Bayes via forward simulation: Approximate Bayesian Computation

Jessi Cisewski-Kehe
Yale University

231st Meeting of the American Astronomical Society
January 7, 2018

What is Approximate Bayesian Computing?

- "Likelihood-free" approach
- Works by simulating from the forward process

What is Approximate Bayesian Computing?

- "Likelihood-free" approach
- Works by simulating from the forward process

Why not just use the likelihood?

The posterior for θ given observed data x_{obs}:

$$\pi(\theta \mid x_{\text{obs}}) = \frac{f(x_{\text{obs}} \mid \theta)\pi(\theta)}{\int f(x_{\text{obs}} \mid \theta)\pi(\theta)d\theta} = \frac{f(x_{\text{obs}} \mid \theta)\pi(\theta)}{f(x_{\text{obs}})}$$

Approximate Bayesian Computation

- "Likelihood-free" approach to approximating $\pi(\theta \mid x_{\text{obs}})$ ($f(x_{\text{obs}} \mid \theta)$ not specified)
- Proceeds via simulation of the forward process

Why would we not know $f(x_{\text{obs}} \mid \theta)$?

1. Physical model too complex
2. Strong dependency in data
3. Observational limitations

ABC for Astronomy

- cosmoabc: Likelihood-free inference via Population Monte Carlo Approximate Bayesian Computation (Ishida et al., 2015)
- Approximate Bayesian Computation for Forward Modeling in Cosmology (Akeret et al., 2015)
- Likelihood-Free Cosmological Inference with Type Ia Supernovae: Approximate Bayesian Computation for a Complete Treatment of Uncertainty (Weyant et al., 2013)
- Likelihood - free inference in cosmology: potential for the estimation of luminosity functions (Schafer and Freeman, 2012)
- Approximate Bayesian Computation for Astronomical Model Analysis: A case study in galaxy demographics and morphological transformation at high redshift (Cameron and Pettitt, 2012)

Basic ABC algorithm

For the observed data x_{obs} and prior $\pi(\theta)$:

Algorithm*

1. Sample θ_{prop} from prior $\pi(\theta)$
2. Generate x_{prop} from forward process $F(x \mid \theta_{\text{prop}})$
3. Accept θ_{prop} if $x_{\text{obs}} = x_{\text{prop}}$
4. Return to step 1

*Introduced in Tavaré et al. (1997) and Pritchard et al. (1999)

Binomial illustration

- Data are a sample of 1's and 0's coming from $Y_i \sim \text{Bern}(p)$ where $n =$ sample size, $p = P(Y = 1)$.
- Likelihood is $L(p \mid y) = \binom{n}{y} p^y (1-p)^{n-y}$, where $y = \sum_{i=1}^{n} y_i$ (but we will pretend we do not know this).

  Need to determine a distance function, ρ. Use the following:

$$\rho(y, x) = \frac{1}{n}|y - x|$$

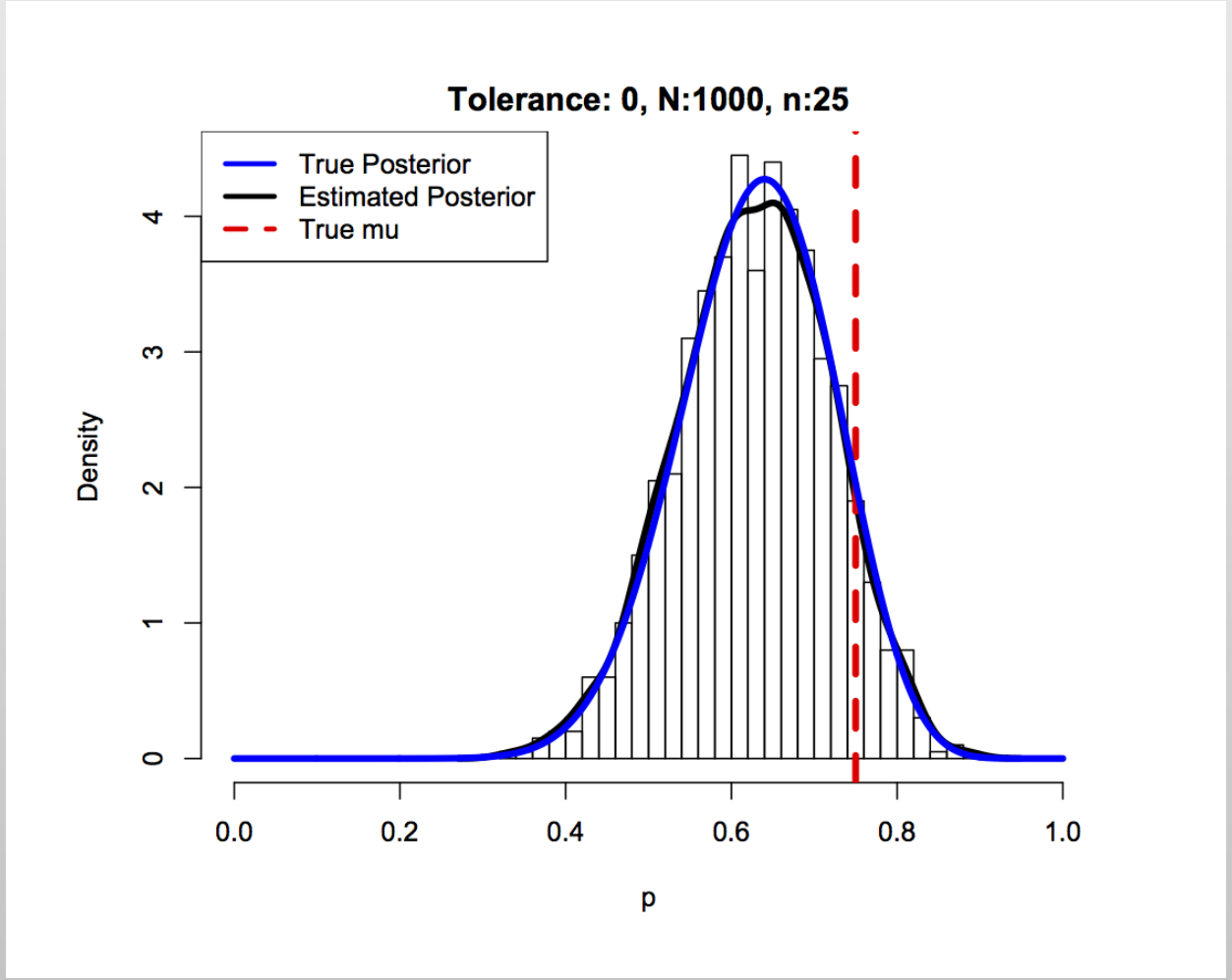
  Hence $\rho(y, x) = 0$ if the generated dataset x has the same number of 1's as y.

Binomial illustration: R code

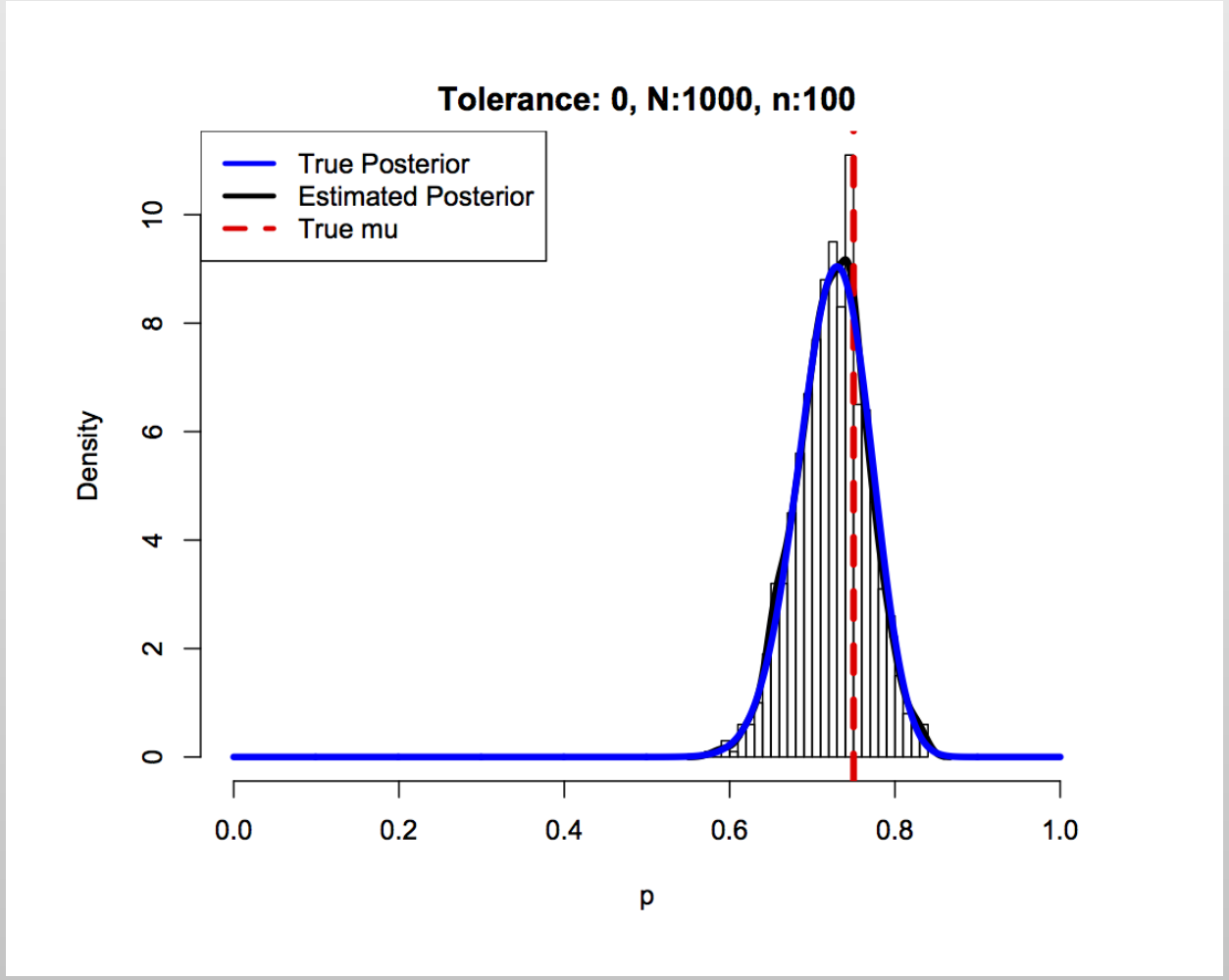
```
n <- 1000  #number of observations
N <- 1000  #generated sample size
true.p <- .75
data <- rbinom(n,1,true.p)
epsilon <- 0
alpha.hyper <- 1
beta.hyper <- 1
p <- numeric(N)
rho <- function(y,x) abs(sum(y)-sum(x))/n
for(i in 1:N){
    d <- epsilon+1
    while(d>epsilon) {
        proposed.p <- rbeta(1,alpha.hyper,beta.hyper)
        x <- rbinom(n,1,proposed.p)
        d <- rho(data,x)}
    p[i] <- proposed.p}
```

Reference: Turner and Zandt (2012)

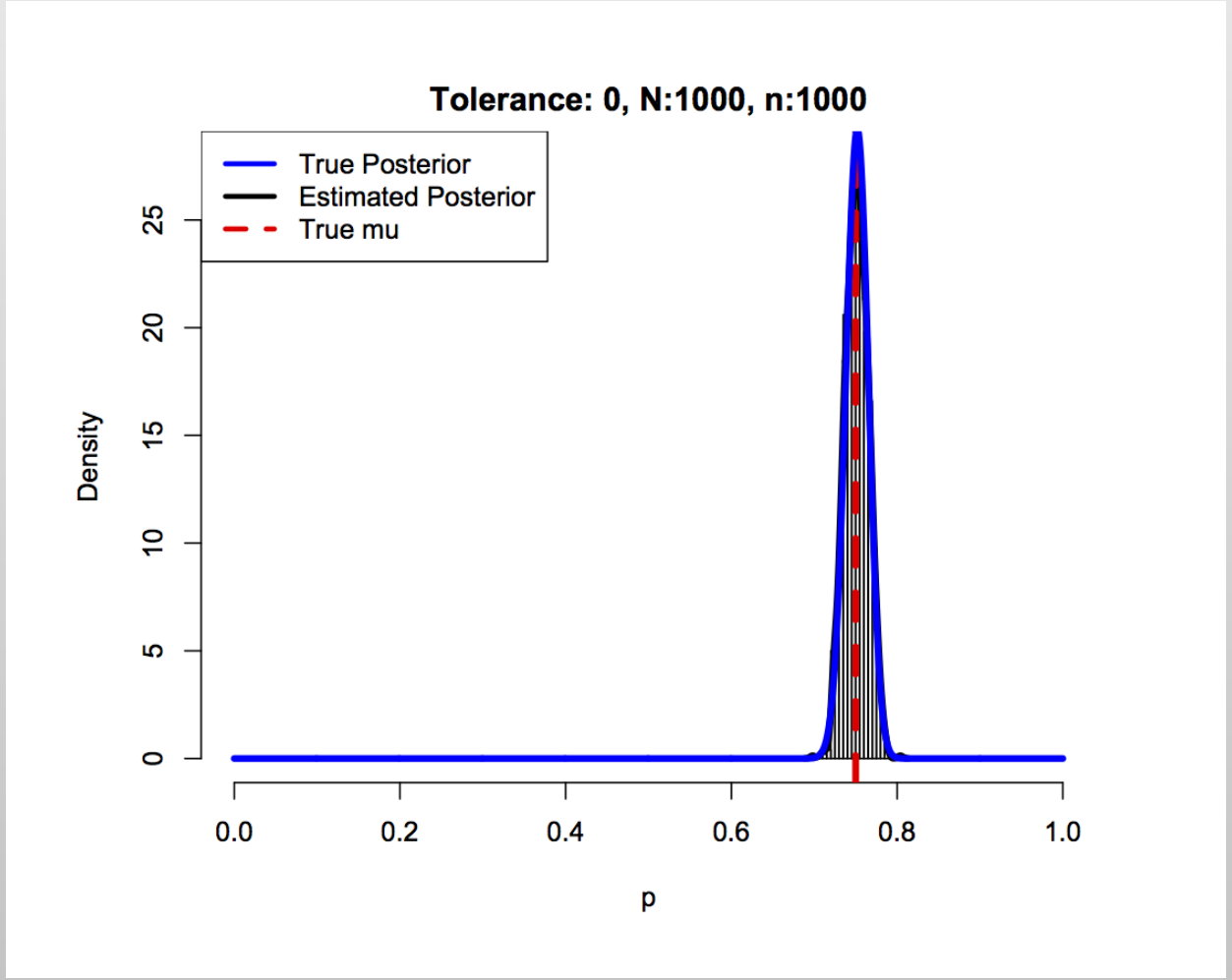
Binomial illustration: posterior



Binomial illustration: posterior



Binomial illustration: posterior



If one wants to generate a draw from the posterior:

1. Draw θ_{prop} from the prior $\pi(\theta)$.
2. Draw x_{prop} from the density $f(x \mid \theta_{\text{prop}})$.
3. Accept θ_{prop} if $x_{\text{prop}} = x_{\text{obs}}$.

Why?

If one wants to generate a draw from the posterior:

1. Draw θ_{prop} from the prior $\pi(\theta)$.
2. Draw x_{prop} from the density $f(x \mid \theta_{\text{prop}})$.
3. Accept θ_{prop} if $x_{\text{prop}} = x_{\text{obs}}$.

Why? Let θ_{acc} denote an accepted θ_{prop}. Then, for any θ,

$$\begin{aligned} P(\theta_{\text{acc}} = \theta) &= P(\theta_{\text{prop}} = \theta \mid x_{\text{prop}} = x_{\text{obs}}) \\ &= P(x_{\text{prop}} = x_{\text{obs}} \mid \theta_{\text{prop}} = \theta) P(\theta_{\text{prop}} = \theta) \Big/ P(x_{\text{prop}} = x_{\text{obs}}) \\ &= f(x_{\text{obs}} \mid \theta)\pi(\theta) \Big/ f(x_{\text{obs}}) = \pi(\theta \mid x_{\text{obs}}) \end{aligned}$$

in the case where θ **is discrete.**

Illustration from Chad Schafer (CMU)

If one wants to generate a draw from the posterior:

1. Draw θ_{prop} from the prior $\pi(\theta)$.
2. Draw x_{prop} from the density $f(x \mid \theta_{\text{prop}})$.
3. Accept θ_{prop} if $x_{\text{prop}} = x_{\text{obs}}$.

Why? Let θ_{acc} denote an accepted θ_{prop}. Then, for any $T \subseteq \mathbb{R}$,

$$\begin{aligned} P(\theta_{\text{acc}} \in T) &= P(\theta_{\text{prop}} \in T \mid x_{\text{prop}} = x_{\text{obs}}) \\ &= \int_T P(x_{\text{prop}} = x_{\text{obs}} \mid \theta_{\text{prop}} = \theta)\pi(\theta)\, d\theta \Big/ P(x_{\text{prop}} = x_{\text{obs}}) \\ &= \int_T f(x_{\text{obs}} \mid \theta)\pi(\theta)\, d\theta \Big/ f(x_{\text{obs}}) = \int_T \pi(\theta \mid x_{\text{obs}})\, d\theta \end{aligned}$$

in the case where θ **is continuous.**

If one wants to generate a draw from the posterior:

1. Draw θ_{prop} from the prior $\pi(\theta)$.
2. Draw x_{prop} from the density $f(x \mid \theta_{\text{prop}})$.
3. Accept θ_{prop} if $x_{\text{prop}} = x_{\text{obs}}$.

Why? Let θ_{acc} denote an accepted θ_{prop}. Then, for any $T \subseteq \mathbb{R}$,

$$\begin{aligned} P(\theta_{\text{acc}} \in T) &= P(\theta_{\text{prop}} \in T \mid x_{\text{prop}} = x_{\text{obs}}) \\ &= K \int_T P(x_{\text{prop}} = x_{\text{obs}} \mid \theta_{\text{prop}} = \theta)\pi(\theta)\, d\theta \\ &= K \int_T f(x_{\text{obs}} \mid \theta)\pi(\theta)\, d\theta = \int_T \pi(\theta \mid x_{\text{obs}})\, d\theta \end{aligned}$$

in the case where θ **is continuous.**

If one wants to generate a draw from the posterior:

1. Draw θ_{prop} from the prior $\pi(\theta)$.
2. Draw x_{prop} from the density $f(x \mid \theta_{\text{prop}})$.
3. Accept θ_{prop} if **???**

Why? Let θ_{acc} denote an accepted θ_{prop}. Then, for any $T \subseteq \mathbb{R}$,

$$\begin{aligned} P(\theta_{\text{acc}} \in T) &= P(\theta_{\text{prop}} \in T \mid \text{Accept } \theta_{\text{prop}}) \\ &= K \int_T P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta)\pi(\theta)\, d\theta \\ &\stackrel{?}{=} K' \int_T f(x_{\text{obs}} \mid \theta)\pi(\theta)\, d\theta = \int_T \pi(\theta \mid x_{\text{obs}})\, d\theta \end{aligned}$$

in the case where θ **is continuous.**

The Point:

θ_{acc} is a draw from the posterior if

$$P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta) \propto f(x_{\text{obs}}|\theta) \quad \text{(the likelihood)}$$

This creates a basis for assessing the quality of the approximation, irrespective of the prior.

To achieve this, we could accept θ_{prop} if $x_{\text{prop}} = x_{\text{obs}}$.
Of course, this is not practical.

Clear alternative is to accept θ_{prop} if x_{prop} is "close to" x_{obs} using some chosen distance metric Δ.

Clear alternative is to accept θ_{prop} if x_{prop} is "close to" x_{obs} using some chosen distance metric Δ.

What is the price of this approximation?

Clear alternative is to accept θ_{prop} if x_{prop} is "close to" x_{obs} using some chosen distance metric Δ.

What is the price of this approximation?

Define:

$$\phi_\epsilon(x_{\text{prop}}, x_{\text{obs}}) = \begin{cases} 1, & \text{if } \Delta(x_{\text{prop}}, x_{\text{obs}}) < \epsilon \\ 0, & \text{if } \Delta(x_{\text{prop}}, x_{\text{obs}}) \geq \epsilon \end{cases}$$

In other words, $\phi_\epsilon(x_{\text{prop}}, x_{\text{obs}})$ is an indicator as to whether or not x_{prop} is close to x_{obs}.

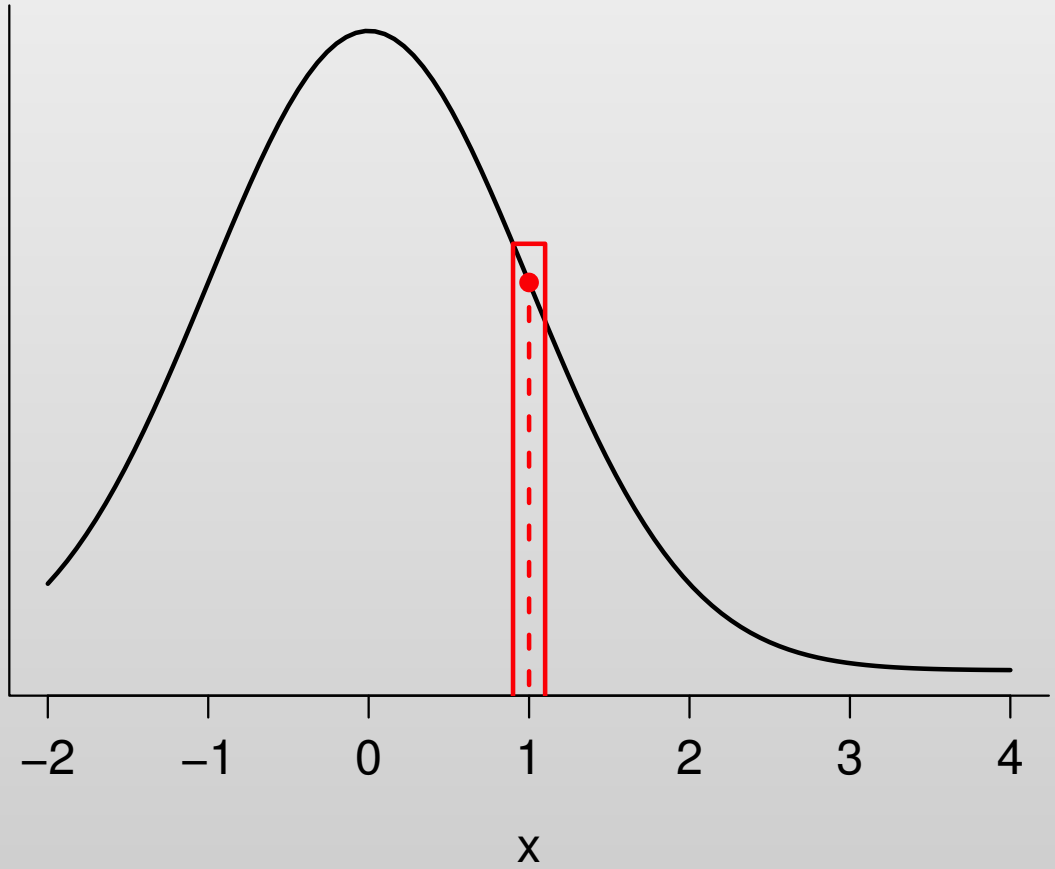
Hence,

$$\begin{aligned} P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta) &= P(\Delta(x_{\text{prop}}, x_{\text{obs}}) < \epsilon \mid \theta_{\text{prop}} = \theta) \\ &= \int \phi_\epsilon(x, x_{\text{obs}}) f(x \mid \theta)\, dx \\ &\longrightarrow Kf(x_{\text{obs}} \mid \theta) \quad \text{as } \epsilon \to 0 \end{aligned}$$

Hence, for ϵ small,

$$P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta) \approx Kf(x_{\text{obs}} \mid \theta)$$

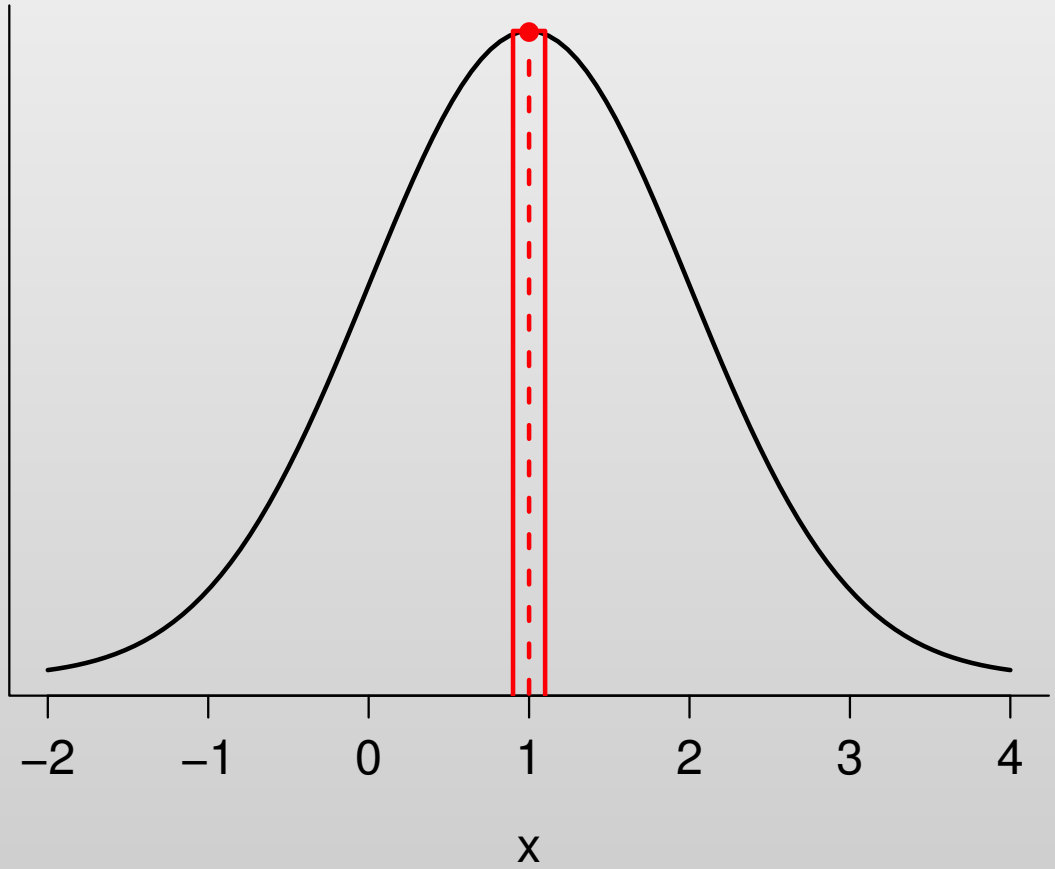
Toy Example: Assume we have a single observation, x_{obs}, from a Gaussian with mean θ and variance one.



Depicts the convolution

$$\int \phi_\epsilon(x, x_{\text{obs}}) f(x \mid \theta)\, dx = P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta)$$

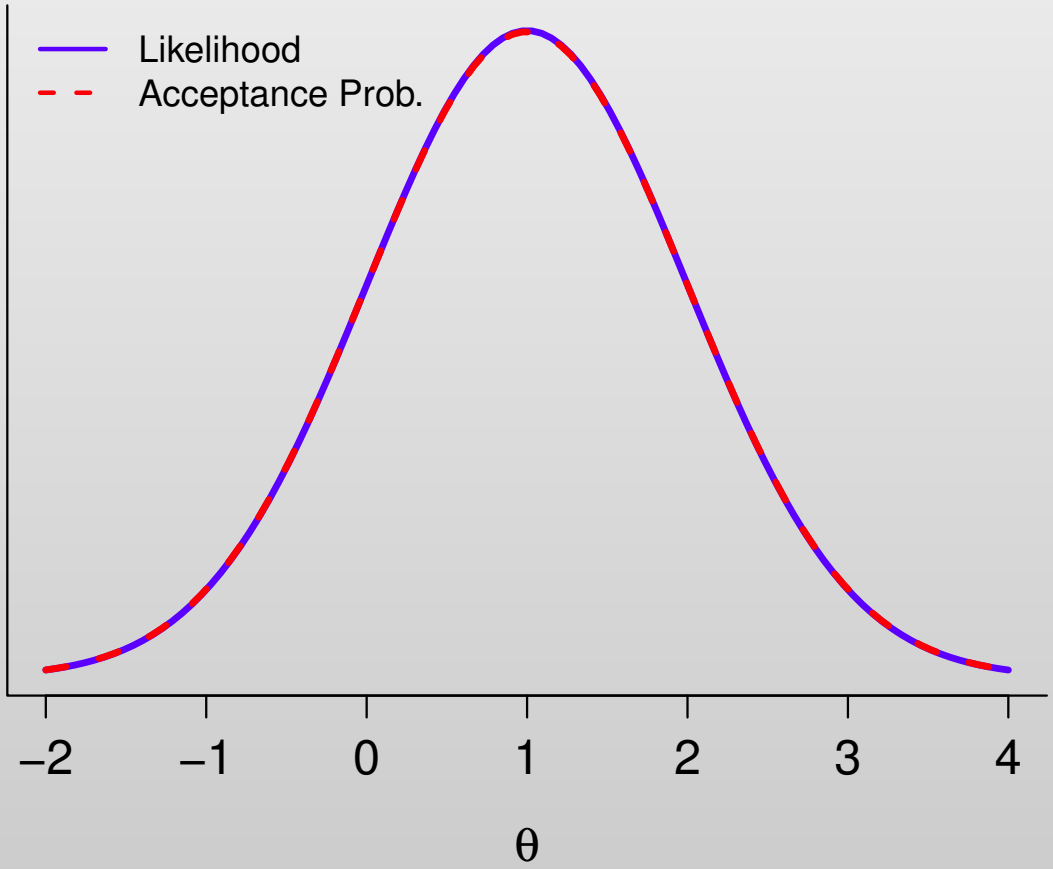
for case where $x_{\text{obs}} = 1$, $\theta = 0$, $\epsilon = 0.1$.



Depicts the convolution

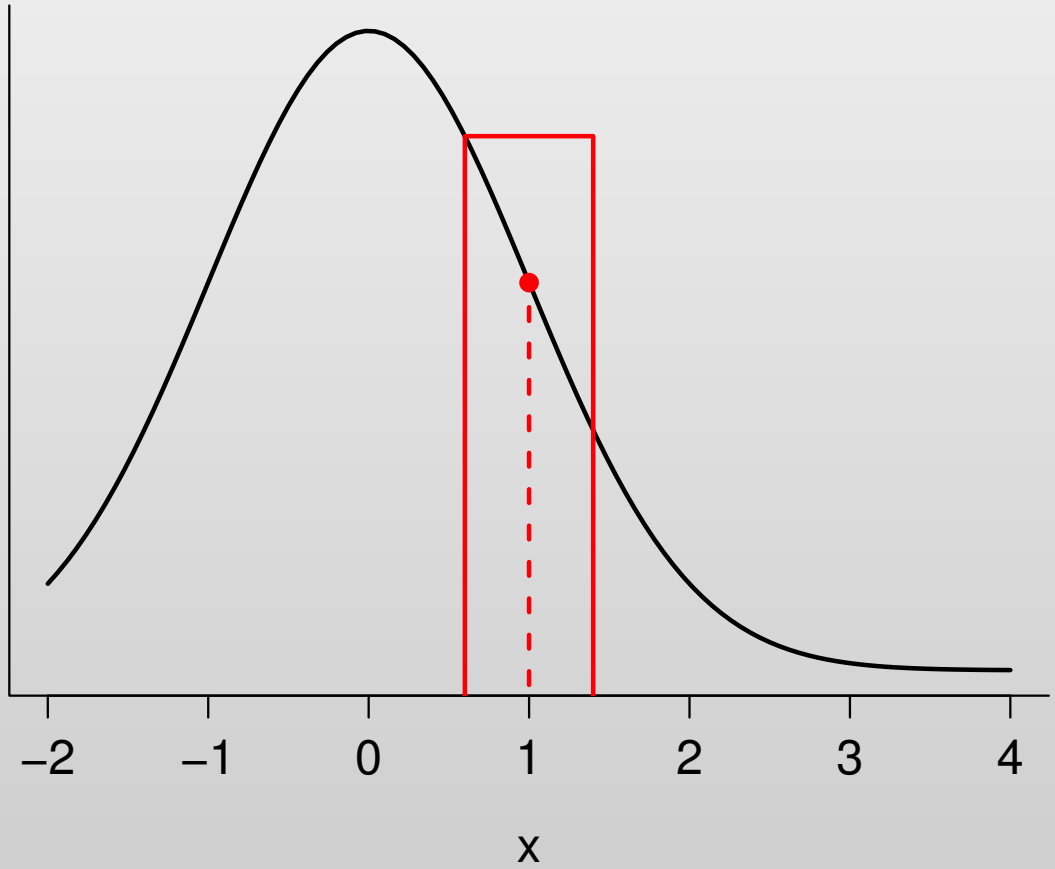
$$\int \phi_\epsilon(x, x_{\text{obs}}) f(x \mid \theta)\, dx = P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta)$$

for case where $x_{\text{obs}} = 1$, $\theta = 1$, $\epsilon = 0.1$.



Compare these quantities for all θ. Case where $x_{\text{obs}} = 1$, $\epsilon = 0.1$.

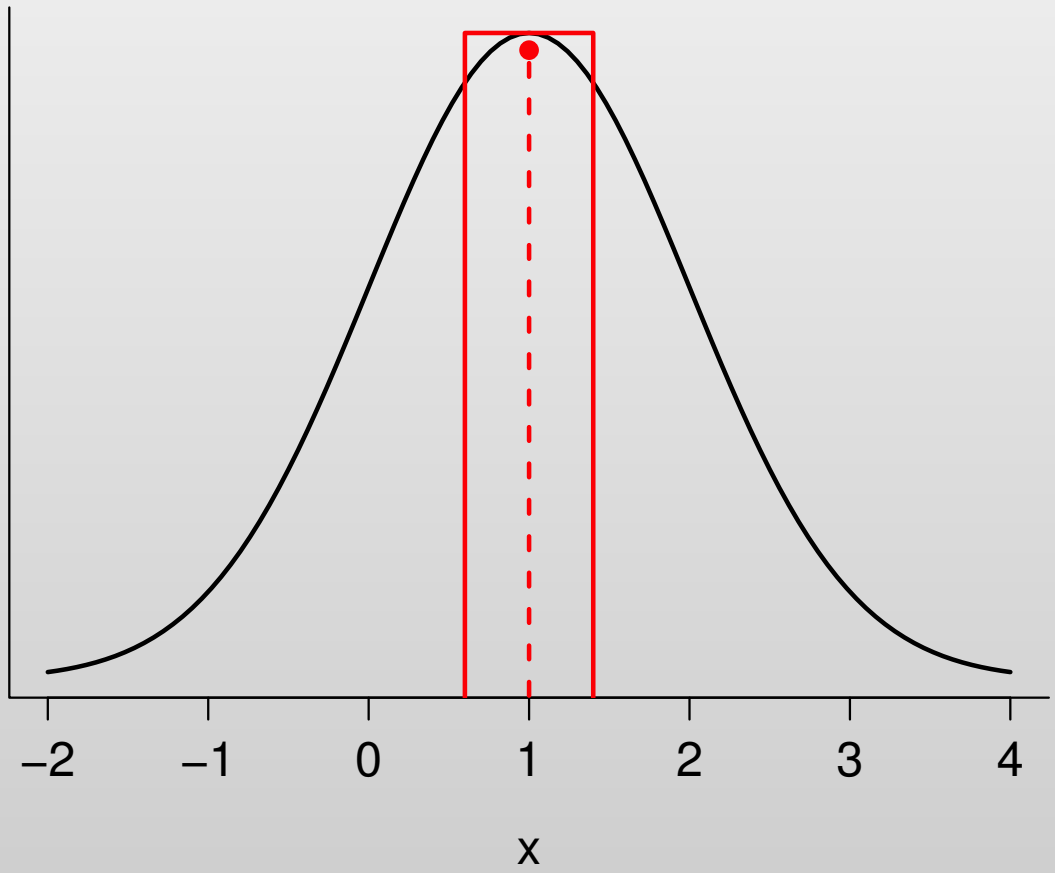
Note: Acceptance probability curve has been normalized so the area under the curve is 1.



Depicts the convolution

$$\int \phi_\epsilon(x, x_{\text{obs}}) f(x \mid \theta)\, dx = P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta)$$

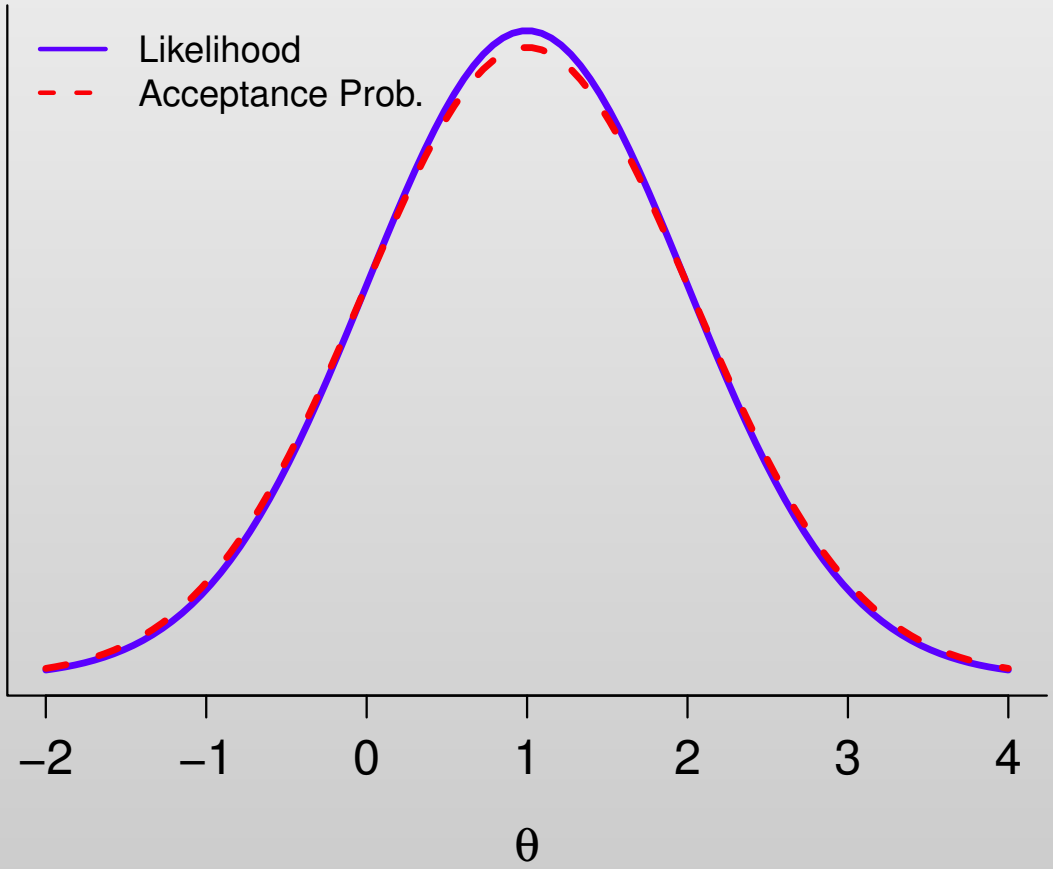
for case where $x_{\text{obs}} = 1$, $\theta = 0$, $\epsilon = 0.4$.



Depicts the convolution

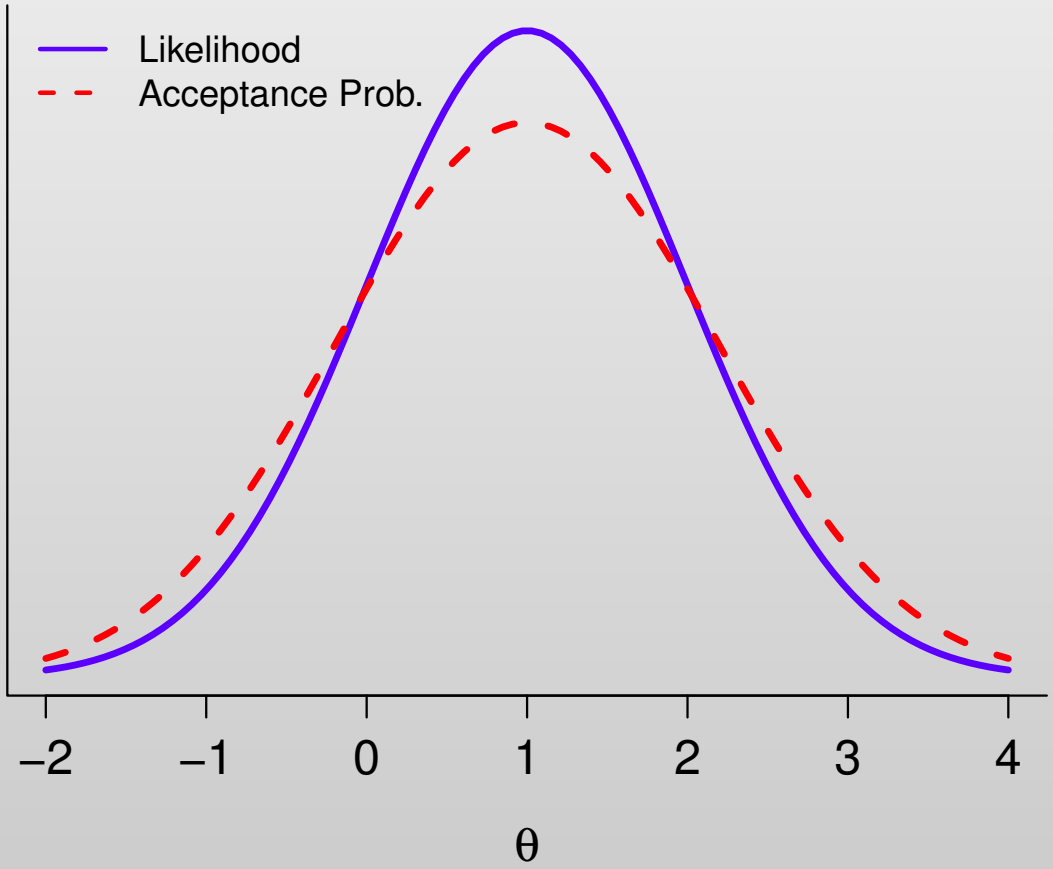
$$\int \phi_\epsilon(x, x_{\text{obs}}) f(x \mid \theta)\, dx = P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta)$$

for case where $x_{\text{obs}} = 1$, $\theta = 1$, $\epsilon = 0.4$.



Compare these quantities for all θ. Case where $x_{\text{obs}} = 1$, $\epsilon = 0.4$.

Note: Acceptance probability curve has been normalized so the area under the curve is 1.



Case where $x_{\text{obs}} = 1$, $\epsilon = 1$.

Note: Acceptance probability curve has been normalized so the area under the curve is 1.

Is comparing x_{prop} with x_{obs} realistic?

Is comparing x_{prop} with x_{obs} realistic?

No. When x is high-dimensional, ϵ will need to be too large in order to keep the acceptance probability reasonable.

Instead, need to compare summaries, $S(x_{\text{prop}})$ and $S(x_{\text{obs}})$.

Gaussian illustration

Mean of a Gaussian with known variance

Consider the following model:

$$\mu \sim N(\mu_0, \sigma_0^2)$$
$$Y_i \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

The (true) **posterior** is

$$\pi(\mu \mid y_{1:n}) \sim N(\mu_1, \sigma_1^2)$$

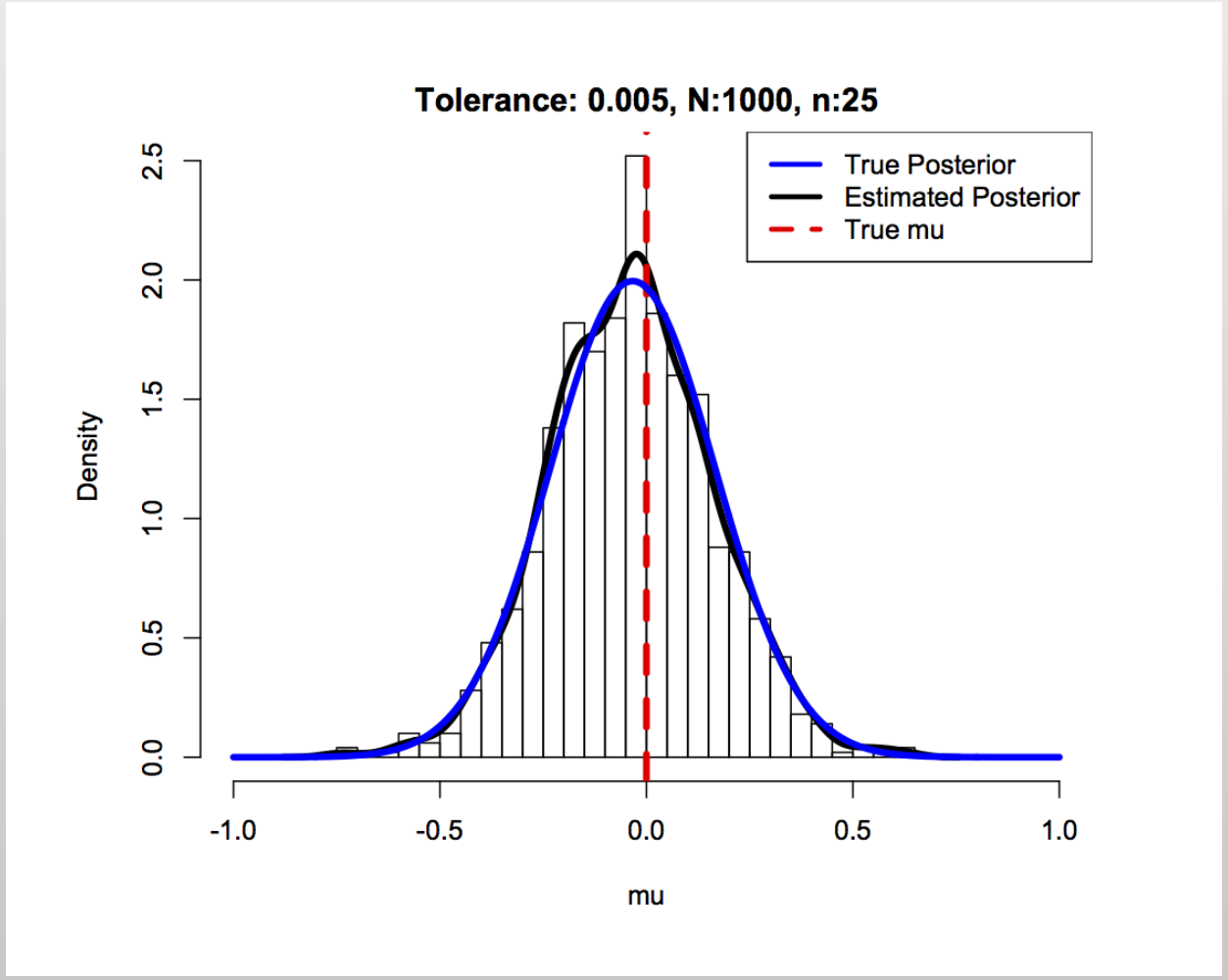
where

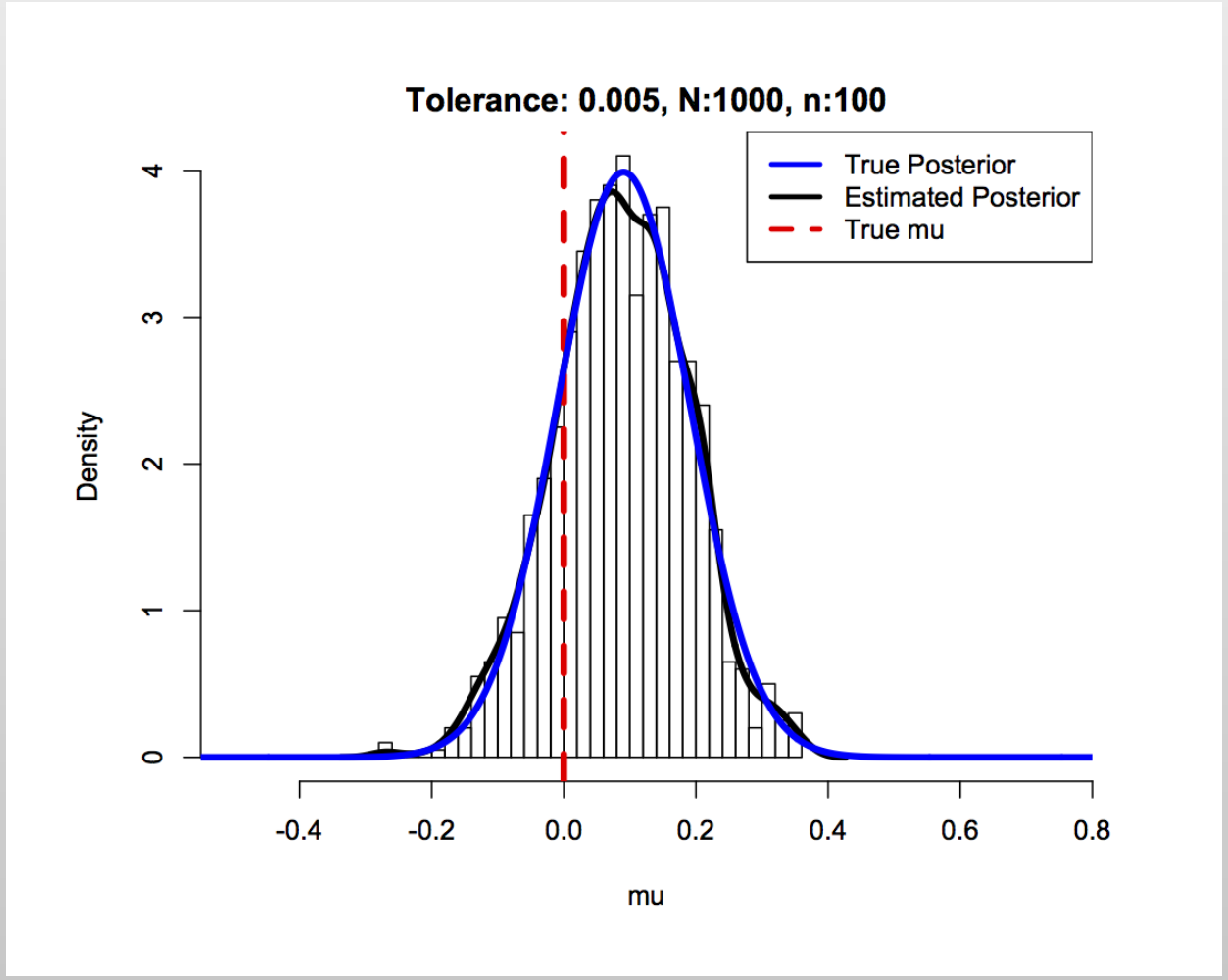
$$\mu_1 = \frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum y_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}, \qquad \sigma_1^2 = \frac{1}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}$$

Gaussian illustration: R code

```
n <- 25          #number of observations
N <-1000         #particle sample size
true.mu <- 0; sigma <- 1
mu.hyper <- 0; sigma.hyper <- 10
data <- rnorm(n,true.mu,sigma)
epsilon <- 0.005
mu <- numeric(N)
rho <- function(y,x) abs(sum(y)-sum(x))/n

for(i in 1:N){
d <- epsilon+1
   while(d>epsilon) {
       proposed.mu <- rnorm(1,mu.hyper,sigma.hyper) #<--prior draw
       x <- rnorm(n, proposed.mu, sigma)
       d <- rho(data,x)
       }
mu[i] <- proposed.mu
}
```


Tolerance: 0.005, N:1000, n:25
True Posterior
Estimated Posterior
True mu
Density
0.0
0.5
1.0
1.5
2.0
2.5
-1.0
-0.5
0.0
0.5
1.0
mu


Tolerance: 0.005, N:1000, n:100
True Posterior
Estimated Posterior
True mu
Density
0
1
2
3
4
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
mu

ABC in a nutshell

"The basic idea behind ABC is that using a representative (enough) summary statistic η coupled with a small (enough) tolerance ϵ should produce a good (enough) approximation to the posterior..."

Marin et al. (2012)

Summary of basic ABC

- Decisions that need to be made:
  1. Select distance function (ρ) and summary statistic(s)
  2. Tolerance (ϵ)
- Finding the "right" ϵ can be inefficient
  $\longrightarrow$ we end up throwing away many of the theories proposed from the selected priors
- How can we improve this basic algorithm?

Sequential ABC

Main idea

Instead of starting the ABC algorithm over with a smaller tolerance (ϵ), use the already sampled particle system as a proposal distribution *rather* than drawing from the prior distribution.

Particle system:

(1) retained sampled values, (2) importance weights

Some references:

Beaumont et al. (2009); Moral et al. (2011); Bonassi and West (2004)

Algorithm 1 ABC - Population Monte Carlo algorithm*

At iteration $t = 1$
Basic ABC sampler to obtain $\{\theta_1^{(i)}\}_{i=1}^N$
Set importance weights $W_1^{(i)} = 1/N$ for $i = 1, \ldots, N$
for $t = 2$ to T **do**
  Set $\tau_t^2 = 2 \cdot \text{var}\left(\{\theta_{t-1}^{(i)}, W_{t-1}^{(i)}\}_{i=1}^N\right)$
  for $i = 1$ to N **do**
    while $\rho\left(S(x_{\text{obs}}), S(x_{\text{prop}})\right) > \epsilon_t$ **do**
      Draw θ_0 from $\{\theta_{t-1}^{(i)}\}_{i=1}^N$ with probabilities $\{W_{t-1}^{(i)}\}_{i=1}^N$
      Propose $\theta_{\text{prop}} \sim N(\theta_0, \tau_t^2)$
      Generate x_{prop} from $F(x \mid \theta_{\text{prop}})$
      Calculate summary statistics $\{S(x_{\text{obs}}), S(x_{\text{prop}})\}$
    end while
    $\theta_t^{(i)} \leftarrow \theta_{\text{prop}}$
    $$\widetilde{W}_t^{(i)} \leftarrow \frac{\pi\left(\theta_t^{(i)}\right)}{\sum_{j=1}^N W_{t-1}^{(j)} \phi\left[\tau_t^{-1}(\theta_t^{(i)} - \theta_{t-1}^{(j)})\right]}$$
  end for
  $\{W_t^{(i)}\}_{i=1}^N \leftarrow \{\widetilde{W}_t^{(i)}\}_{i=1}^N / \sum_{i=1}^N \widetilde{W}_t^{(i)}$
end for

Decreasing tolerances $\epsilon_1 \geq \cdots \geq \epsilon_T$, $\phi(\cdot)$ is the density function of a $N(0, 1)$

*From Beaumont et al. (2009)

Recall: Mean of a Gaussian with known σ^2

Given the following model:

$$\mu \sim N(\mu_0, \sigma_0^2)$$
$$Y_i \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

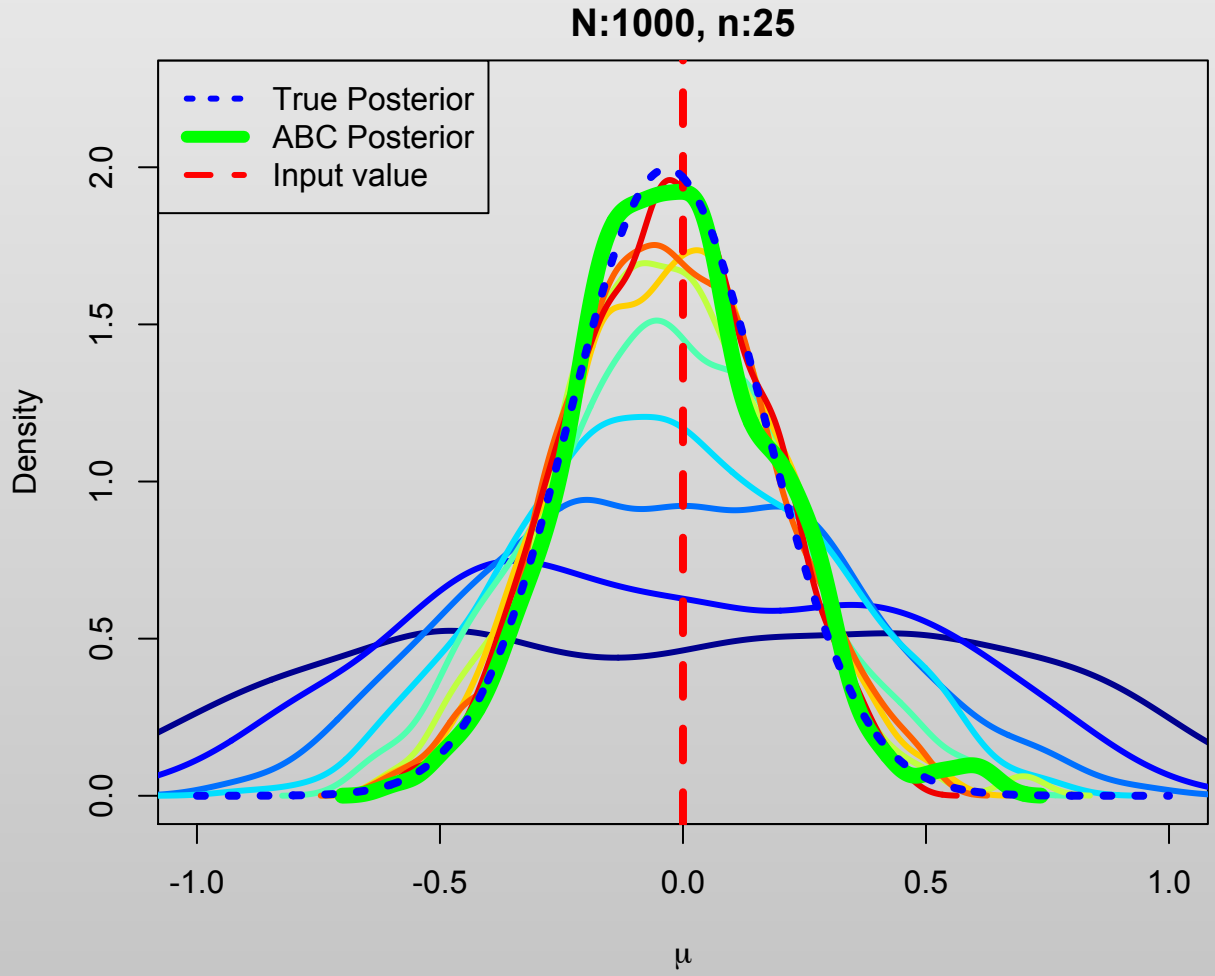
The **posterior** is

$$\pi(\mu \mid y_{1:n}) \sim N(\mu_1, \sigma_1^2)$$

where

$$\mu_1 = \frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum y_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}, \qquad \sigma_1^2 = \frac{1}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}$$

Gaussian illustration: sequential posteriors



Tolerance sequence, $\epsilon_{1:10}$:

1.00 0.75 0.53 0.38 0.27 0.19 0.15 0.11 0.08 0.06

Gaussian illustration: Sequential R code

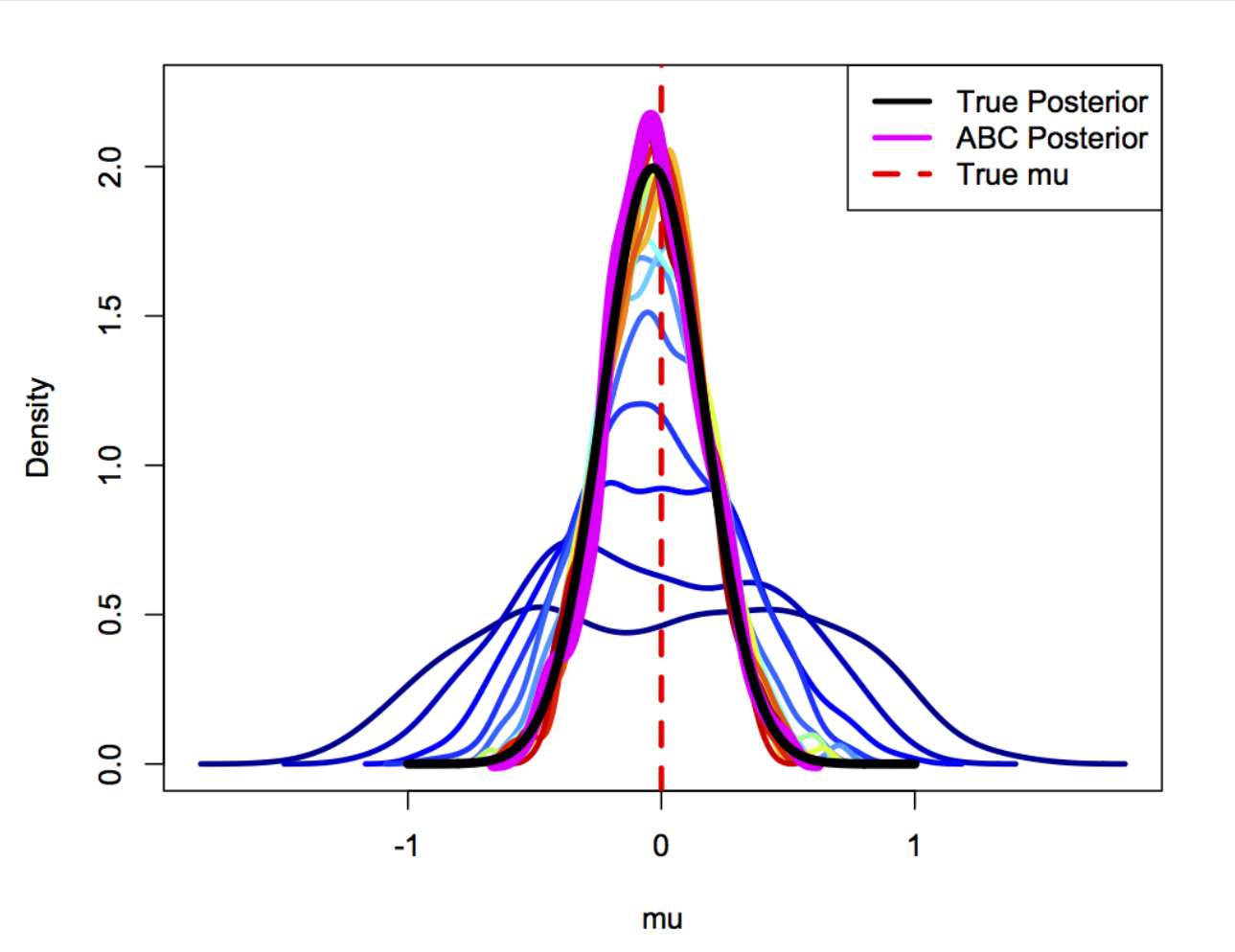
```
# INPUTS
n <- 25  #number of observations
N <- 2500  #particle sample size
true.mu <- 0
sigma <- 1
mu.hyper <- 0
sigma.hyper <- 10
data <- rnorm(n,true.mu,sigma)
epsilon <- 1
time.steps <- 20
weights <- matrix(1/N,time.steps,N)
mu <- matrix(NA,time.steps,N)
d <- matrix(NA,time.steps,N)
rho <- function(y,x) abs(sum(y)-sum(x))/n
```

Mean of a Gaussian with σ^2 known: Sequential R code

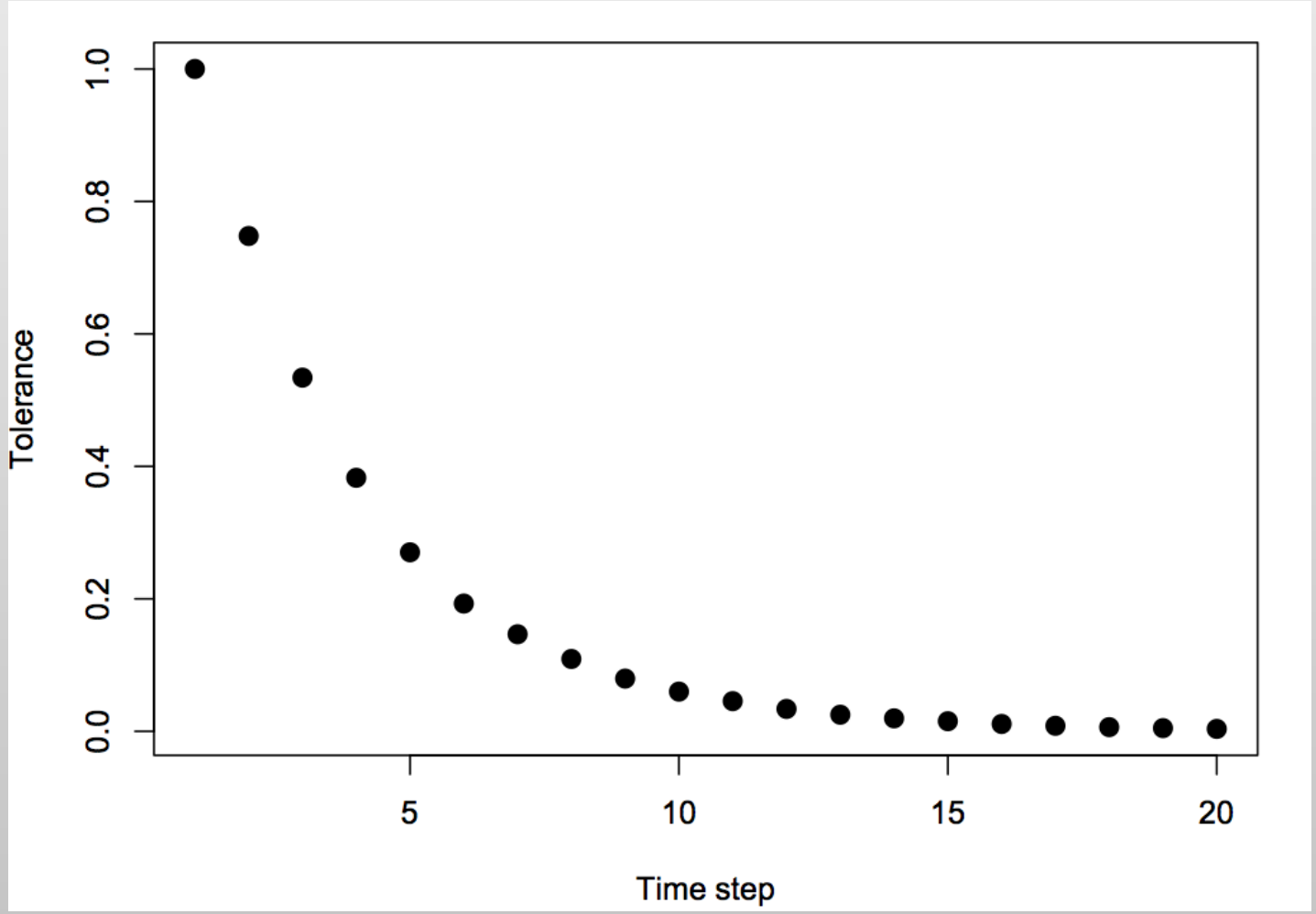
```
for(t in 1:time.steps){
   if(t==1){
      for(i in 1:N){
         d[t,i] <- epsilon +1
         while(d[t,i]>epsilon) {
            proposed.mu <- rnorm(1,0,sigma.hyper) #<--prior draw
            x <- rnorm(n, proposed.mu, sigma)
            d[t,i] <- rho(data,x)}
            mu[t,i] <- proposed.mu
   }} else{[NEXT SLIDE]}
}
```

```
for(t in 1:time.steps){ if(t==1){[PREVIOUS SLIDE]} else{
   epsilon <- c(epsilon,quantile(d[t-1,],.75))
   mean.prev <- sum(mu[t-1,]*weights[t-1,])
   var.prev <- sum((mu[t-1,] - mean.prev)^2*weights[t-1,])
   for(i in 1:N){d[t,i] <- epsilon[t]+1
      while(d[t,i]>epsilon[t]) {
      sample.particle <- sample(N, 1, prob = weights[t-1,])
      proposed.mu0 <- mu[t-1, sample.particle]
      proposed.mu <- rnorm(1, proposed.mu0, sqrt(2*var.prev))
      x <- matrix(rnorm(n,proposed.mu, sigma),n,1)
      d[t,i] <- rho(data,x) }
   mu[t,i] <- proposed.mu
   mu.weights.denominator<-
        sum(weights[t-1,]*dnorm(proposed.mu,mu[t-1,],sqrt(2*var.prev)))
   mu.weights.numerator<-dnorm(proposed.mu,0,sigma.hyper)
   weights[t,i] <- mu.weights.numerator/mu.weights.denominator
   }}
weights[t,] <- weights[t,]/sum(weights[t,])}
```

$N = 1000,\ n = 25$


True Posterior
ABC Posterior
True mu
Density
0.0
0.5
1.0
1.5
2.0
-1
0
1
mu

$N = 1000,\ n = 25$



Sequential setting: decisions

1. Determining the sequence of tolerances, $\epsilon_{1:t}$

   One possibilitiy use a quantile (e.g. 50th percentile) of the distribution of accepted distances from the previous time step

2. Moving the particles between time steps

   Need to ensure any constraints on the parameter space are satisfied

3. Calculating the particle weights

   Relies on ideas from Importance Sampling (next)

Before we get into importance sampling, let's recall **Monte Carlo integration**...

$$I = \int_a^b h(y)dy$$

- Goal: evaluate this integral
- Sometimes we can't directly calculate I and need a way to approximate it. Monte Carlo is one approach for doing this.

General idea

Monte Carlo methods are a form of stochastic integration used to approximate expectations by invoking the law of large numbers.

$$I = \int_a^b h(y)dy = \int_a^b w(y)f(y)dy = E_f(w(Y))$$

where $f(y) = \frac{1}{b-a}$ and $w(y) = h(y) \cdot (b-a)$

- $f(y) = \frac{1}{b-a}$ is the pdf of a Uniform(a,b) random variable
- By the LLN, if we take an iid sample of size N from Uniform(a,b), we can estimate I as

$$\hat{I} = N^{-1} \sum_{i=1}^{N} w(Y_i) \longrightarrow E(w(Y)) = I$$

Monte Carlo Integration: Gaussian CDF example

- Goal: estimate $F_Y(y) = P(Y \leq y) = E\left[I_{(-\infty,y)}(Y)\right]$ where $Y \sim N(0,1)$:

$$F(Y \leq y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \int_{-\infty}^{\infty} h(t) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

where $h(t) = 1$ if $t < y$ and $h(t) = 0$ if $t \geq y$

Monte Carlo Integration: Gaussian CDF example

- Goal: estimate $F_Y(y) = P(Y \leq y) = E\left[I_{(-\infty,y)}(Y)\right]$ where $Y \sim N(0,1)$:

$$F(Y \leq y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \int_{-\infty}^{\infty} h(t) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

  where $h(t) = 1$ if $t < y$ and $h(t) = 0$ if $t \geq y$

- Draw an iid sample $Y_1, \ldots, Y_N$ from a $N(0,1)$, then the estimator is

$$\hat{I} = N^{-1} \sum_{i=1}^{N} h(Y_i) = \frac{\#\text{ draws } < x}{N}$$

★ Example 24.2 of Wasserman (2004)

Importance Sampling: motivation

- Standard Monte Carlo integration is great if you can sample from the *target* distribution (i.e. the desired distribution)

  $\longrightarrow$ But what if you can't sample from the target?

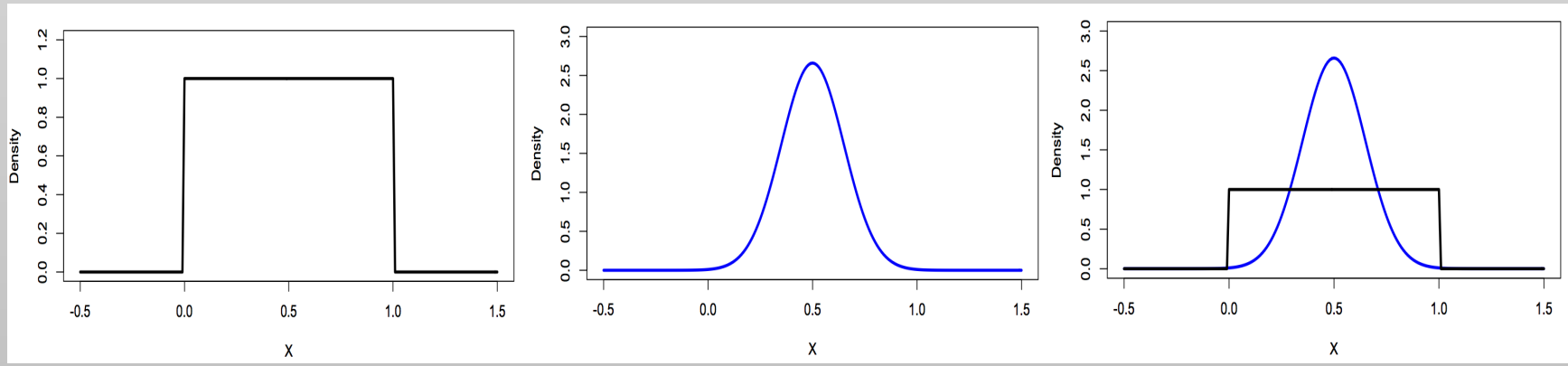
Importance Sampling: motivation

- Standard Monte Carlo integration is great if you can sample from the *target* distribution (i.e. the desired distribution)

  $\longrightarrow$ But what if you can't sample from the target?

- Idea of importance sampling: draw the sample from a *proposal* distribution and re-weight the integral using *importance weights* so that the correct distribution is targeted

Target Proposal IS



Monte Carlo Integration $\longrightarrow$ Importance Sampling

$$I = \int h(y) f(y) dy$$

- h is some function and f is the probability density function of Y
- When the density f is difficult to sample from, importance sampling can be used

Monte Carlo Integration $\longrightarrow$ Importance Sampling

$$I = \int h(y) f(y) dy$$

- h is some function and f is the probability density function of Y
- When the density f is difficult to sample from, importance sampling can be used
- Rather than sampling from f, you specify a different probability density function, g, as the proposal distribution.

$$I = \int h(y) f(y) dy = \int h(y) \frac{f(y)}{g(y)} g(y) dy = \int \frac{h(y) f(y)}{g(y)} g(y) dy$$

Importance Sampling

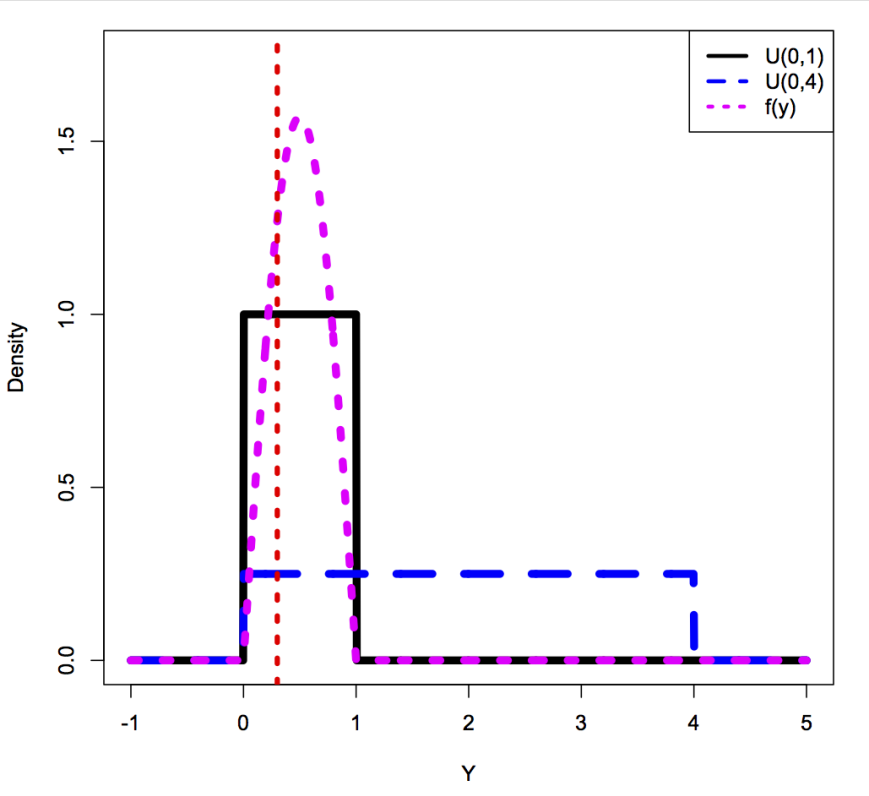
$$I = E_f\left[h(Y)\right] = \int \frac{h(y)f(y)}{g(y)} g(y)dy = E_g\left[\frac{h(Y)f(Y)}{g(Y)}\right]$$

Hence, given an iid sample $Y_1, \ldots, Y_N$ from g, our estimator of I becomes

$$\hat{I} = N^{-1} \sum_{i=1}^{N} \frac{h(Y_i)f(Y_i)}{g(Y_i)} \longrightarrow E_g\left[\frac{h(Y)f(Y)}{g(Y)}\right] = I$$

Importance sampling: Illustration

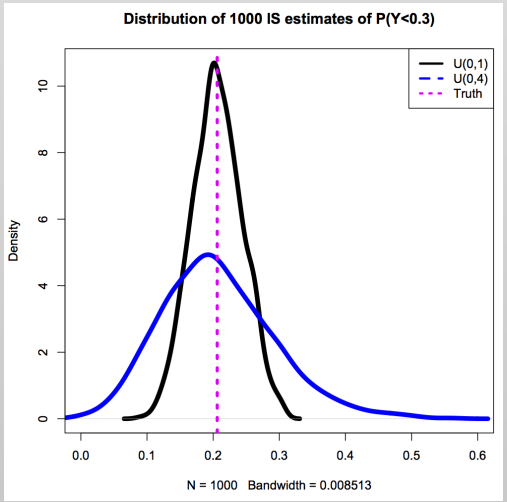
- Goal: estimate $P(Y < 0.3)$ where $Y \sim f$
- Try two proposal distributions: U(0,1) and U(0,4)



Importance sampling: Illustration, continued.

If we take 1000 samples of size 100, and find the IS estimates, we get the following *estimated* expected values and variances.

	Expected Value	Variance
Truth	0.206	0
g_1: U(0,1)	0.206	0.0014
g_2: U(0,4)	0.211	0.0075



Extensions of Importance Sampling

- Sequential Importance Sampling
- Sequential Monte Carlo (Particle Filtering)
  $\longrightarrow$ See Doucet et al. (2001)
- Approximate Bayesian Computation

ABC-PMC: importance weights

1: At iteration $t = 1$
2: Basic ABC sampler to obtain $\{\theta_1^{(i)}\}_{i=1}^N$
3: Set importance weights $W_1^{(i)} = 1/N$ for $i = 1, \ldots, N$
4: **for** $t = 2$ to T **do**
5: Set $\tau_t^2 = 2 \cdot \text{var}\left(\{\theta_{t-1}^{(i)}, W_{t-1}^{(i)}\}_{i=1}^N\right)$
6: **for** $i = 1$ to N **do**
7: **while** $\rho\left(S(x_{\text{obs}}), S(x_{\text{prop}})\right) > \epsilon_t$ **do**
8: Draw θ_0 from $\{\theta_{t-1}^{(i)}\}_{i=1}^N$ with probabilities $\{W_{t-1}^{(i)}\}_{i=1}^N$
9: Propose $\theta_{\text{prop}} \sim N(\theta_0, \tau_t^2)$
10: Generate x_{prop} from $F(x \mid \theta_{\text{prop}})$
11: Calculate summary statistics $\{S(x_{\text{obs}}), S(x_{\text{prop}})\}$
12: **end while**
13: $\theta_t^{(i)} \leftarrow \theta_{\text{prop}}$
14: $\widetilde{W}_t^{(i)} \leftarrow \dfrac{\pi\left(\theta_t^{(i)}\right)}{\sum_{j=1}^N W_{t-1}^{(j)} \phi\left[\tau_t^{-1}(\theta_t^{(i)} - \theta_{t-1}^{(j)})\right]}$
15: **end for**
16: $\{W_t^{(i)}\}_{i=1}^N \leftarrow \{\widetilde{W}_t^{(i)}\}_{i=1}^N / \sum_{i=1}^N \widetilde{W}_t^{(i)}$
17: **end for**

The importance weights for time step t and particle i in the ABC-PMC algorithm are defined as

$$\widetilde{W}_t^{(i)} = \frac{\pi\left(\theta_t^{(i)}\right)}{\sum_{j=1}^{N} W_{t-1}^{(j)} \phi\left[\tau_t^{-1}(\theta_t^{(i)} - \theta_{t-1}^{(j)})\right]}$$

Recall the following model:

$$\mu \sim N(\mu_0, \sigma_0^2), \qquad Y_i \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

```
mu.weights.denominator <-
sum(weights[t-1,]*dnorm(proposed.mu,mu[t-1,],sqrt(2*var.prev)))

mu.weights.numerator <- dnorm(proposed.mu, mu.hyper, sigma.hyper)

weights[t,i] <- mu.weights.numerator/mu.weights.denominator
```

(after this is computed for all the particles, $i = 1, \ldots, N$, the weights are normalized to sum to 1)

- There are other variations of ABC that may prove useful in your setting (Marin et al., 2012)
- Beaumont et al. (2002) introduces a post-processing adjustment (using local regression) to the simulation output in order to use more of the simulated draws (with extensions in Blum and François (2010))

Concluding remarks

1. Approximate Bayesian Computation could be a useful tool in astronomy, but it must be handled with care
2. There are three main decisions that need to be made in the standard ABC algorithm: summary statistic, distance function, and tolerance
3. Considering a sequence of tolerances can lead to more efficient sampling, but results in more decisions: how to decrease the tolerance, when to stop the sampling, how to "move" or "mix" the particles between sampling steps

Additional resources

- Csilléry et al. (2010): Approximate Bayesian Computation (ABC) in practice
- Csillery et al. (2012): abc: an R package for approximate Bayesian computation (ABC)
- Jabot et al. (2013): EasyABC: performing efficient approximate Bayesian computation sampling schemes (R package)

Akeret, J., Refregier, A., Amara, A., Seehars, S., and Hasner, C. (2015), "Approximate Bayesian computation for forward modeling in cosmology," *Journal of Cosmology and Astroparticle Physics*, 2015, 043.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009), "Adaptive approximate Bayesian computation," *Biometrika*, 96, 983 – 990.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002), "Approximate Bayesian Computation in Population Genetics," *Genetics*, 162, 2025 – 2035.

Blum, M. G. B. and François, O. (2010), "Non-linear regression models for Approximate Bayesian Computation," *Statistics and Computing*, 20, 63 – 73.

Bonassi, F. V. and West, M. (2004), "Sequential Monte Carlo with Adaptive Weights for Approximate Bayesian Computation," *Bayesian Analysis*, 1, 1–19.

Cameron, E. and Pettitt, A. N. (2012), "Approximate Bayesian Computation for Astronomical Model Analysis: A Case Study in Galaxy Demographics and Morphological Transformation at High Redshift," *Monthly Notices of the Royal Astronomical Society*, 425, 44–65.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010), "Approximate Bayesian Computation (ABC) in practice," *Trends in ecology & evolution*, 25, 410 – 418.

Csillery, K., Francois, O., and Blum, M. G. B. (2012), "abc: an R package for approximate Bayesian computation (ABC)," *Methods in Ecology and Evolution*.

Doucet, A., De Freitas, N., and Gordon, N. (2001), *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, New York: Springer-Verlag.

Ishida, E. E. O., Vitenti, S. D. P., Penna-Lima, M., Cisewski, J., de Souza, R. S., Trindade, A. M. M., Cameron, E., and Busti, V. C. (2015), "cosmoabc: Likelihood-free inference via Population Monte Carlo Approximate Bayesian Computation," *Astronomy and Computing*, 13, 1 – 11.

Jabot, F., Faure, T., and Dumoullin, N. (2013), *EasyABC: performing efficient approximate Bayesian computation*.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012), "Approximate Bayesian computational methods," *Statistics and Computing*, 22, 1167 – 1180.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003), "Markov chain Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences*, 100, 15324 – 15328.

Moral, P. D., Doucet, A., and Jasra, A. (2011), "An adaptive sequential Monte Carlo method for approximate Bayesian computation," *Statistics and Computing*, 22, 1009–1020.

Pritchard, J. K., Seielstad, M. T., and Perez-Lezaun, A. (1999), "Population Growth of Human Y Chromosomes: A study of Y Chromosome Microsatellites," *Molecular Biology and Evolution*, 16, 1791 – 1798.

Schafer, C. M. and Freeman, P. E. (2012), *Statistical Challenges in Modern Astronomy V*, Springer, chap. 1, Lecture Notes in Statistics, pp. 3 – 19.

Tavaré, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997), "Inferring coalescence times from DNA sequence data," *Genetics*, 145, 505 – 518.

Turner, B. M. and Zandt, T. V. (2012), "A tutorial on approximate Bayesian computation," *Journal of Mathematical Psychology*, 56, 69 – 85.

Wasserman, L. (2004), *All of statistics: a concise course in statistical inference*, Springer.

Weyant, A., Schafer, C., and Wood-Vasey, W. M. (2013), "Likeihood-free cosmological inference with type Ia supernovae: approximate Bayesian computation for a complete treatment of uncertainty," *The Astrophysical Journal*, 764, 116.