ORIE6217/CS6384: Applied Bayesian Data Analysis for Research

Lecture 4: Sampling introduction

Nikhil Garg

Announcements

Lecture plan

How are Bayesian models fit? Part 1

- Central difficulty
- Naïve solution: random sampling
- Building toward Markov Chain Monte Carlo methods: rejection sampling

First: why is this important?

- Even if you're only ever using Stan, useful to understand model fitting diagnostics
- Some models (especially with discrete parameters) can't be fit in Stan

Bayesian learning

Goal:

$$P(\theta|X) = \frac{\overbrace{P(X|\theta)}^{\text{data generating process}}\ \overbrace{P(\theta)}^{\text{prior}}}{P(X)}$$

easy to calculate

– Pr of data??

$$P(X) = \int P(X|\theta)\,P(\theta)\,d\theta$$

hard to calculate!

integral over potentially high dimensional space

More generally, might care about functions of theta.

$$\mathbb{E}_{\theta \sim p(\theta|x)}[f(\theta)] = \int f(\theta) P(\theta|x)\, d\theta$$

Examples: $f(\theta) = \theta$ posterior mean

$f_p(\theta) = p$th percentile CI

$\Rightarrow (f_{2.5}(\theta), f_{97.5}(\theta))$

If we can sample from $P(\theta|x)$, then done!

Notation

"unnormalized posterior"

easy to calculate

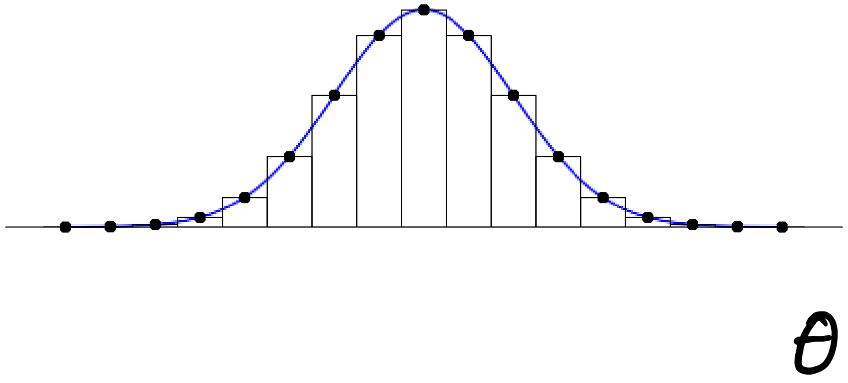
$$q(\theta|x) = p(x|\theta)\,p(\theta) \propto p(\theta|x)$$

$$\Rightarrow$$

$$\underset{\theta \sim p(\theta|x)}{E}\left[f(\theta)\right] = \int f(\theta) \underbrace{\left[\frac{q(\theta|x)}{\int q(\theta)\,d\theta}\right]}_{p(\theta|x)} d\theta$$

Idea 1: Grid sampling (uniform/equal spacing)

In high school calculus, learned grid approximations to integrals



Let G = grid over θ

e.g. np.linspace(0,1,1000)

$$E[f(\theta)] = \int f(\theta) \left[\frac{q(\theta|x)}{\int q(\theta|x)\,d\theta} \right] d\theta \approx \frac{\sum_{\theta \in G} f(\theta)\, q(\theta|x)}{\sum_{\theta \in G} q(\theta|x)}$$

Grid sampling notes

Often want to calculate $q(\theta|x)$ in log space

$$q(\theta|x) = \prod_i q(\theta|x_i) \quad \text{if data is independent}$$

$$\log q(\theta|x) = \sum \log q(\theta|x_i)$$

How good is it? Good if:

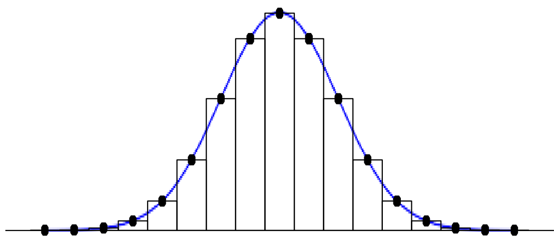
- Grid is "fine-grained" — many points
- Grid overlaps with $q(\theta|x)$

Really computationally expensive in high dimensions!! And how do we know where q is comparatively large?

Grid sampling and curse of dimensionality

- 10 parameters
- if we don't know beforehand where the posterior mass is
  - need to choose wide box for the grid
  - need to have enough grid points to get some of them where essential mass is
- e.g. 50 or 1000 grid points per dimension
  - $\rightarrow$ $50^{10} \approx$ 1e17 grid points
  - $\rightarrow$ $1000^{10} \approx$ 1e30 grid points
- R and my current laptop can compute density of normal distribution about 50 million times per second
  - $\rightarrow$ evaluation in 1e17 grid points would take 60 years
  - $\rightarrow$ evaluation in 1e30 grid points would take 600 billion years

Idea 2: We don't need uniform grid



We just need to *sample* where $q(\theta|x)$ is comparatively large

For any $g(\theta)$:

$$E_{\theta\sim p(\theta|x)}[f(\theta)] = \int f(\theta)\,\frac{1}{g(\theta)}\,\frac{q(\theta)}{\int q(\theta')\,d\theta'}\,g(\theta)\,d\theta$$

$$= \frac{E_{\theta\sim g(\theta)}\left[f(\theta)\,\frac{1}{g(\theta)}\,q(\theta)\right]}{E_{\theta\sim g(\theta)}\left[q(\theta)\,\frac{1}{g(\theta)}\right]}$$

"monte carlo"
random θ

$\Rightarrow$ Suppose G is a set of θ sampled w.p. $g(\theta)$

$$\mathbb{E}_{\theta \sim p(\theta|x)}\left[f(\theta) \right] \approx \frac{\sum_{\theta \in G} f(\theta) \frac{1}{g(\theta)} q(\theta)}{\sum_{\theta \in G} q(\theta)/g(\theta)}$$

Again, want $g(\theta)$ large where $q(\theta|x)$ comparatively large.

$$\boxed{E_{\theta \sim p(\theta|x)}[f(\theta)] = \int f(\theta) \frac{1}{g(\theta)} \frac{q(\theta)}{\int q(\theta')\, d\theta'} g(\theta)\, d\theta}$$

Ideally, $g(\theta) = p(\theta|x)$

why? do the math...

$$\Rightarrow E_{\theta \sim p(\theta|x)}[f(\theta)] \approx \frac{1}{|G|} \sum_{\theta \in G} f(\theta)$$

- how do we sample from $p(\theta|x)$?

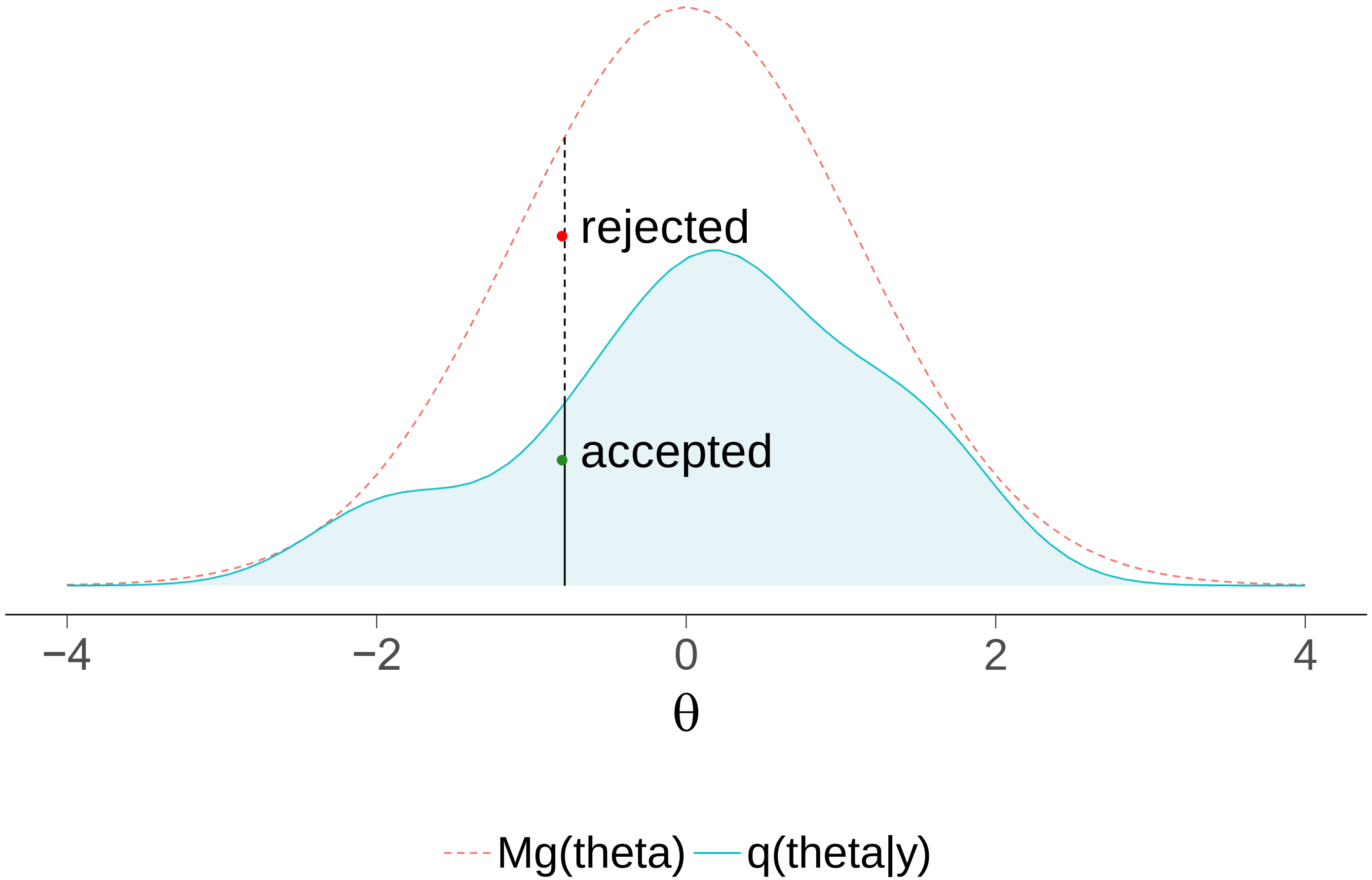
Indirect sampling

- Rejection sampling
- Importance sampling
- Markov chain Monte Carlo (next time)

Rejection sampling

Proposal forms envelope over the target distribution
$q(\theta|y)/Mg(\theta) \leq 1$

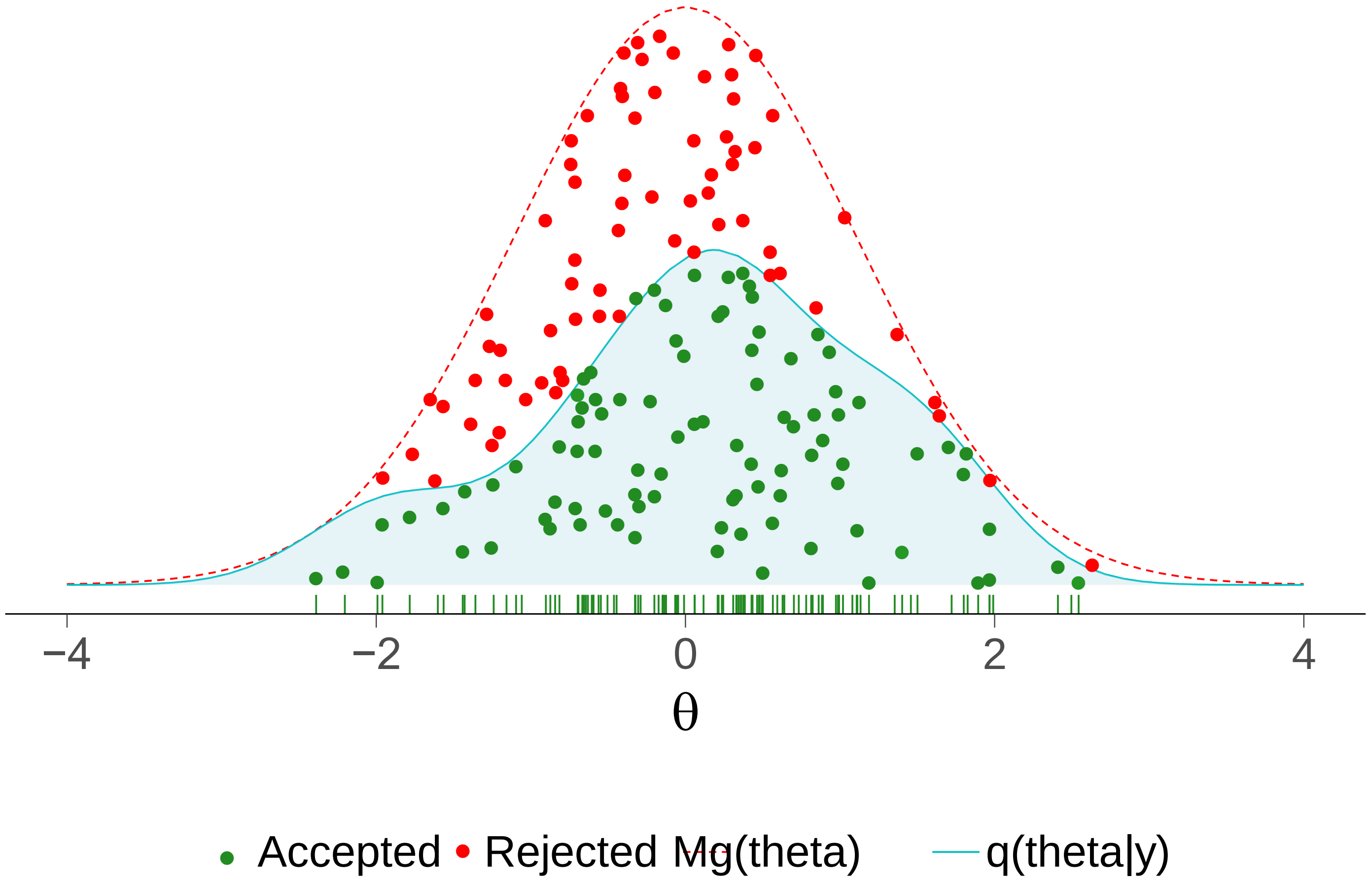
Draw from the proposal and accept with probability
$q(\theta|y)/Mg(\theta)$



Rejection sampling

Proposal forms envelope over the target distribution
$q(\theta|y)/Mg(\theta) \leq 1$

Draw from the proposal and accept with probability
$q(\theta|y)/Mg(\theta)$



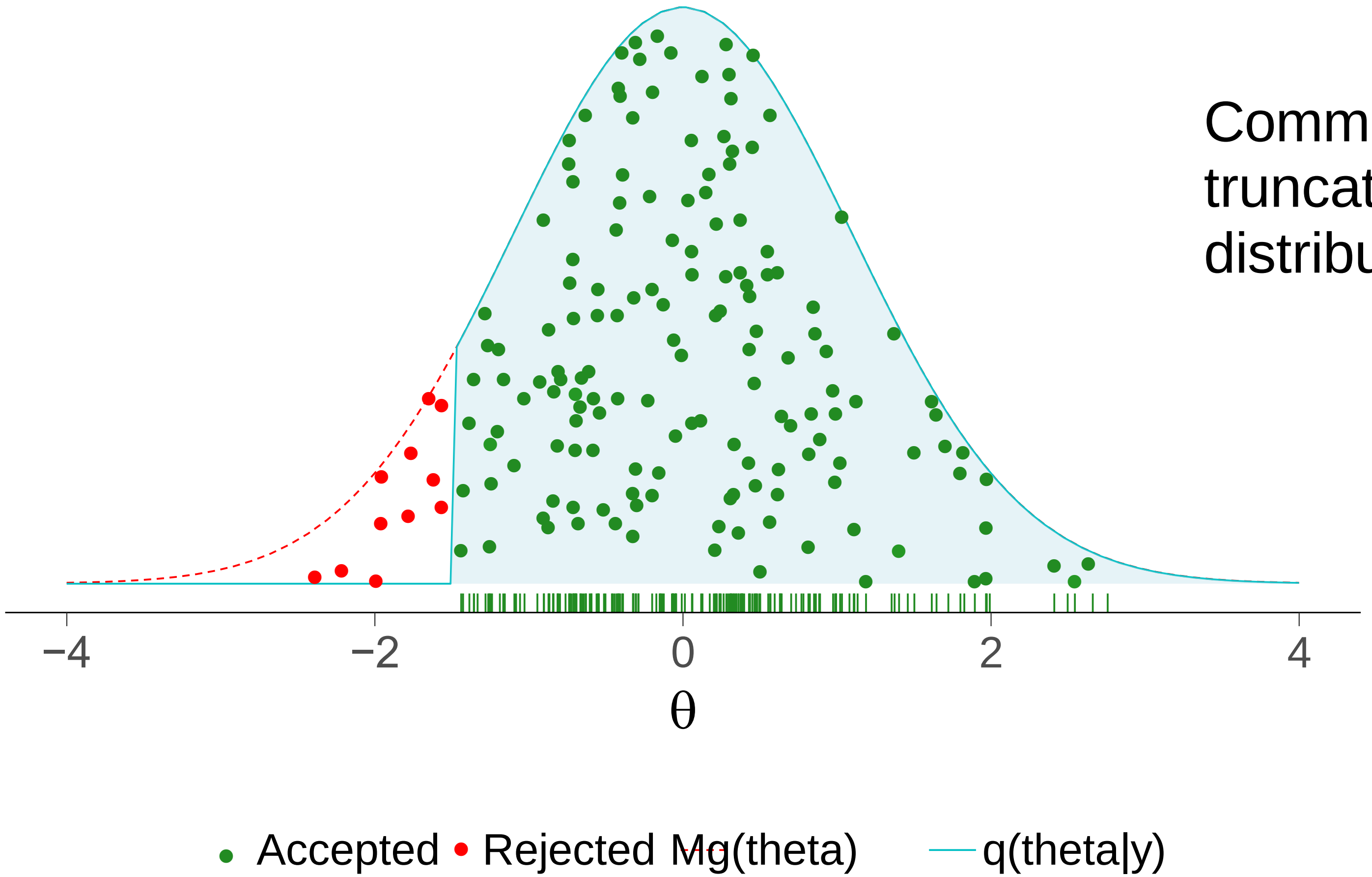
Rejection sampling

Proposal forms envelope over the target distribution

$q(\theta|y)/Mg(\theta) \leq 1$

Draw from the proposal and accept with probability

$q(\theta|y)/Mg(\theta)$



Common for truncated distributions

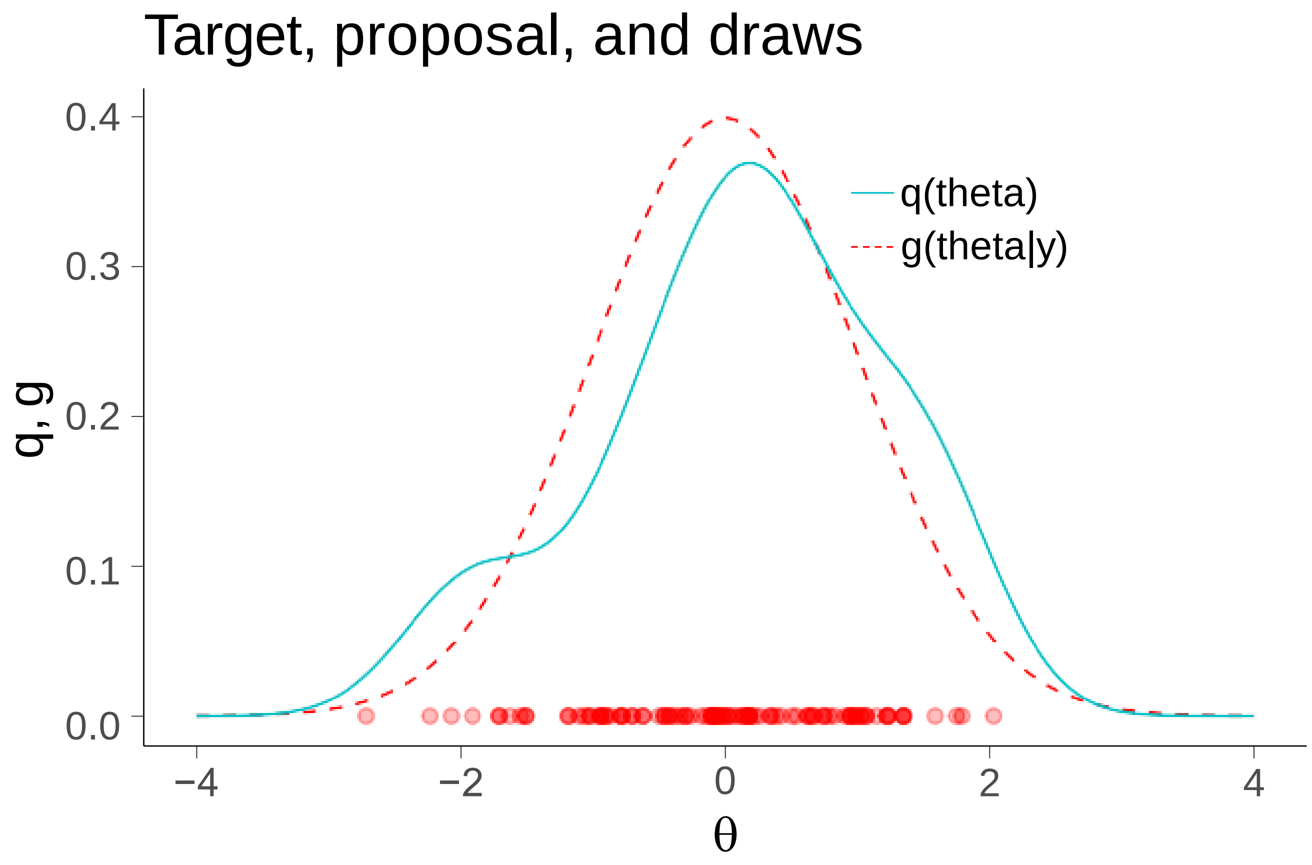
Rejection sampling

- The effective sample size (ESS) is the number of accepted draws
  - with bad proposal distribution may require a lot of trials
  - selection of good proposal gets very difficult when the number of dimensions increase
  - reliable diagnostics and thus can be a useful part

Code?

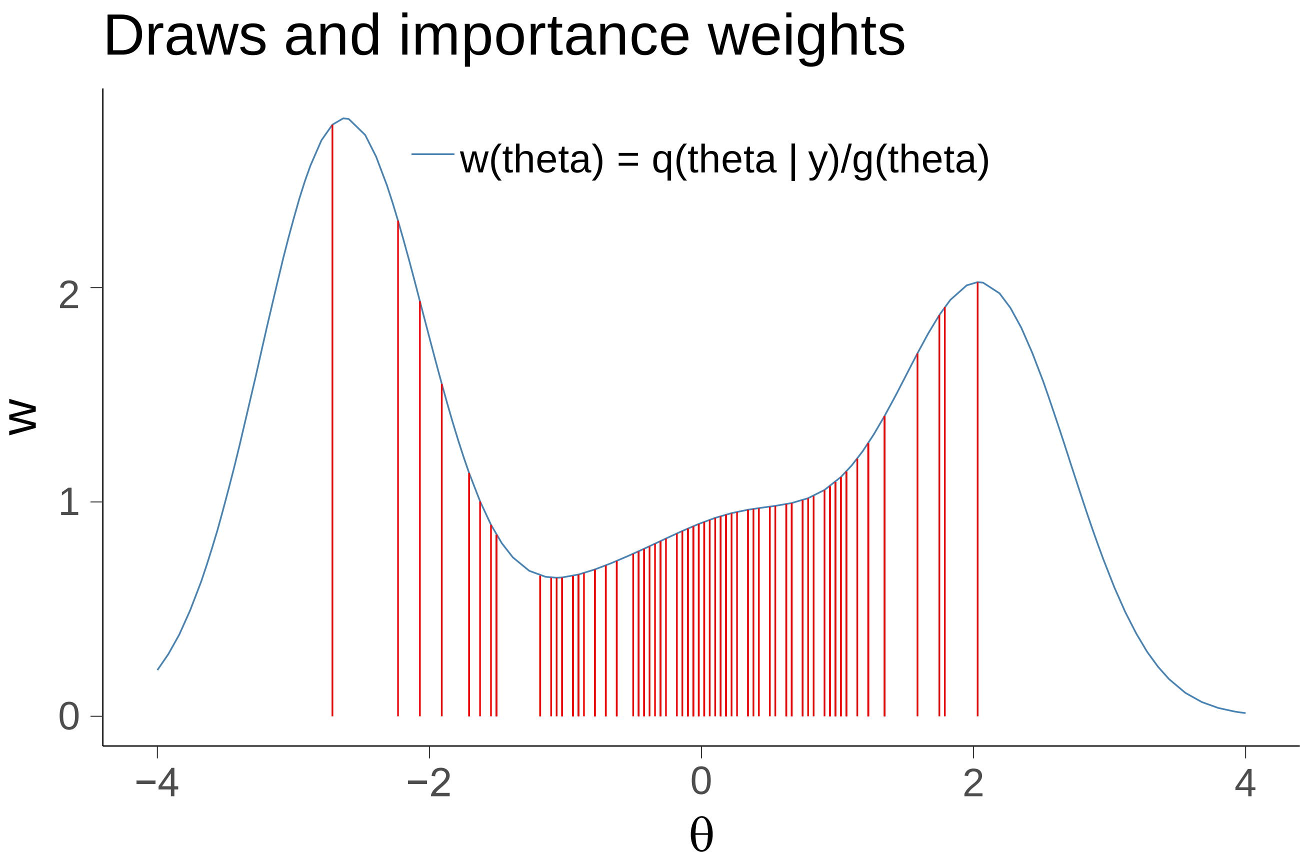
Importance sampling

-Proposal does not need to have a higher value everywhere



Importance sampling

-Proposal does not need to have a higher value everywhere



Some uses of importance sampling

In general selection of good proposal gets more difficult when the number of dimensions increase, but there are many special use case which scale well (e.g. Prof. Aki has used IS up to 10k dimensions)

- Fast leave-one-out cross-validation
- Fast bootstrapping
- Fast prior and likelihood sensitivity analysis
- Conformal Bayesian computation
- Particle filtering
- Improving distributional approximations (e.g Laplace, VI)

Questions?