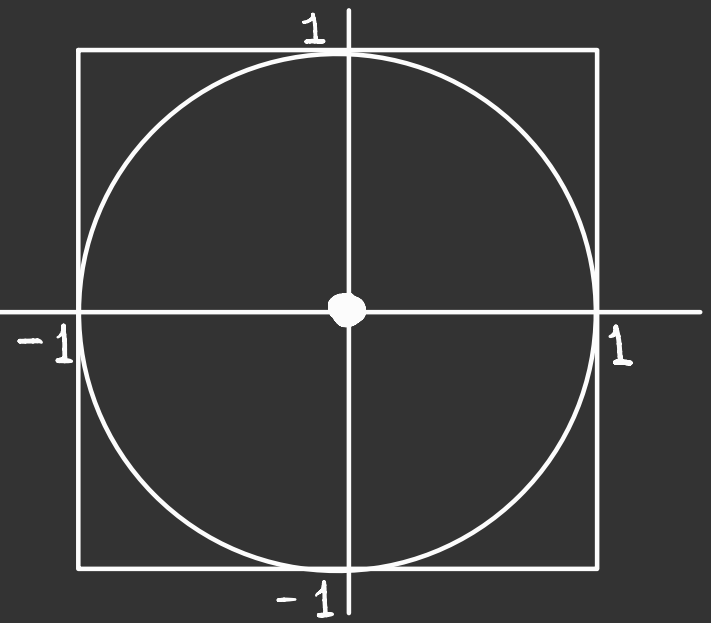


$$\text{Area} = 4\int_0^1 \sqrt{1-x^2}\,dx$$

Use trig rules to solve.

- Monte Carlo approach

Circle or ball of radius 1



- Let $B = \{(x,y) : x^2 + y^2 < 1\}$
- Let S be the square around B (side length 2)
- Let $X, Y \overset{iid}{\sim} \text{Uniform}(-1,1)$
- Then:

$$\text{Area}(B) = \mathbb{P}((X,Y) \in B)\,\text{Area}(S)$$
$$= 4\,\mathbb{P}((X,Y) \in B).$$

$$\text{Area}(B) = 4\,\mathbb{P}((X,Y)\in B) = 4\,\mathbb{E}\left[\mathbb{1}_{((X,Y)\in B)}\right]$$

By LLN, if we draw S samples $X_i \overset{iid}{\sim} \text{Unif}(-1,1)$ and $Y_i \overset{iid}{\sim} \text{Unif}(-1,1)$, then for large S

$$\text{Area}(B) \approx \frac{1}{S}\sum_{i=1}^{S} 4\,\mathbb{1}((X_i,Y_i)\in B)$$

$$= \frac{1}{S}\sum_{i=1}^{S} 4\,\mathbb{1}(X_i^2 + Y_i^2 < 1).$$

More generally, to approximate volume of unit ball

$$B_1(0) = \left\{ \vec{x} \in \mathbb{R}^d : \sum_{i=1}^{d} x_i^2 < 1 \right\},$$

generate $S \times d$ matrix

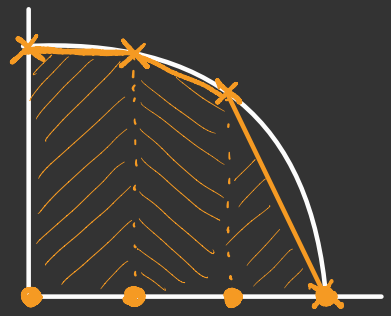
$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1d} \\ \vdots & & & \vdots \\ X_{S1} & X_{S2} & \cdots & X_{Sd} \end{pmatrix}$$

where each $X_{ij} \overset{iid}{\sim} \text{Unif}(-1, 1)$.

Then

$$\mathrm{Vol}\left(B_1(0)\right) \approx \frac{2^d}{S} \sum_{i=1}^{S} \mathbb{1}\left(X_{i1}^2 + \cdots + X_{id}^2 < 1\right).$$

Monte Carlo vs. deterministic grid.



- Grid approach to numerical integration

$$\int_a^b f(x)\,dx$$

  - Split $[a,b]$ into $(x_0, x_1, \ldots, x_g)$ where $x_0 = a$ and $x_g = b$ and $x_i < x_{i+1} \quad \forall\, i = 0, \ldots, g-1$.
  - Then

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{g-1} f(x_i)(x_{i+1} - x_i)$$

In d-dimensions, have g^d grid points!

$5^1 = 5$

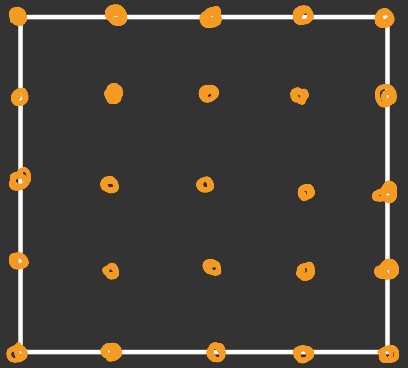
$5^2 = 25$

$5^3 = 125$

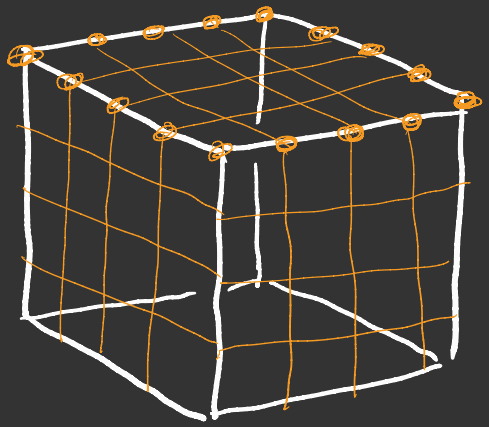
$5^{10} \approx 10{,}000{,}000$

$5^{20} \approx 100$ trillion

$$\int_a^b f(x)\,dx$$



$$\int_{a_1}^{b_1}\int_{a_2}^{b_2} f(x,y)\,dx\,dy$$



$$\int_{a_1}^{b_1}\int_{a_2}^{b_2}\int_{a_3}^{b_3} f(x,y,z)\,dx\,dy\,dz$$

Monte Carlo convergence rate is $\frac{1}{\sqrt{S}}$.

Depends only on # of samples, "dimension free"

Beta-binomial example.

- Prior: $p(\theta) = \text{Beta}(1,1)$
- Observations $Y_1, \ldots, Y_{10} | \theta \overset{iid}{\sim} \text{Bernoulli}(\theta)$ with
$$\sum Y_i = 3$$
- $p(\theta | \vec{Y}_n) = \text{Beta}(4,8)$.
- Since the posterior has closed form, we can compute certain posterior quantities exactly. Let's compare these to Monte

Carlo approximations.

- Posterior mean : $\mathbb{E}[\theta | \vec{Y}_n] = \dfrac{a + \Sigma y_i}{a + b + n}$

- Posterior variance

- $\mathbb{P}(\theta < \frac{3}{4} | y)$

- A 0.1 quantile

Posterior inference of arbitrary functions

If interested in a function $g(\theta)$ rather than θ itself, can generate iid samples via

for $1 \leq i \leq S$:

- Sample $\theta_i \overset{iid}{\sim} p(\theta|y)$
- Set $\gamma_i = g(\theta_i)$.

Then $\gamma_1, \ldots, \gamma_S \overset{iid}{\sim} p(g(\theta)|y)$.

- Example (see Hoff, Section 4.2 for details):

$$\log \text{odds}(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$$

Sampling from joint posterior

If $\vec{\theta}_d = (\theta_1, \cdots, \theta_d) \in \mathbb{R}^d$ AND the parameters θ_j are conditionally independent given the data y, then we can generate iid samples $\vec{\theta}_{d,i} \overset{iid}{\sim} p(\vec{\theta}_d | y)$ via

- for $1 \leq i \leq S$:
  - for $1 \leq j \leq d$:
    - draw $\theta_{ij} \sim p(\theta_j | y)$.

Sampling from posterior predictive distribution

PPD is (assuming $Y_1, \ldots, Y_{n+1}$ cond. ind. given θ)

$$p(y_{n+1} \mid \vec{y}_n) = \int_{\Theta} p(y_{n+1} \mid \theta)\, p(\theta \mid \vec{y}_n)\, d\theta.$$

for $1 \leq i \leq S$:

- Draw $\theta_i \sim p(\theta \mid \vec{y}_n)$
- Draw $\tilde{y}_i \sim p(y_{n+1} \mid \theta_i)$

$\Rightarrow \{(\theta_1, \tilde{y}_1), \ldots, (\theta_S, \tilde{y}_S)\} = S$ independent

samples from $p(\theta, y_{n+1} \mid \vec{y}_n)$ and

$\{\tilde{y}_1, \cdots, \tilde{y}_S\}$ iid samples from $p(y_{n+1} \mid \vec{y}_n)$.

Posterior predictive model checking

- How plausible is our model?
- Suppose we're interested in some test statistic $t(\vec{Y}_n)$ of the data, e.g.,

$$t(\vec{Y}_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i .$$

If our proposed Bayesian model

$$p(\theta|\vec{y}_n) \propto p(\vec{y}_n|\theta)p(\theta)$$

is reasonable, the test values $t(\vec{Y}_n^{ppd})$ w/ datasets $\vec{Y}_n^{ppd}$ drawn from the PPD

$$p(\vec{y}_n^{ppd}|\vec{y}_n) = \int_{\Theta} p(\vec{y}_n^{ppd}|\theta)\, p(\theta|\vec{y}_n)\, d\theta$$

Should resemble the observed $t(\vec{Y}_n)$.

Basic idea:

- Simulate datasets from PPD.
- Evaluate t on simulated datasets.
- Check whether observed t matches those evaluated on simulated data.

Monte Carlo approach.

for $1 \leq i \leq S$:

- Sample $\theta_i \sim p(\theta \mid \vec{y}_n)$
- Sample $\vec{y}^{ppd}_{n,i} = (y^{ppd}_{1i}, \ldots, y^{ppd}_{n,i}) \overset{iid}{\sim} p(y \mid \theta_i)$
- Compute $t_i = t(\vec{y}^{ppd}_{n,i})$

Return: Samples $\{t_1, \ldots, t_S\}$ from PPD of $t(\vec{Y}^{ppd}_n)$

See dental health code for example.

End Ch. 4