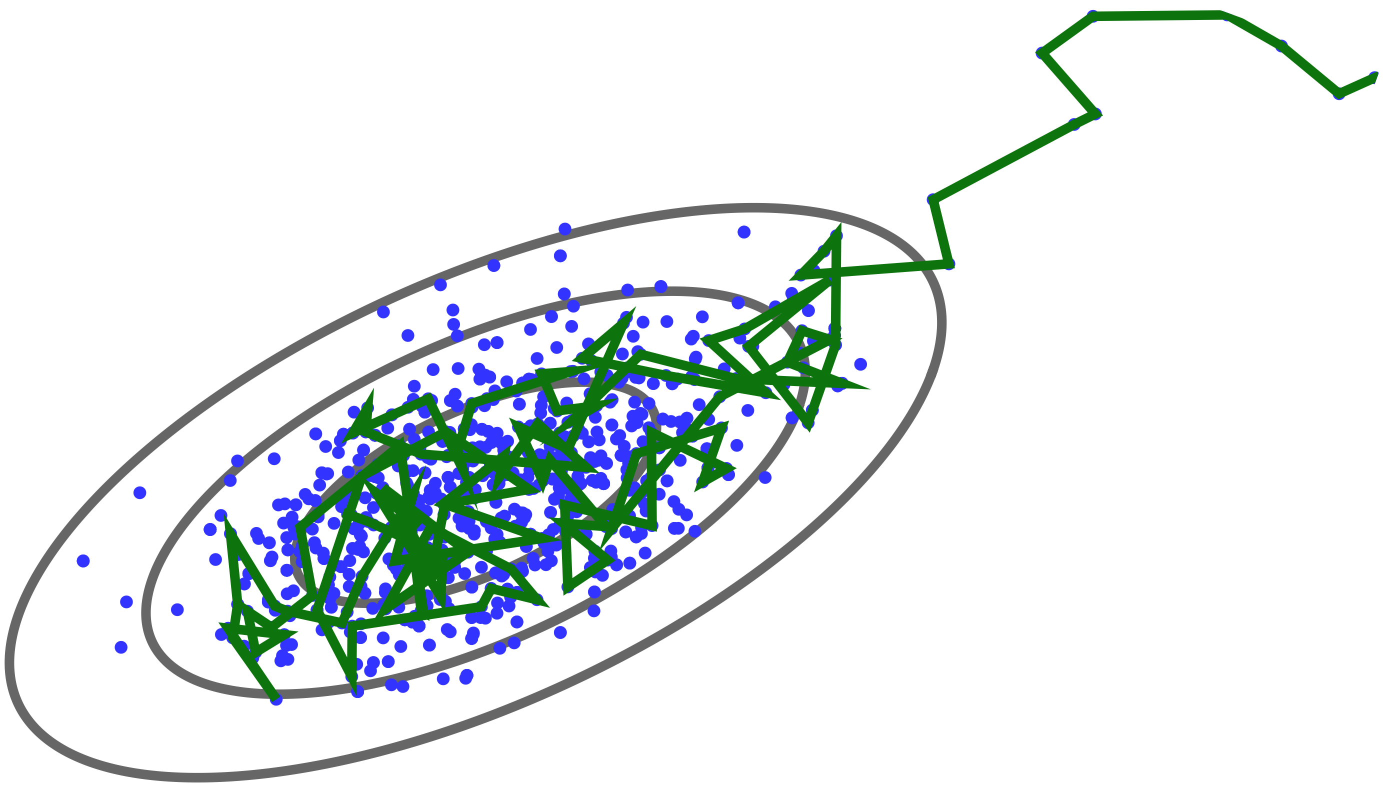
Monte Carlo Inference Methods



Overview

Motivation: Integration

Expectation, Normalization, Marginalization

Sampling methods:

Importance, Rejection, Metropolis–Hastings, Gibbs, Slice, Hybrid Monte Carlo

Open problems:

Simple Monte Carlo Integration

$$\int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta \;=\; \text{“average over } \pi \text{ of } f\text{”}$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)}), \qquad \theta^{(s)} \sim \pi$$

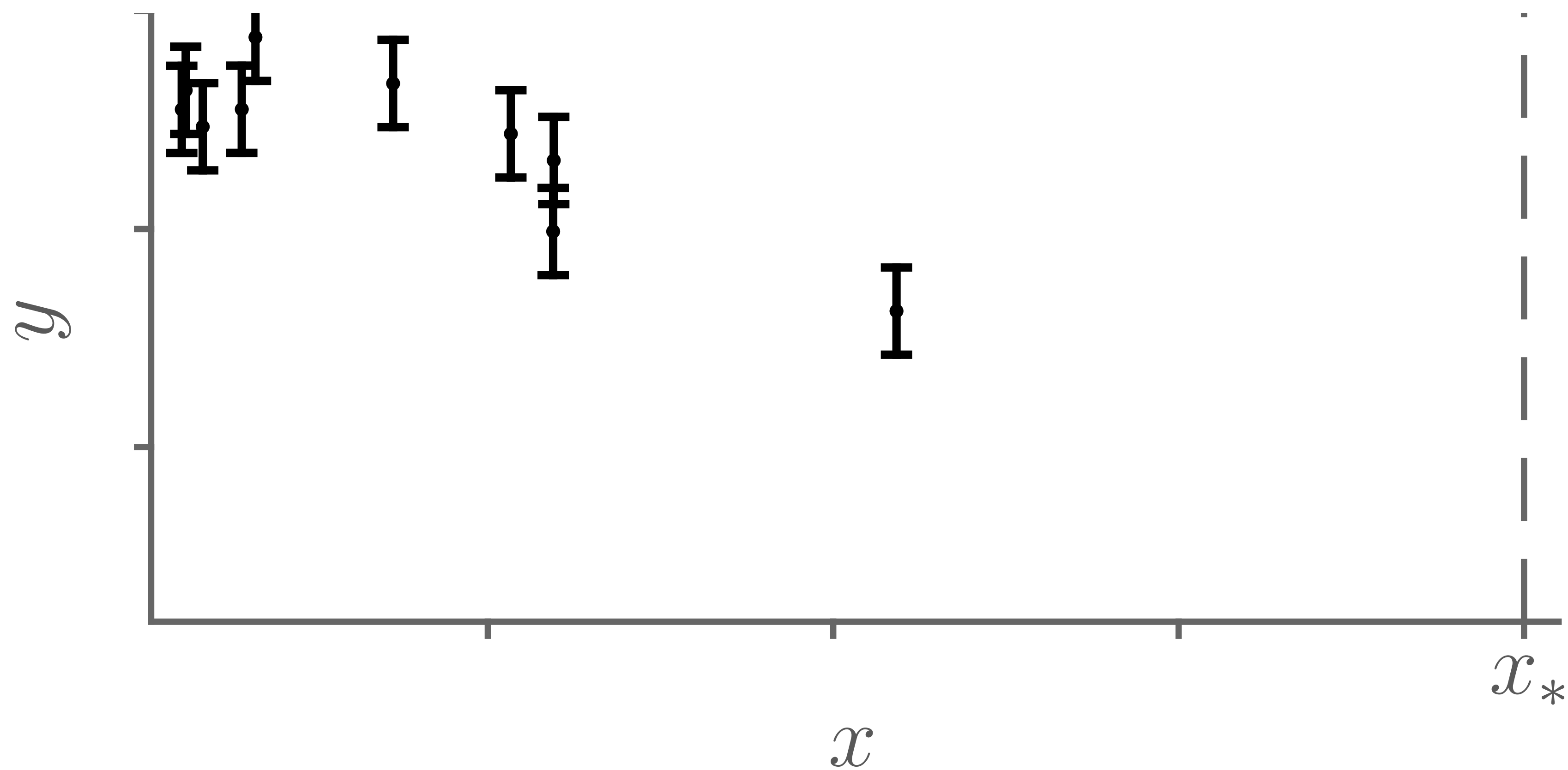
Unbiased

Variance $\sim 1/S$

Prediction

$$p(y_* \mid x_*, \mathcal{D}) = \int p(y_* \mid x_*, \theta)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta$$

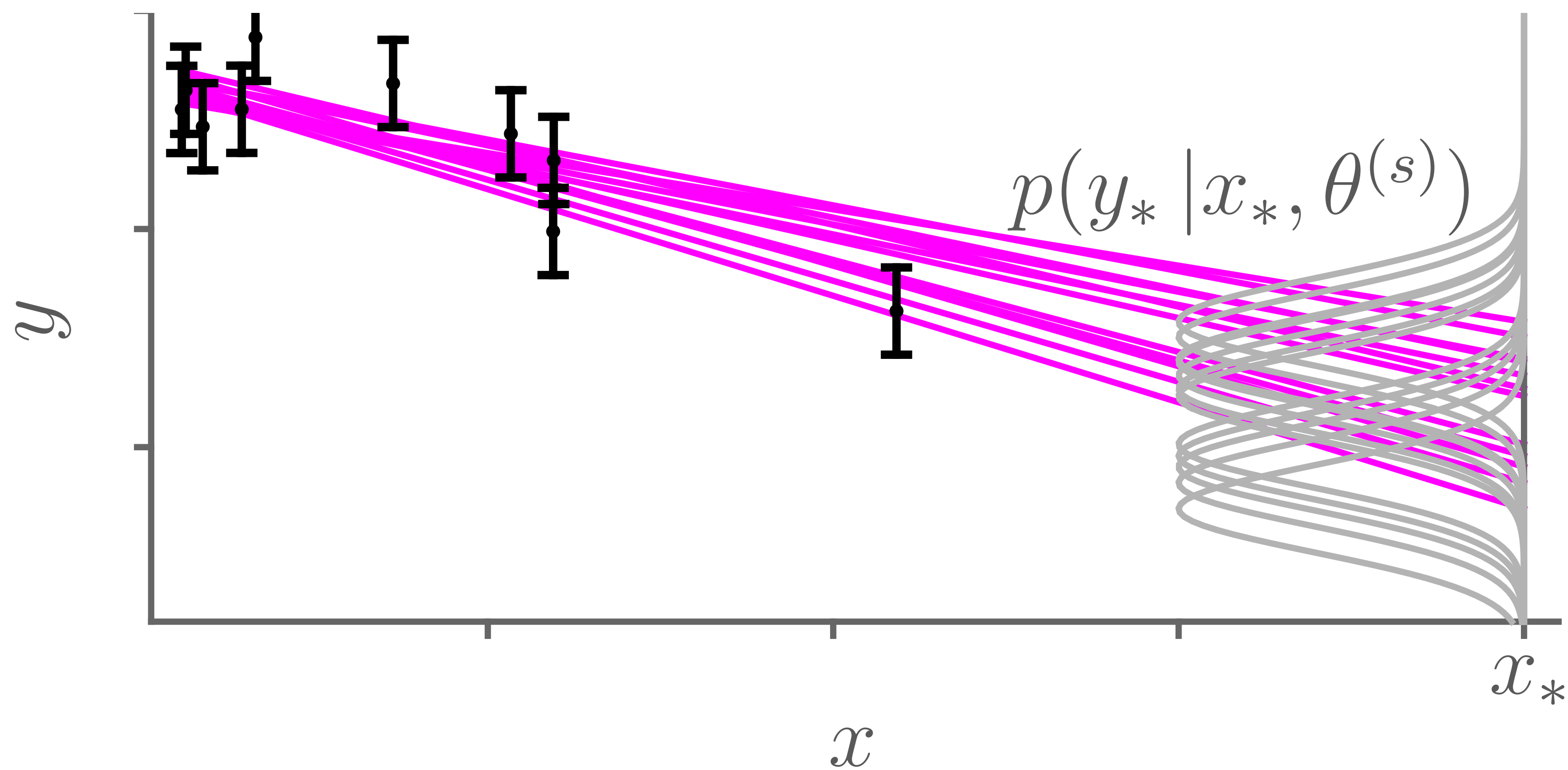
$$\approx \frac{1}{S} \sum_s p(y_* \mid x_*, \theta^{(s)}), \qquad \theta^{(s)} \sim p(\theta \mid \mathcal{D})$$



Prediction

$$p(y_* \,|\, x_*, \mathcal{D}) = \int p(y_* \,|\, x_*, \theta)\, p(\theta \,|\, \mathcal{D})\, \mathrm{d}\theta$$

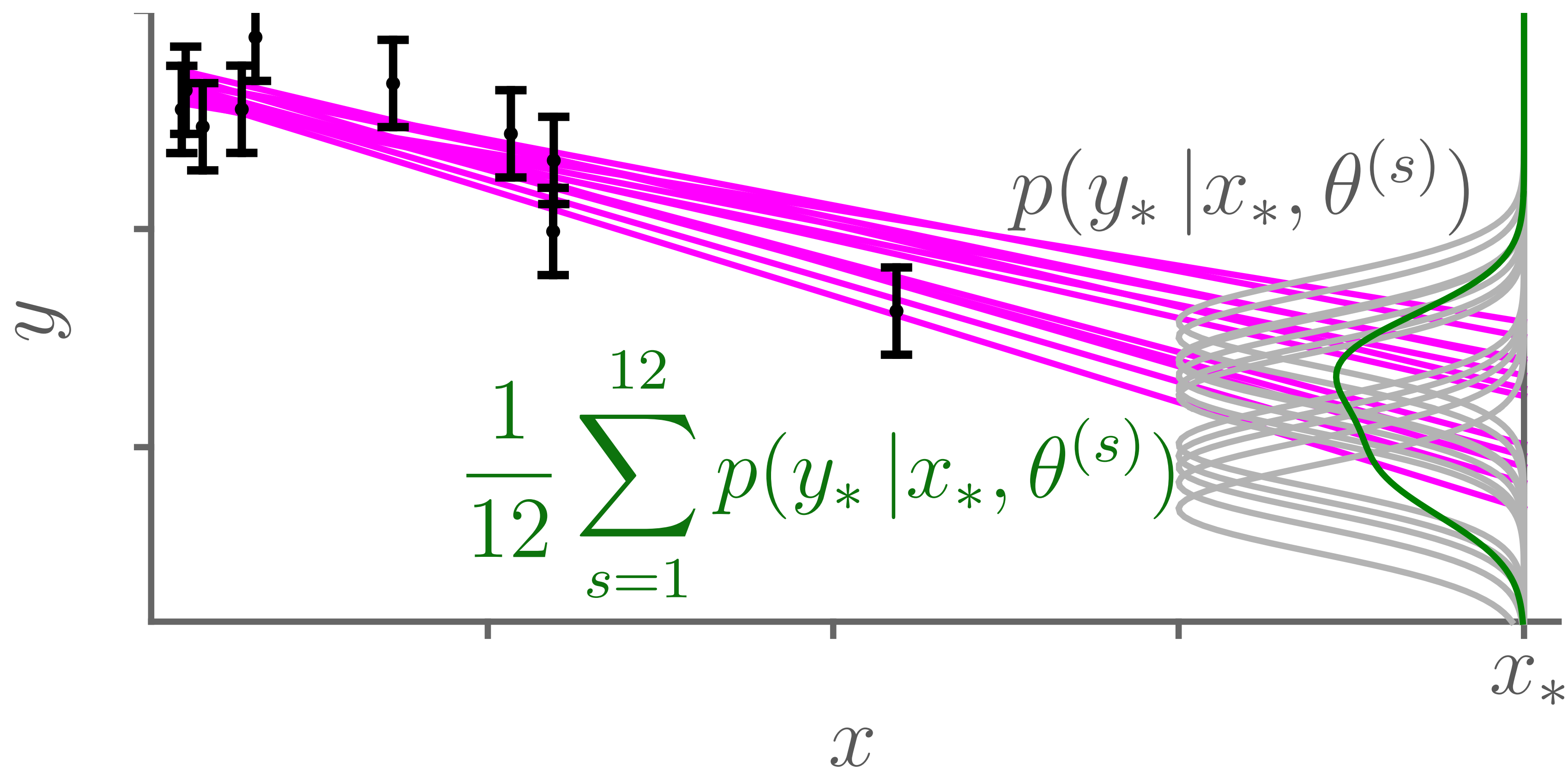
$$\approx \frac{1}{S} \sum_s p(y_* \,|\, x_*, \theta^{(s)}), \qquad \theta^{(s)} \sim p(\theta \,|\, \mathcal{D})$$



Prediction

$$p(y_* \mid x_*, \mathcal{D}) = \int p(y_* \mid x_*, \theta)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta$$

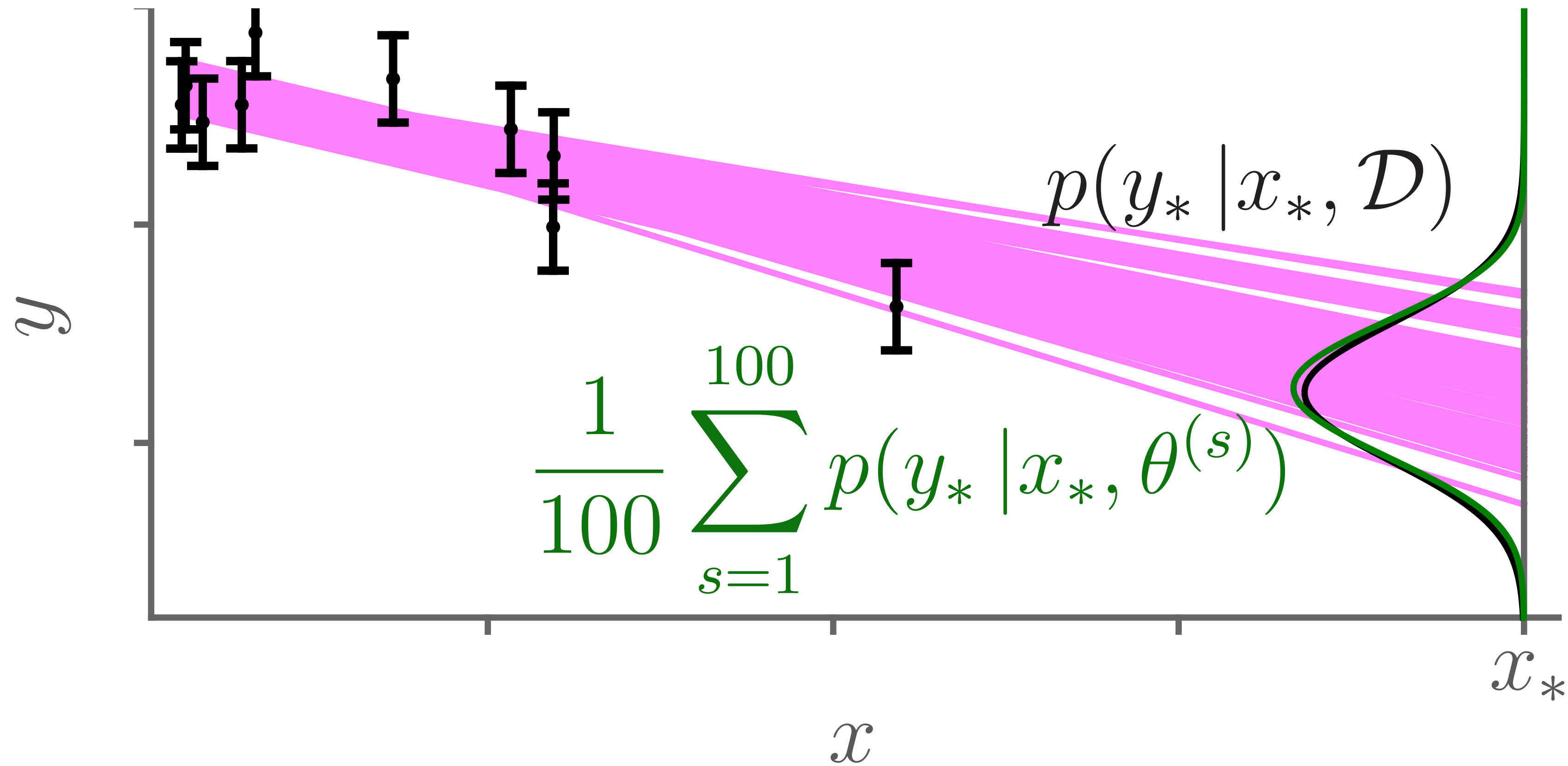
$$\approx \frac{1}{S} \sum_s p(y_* \mid x_*, \theta^{(s)}), \qquad \theta^{(s)} \sim p(\theta \mid \mathcal{D})$$



Prediction

$$p(y_* \mid x_*, \mathcal{D}) = \int p(y_* \mid x_*, \theta)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta$$

$$\approx \frac{1}{S} \sum_{s} p(y_* \mid x_*, \theta^{(s)}), \quad \theta^{(s)} \sim p(\theta \mid \mathcal{D})$$



Inference

Observe data: $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}$

Unknowns: $\theta = \{\mathbf{w}, \alpha, \epsilon, \Sigma, \{z^{(n)}\}, \ldots\}$

$$p(\theta \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \theta)\, p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D}, \theta)$$

Marginalization

Interested in particular parameter θ_i

$$p(\theta_i \mid \mathcal{D}) = \int p(\theta \mid \mathcal{D}) \, \mathrm{d}\theta_{\setminus i}$$

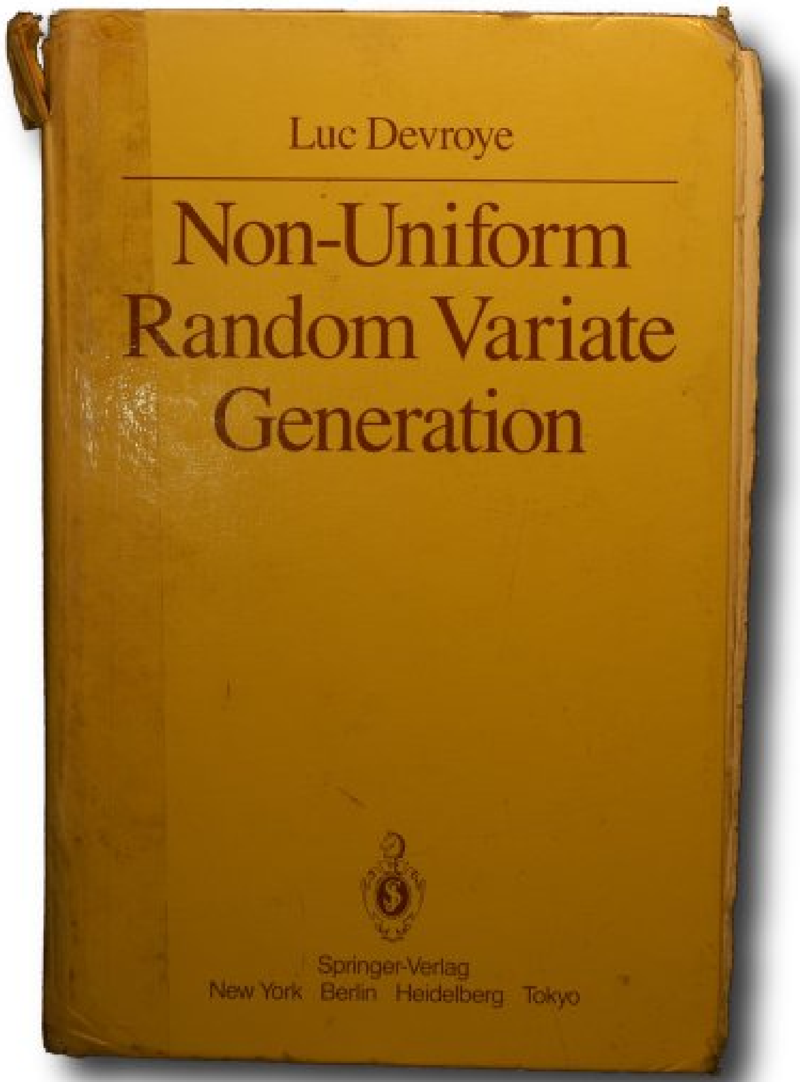
Sampling solution:

— Sample everything: $\theta^{(s)} \sim p(\theta \mid \mathcal{D})$

— $\theta_i^{(s)}$ comes from marginal $p(\theta_i \mid \mathcal{D})$

(But see also 'Rao–Blackwellization')

Sampling simple distributions



Use library routines for univariate distributions

(and some other special cases)

This book (free online) explains how some of them work

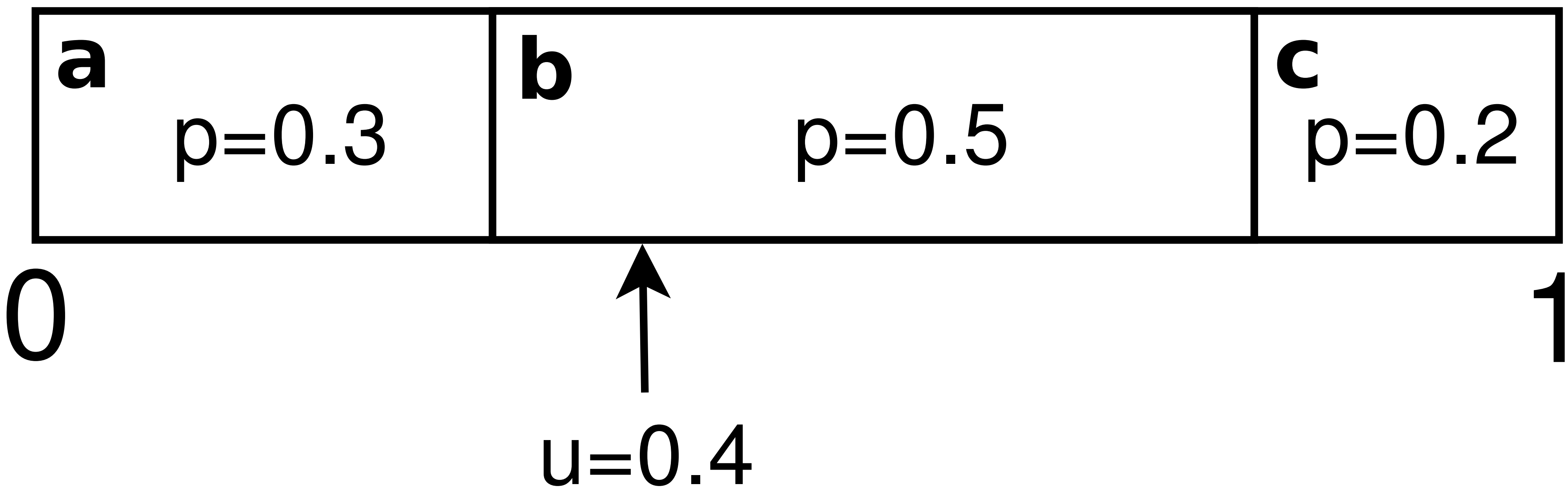
`http://luc.devroye.org/rnbookindex.html`

Target distribution

$$\pi(\theta) = \frac{\pi^*(\theta)}{\mathcal{Z}},$$

$$\text{e.g., } \pi^*(\theta) = p(\mathcal{D} \mid \theta)\, p(\theta)$$

Sampling discrete values



$$u \sim \mathrm{Uniform}[0, 1]$$

$$u = 0.4 \quad \Rightarrow \quad \theta = \mathsf{b}$$

Large number of samples? 1) Alias method, Devroye book; 2) Ex 6.3 MacKay book

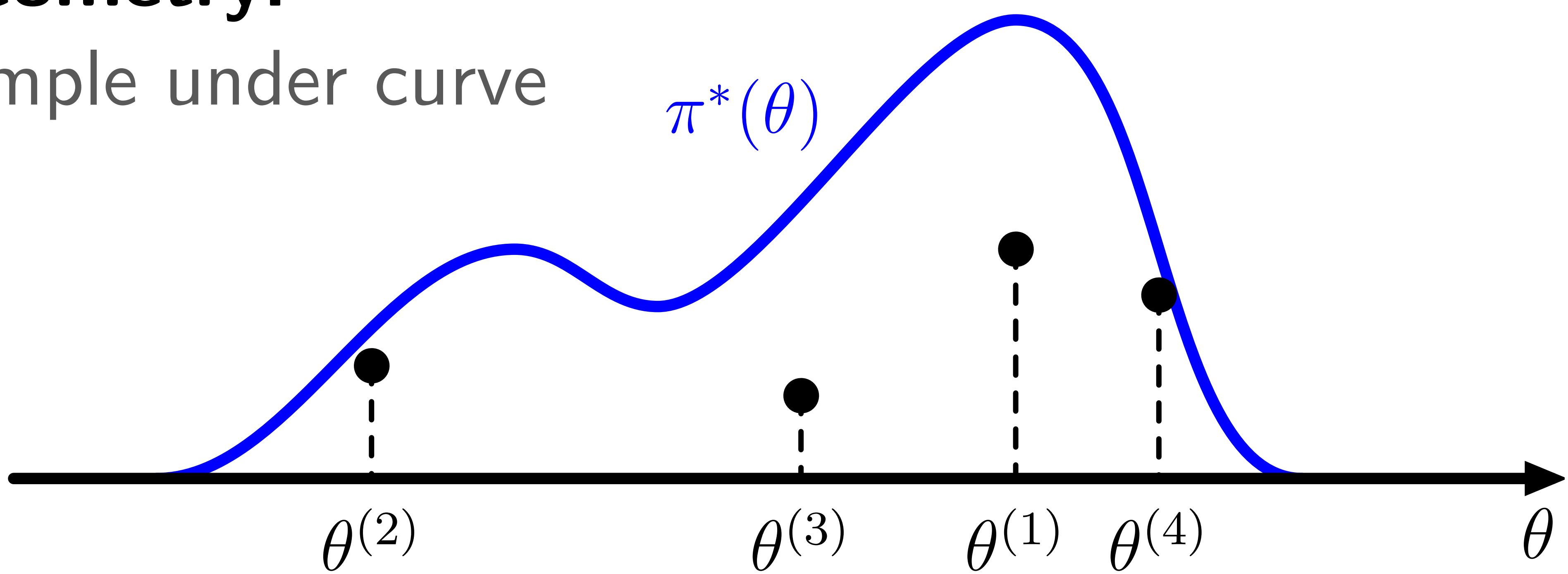
Sampling from a density

Math: $A^{(s)} \sim \mathrm{Uniform}[0,1], \quad \theta^{(s)} = \Phi^{-1}(A^{(s)})$

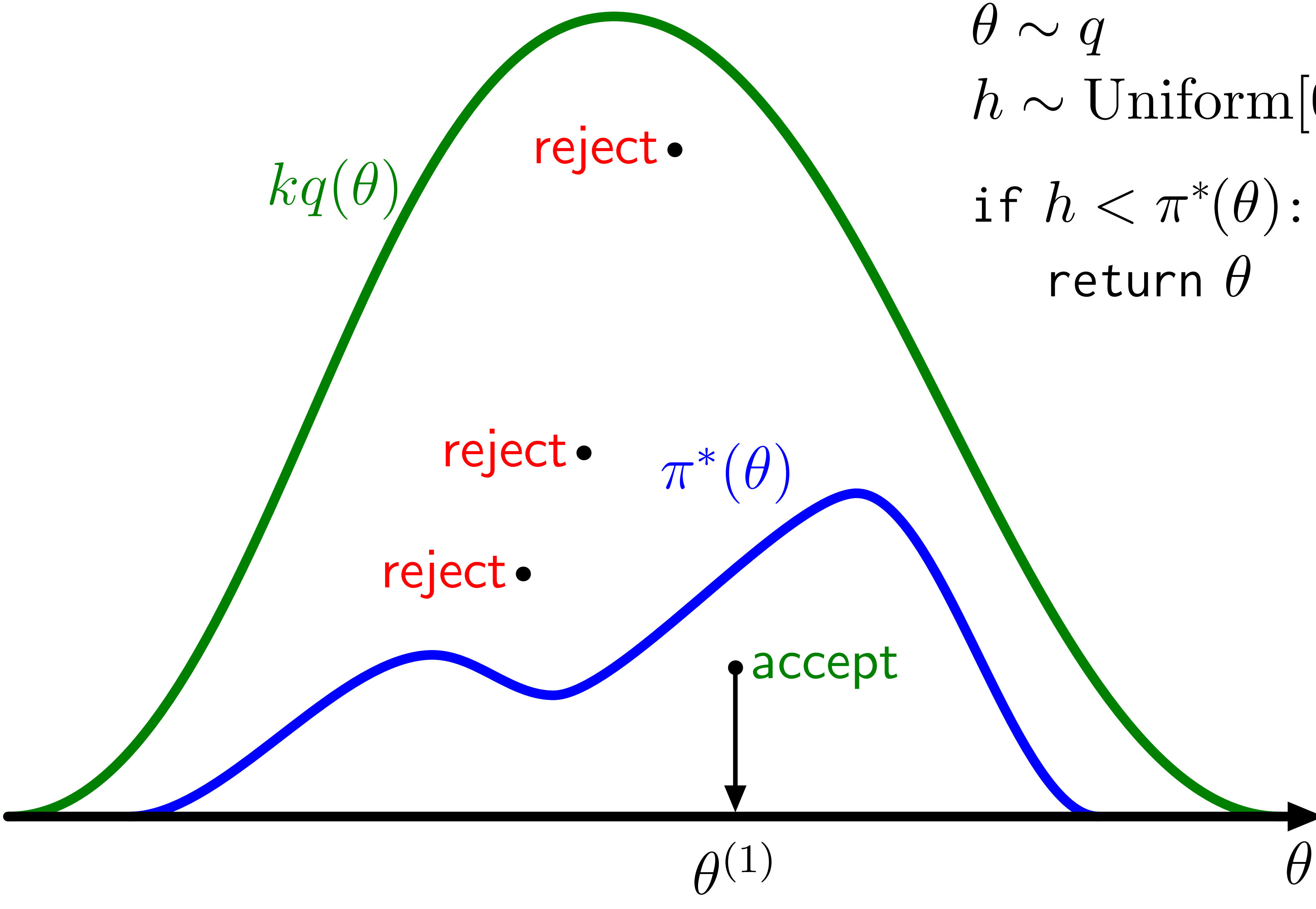
where cdf $\Phi(\theta) = \int_{-\infty}^{\theta} \pi(\theta')\, \mathrm{d}\theta'$

Geometry:

sample under curve

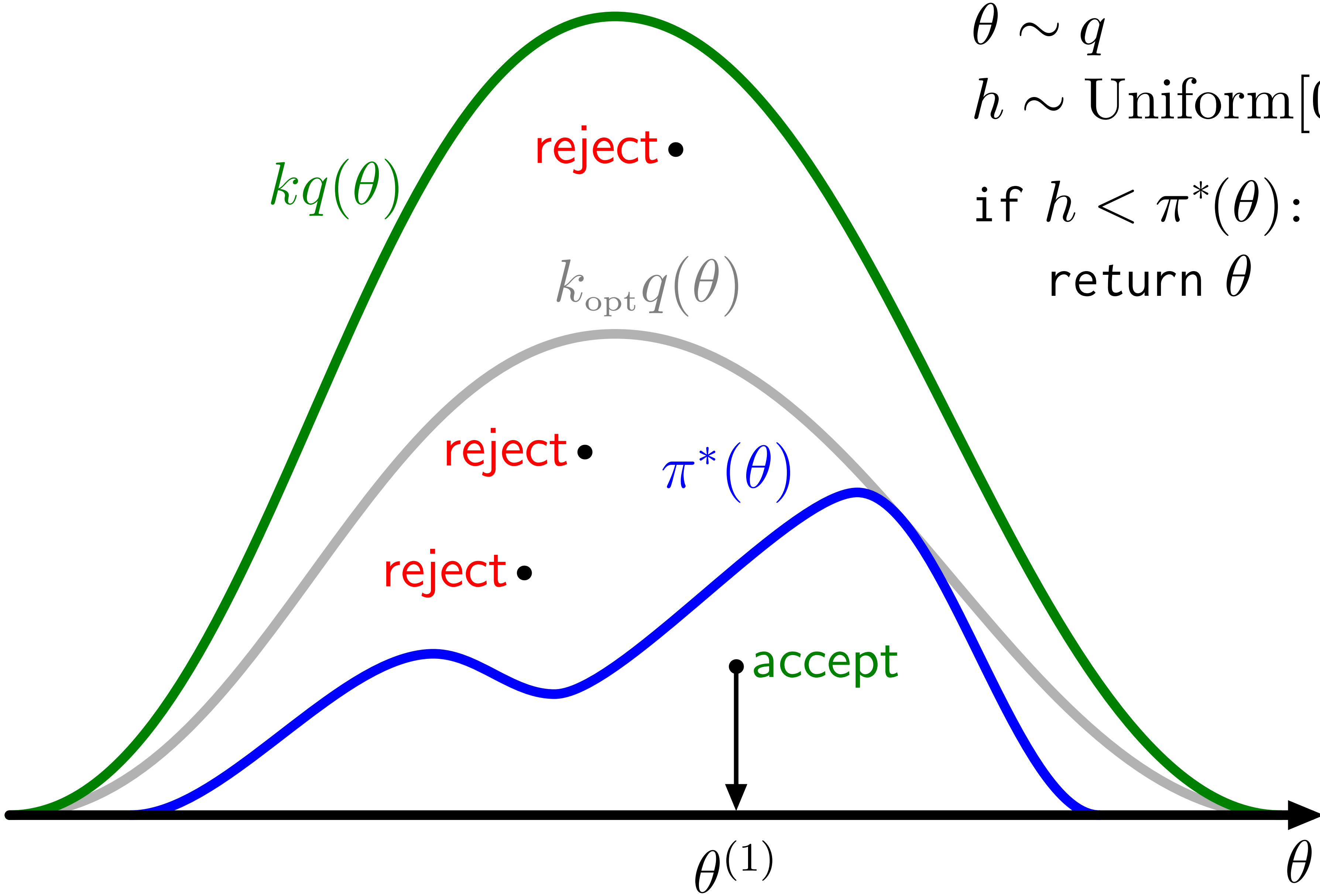


Rejection Sampling



while true:
  $\theta \sim q$
  $h \sim \mathrm{Uniform}[0, kq(\theta)]$
  if $h < \pi^*(\theta)$:
    return θ

Rejection Sampling



while true:

  $\theta \sim q$

  $h \sim \mathrm{Uniform}[0, kq(\theta)]$

  if $h < \pi^*(\theta)$:

    return θ

Importance Sampling

Rewrite integral: expectation under simple distribution q:

$$\int f(\theta)\,\pi(\theta)\;\mathrm{d}\theta \;=\; \int f(\theta)\,\frac{\pi(\theta)}{q(\theta)}\,q(\theta)\;\mathrm{d}\theta,$$

$$\approx\; \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)})\,\frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}, \qquad \theta^{(s)} \sim q$$

Unbiased if $q(\theta)>0$ where $\pi(\theta)>0$. Can have infinite variance.

Importance Sampling 2

Can't evaluate $\pi(\theta) = \dfrac{\pi^*(\theta)}{\mathcal{Z}}$, $\quad \mathcal{Z} = \int \pi^*(\theta)\, \mathrm{d}\theta$

Alternative version:

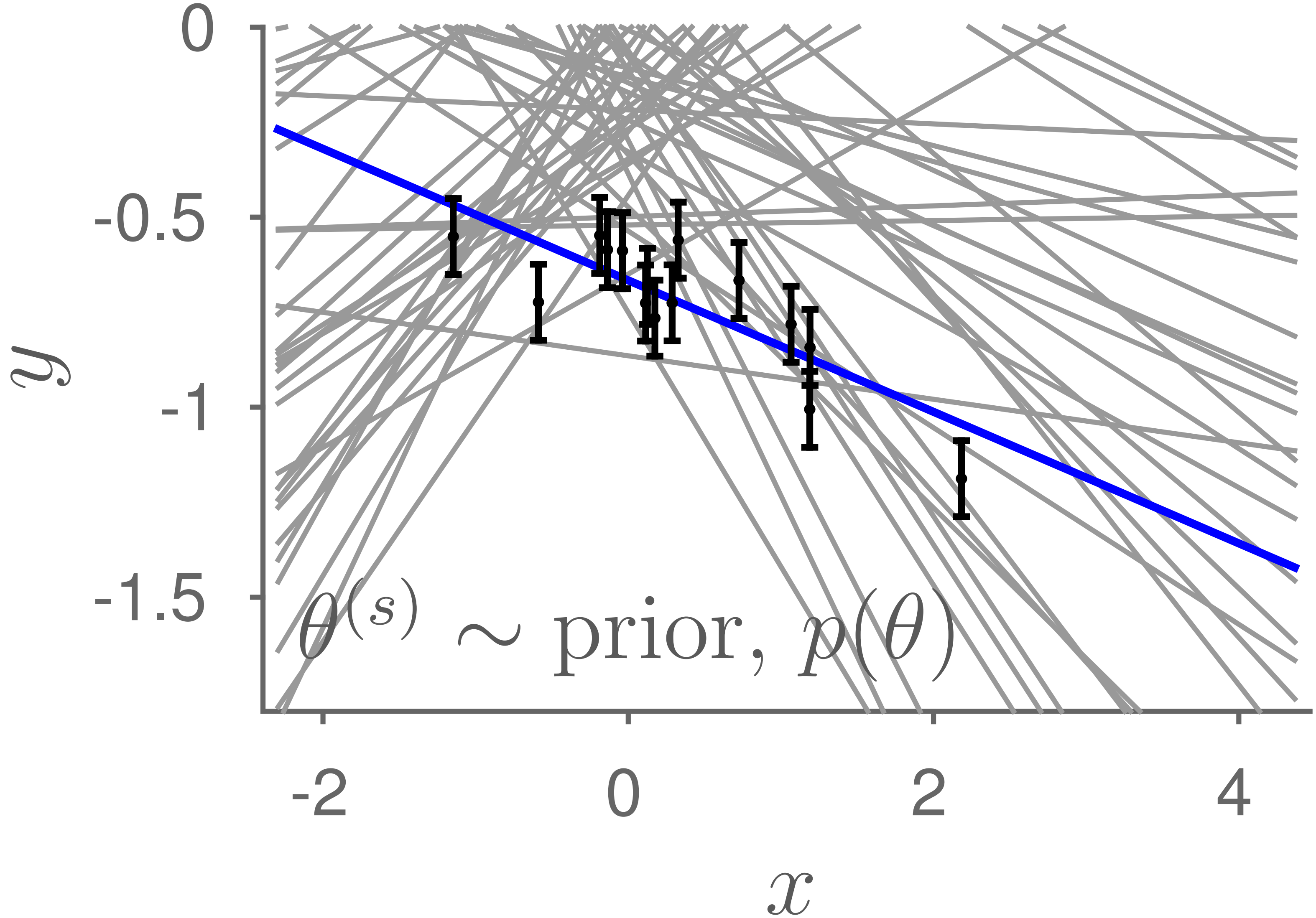
$$\theta^{(s)} \sim q, \qquad r^{*(s)} = \frac{\pi^*(\theta^{(s)})}{q(\theta^{(s)})}, \qquad r^{(s)} = \frac{r^{*(s)}}{\sum_{s'} r^{*(s')}}$$

Biased but consistent estimator:

$$\int f(\theta)\, \pi(\theta)\, \mathrm{d}\theta \;\approx\; \sum_{s=1}^{S} f(\theta^{(s)})\, r^{(s)}$$

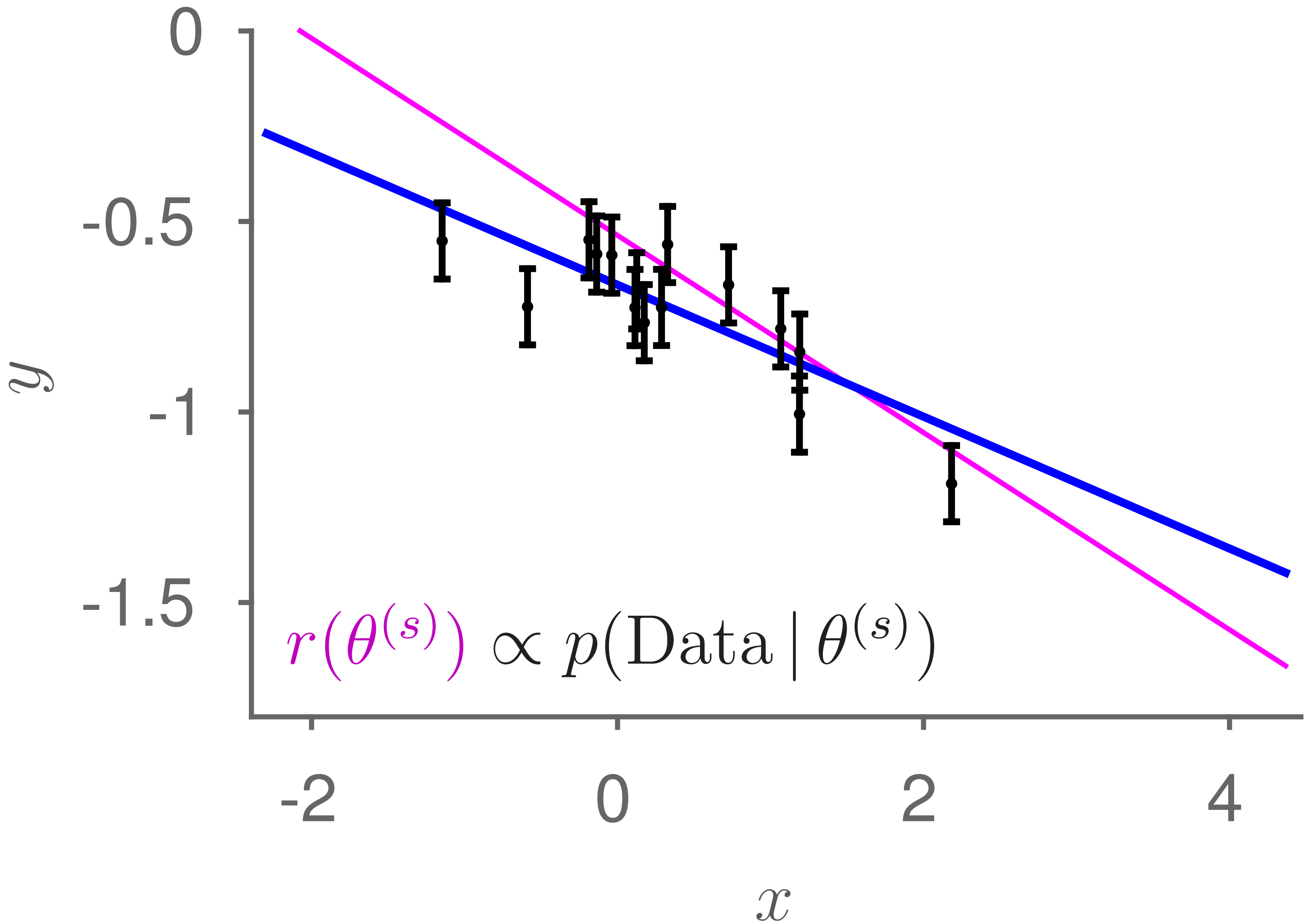
Linear regression

60 samples from prior:



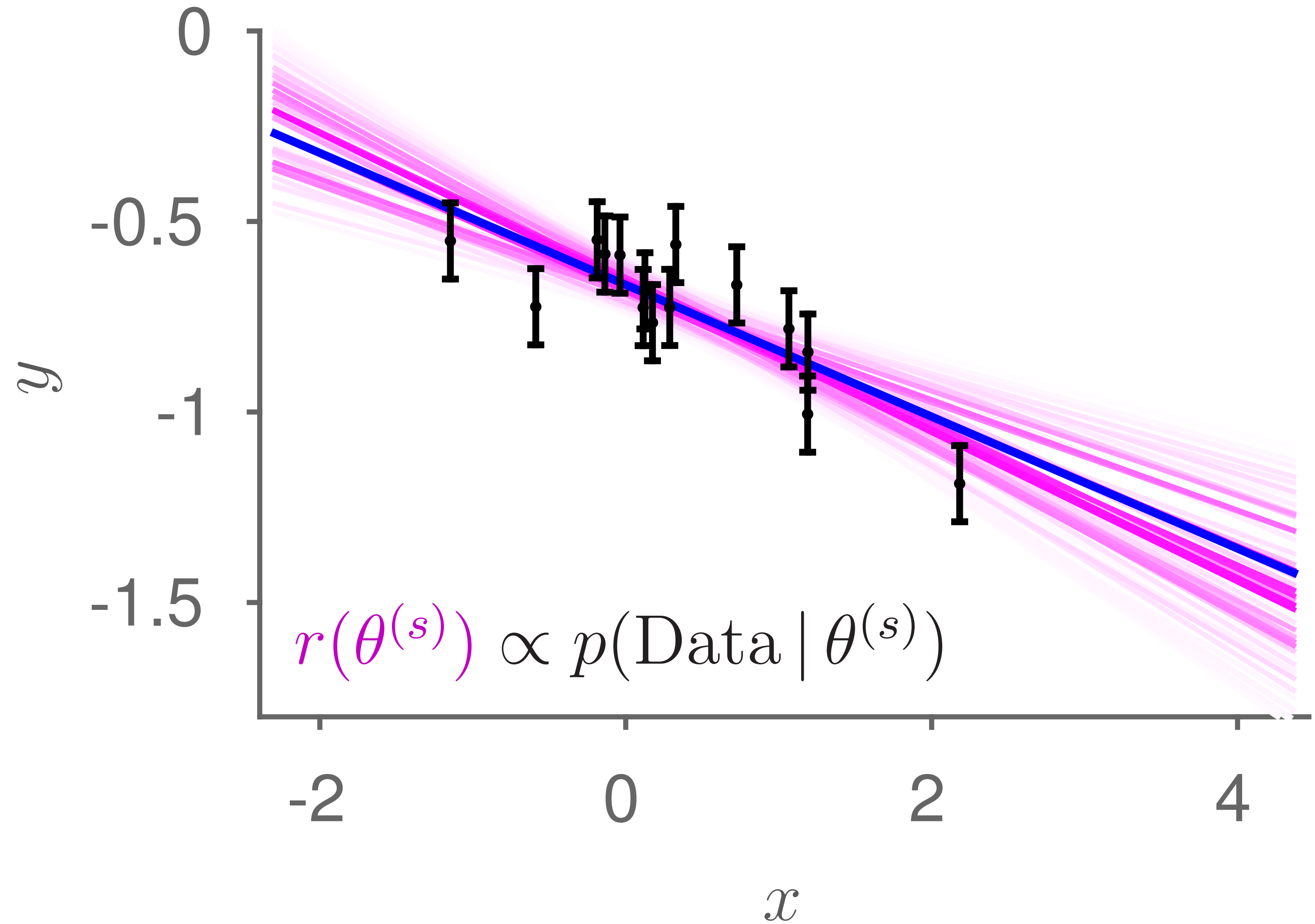
Linear regression

60 samples from prior, importance reweighted:



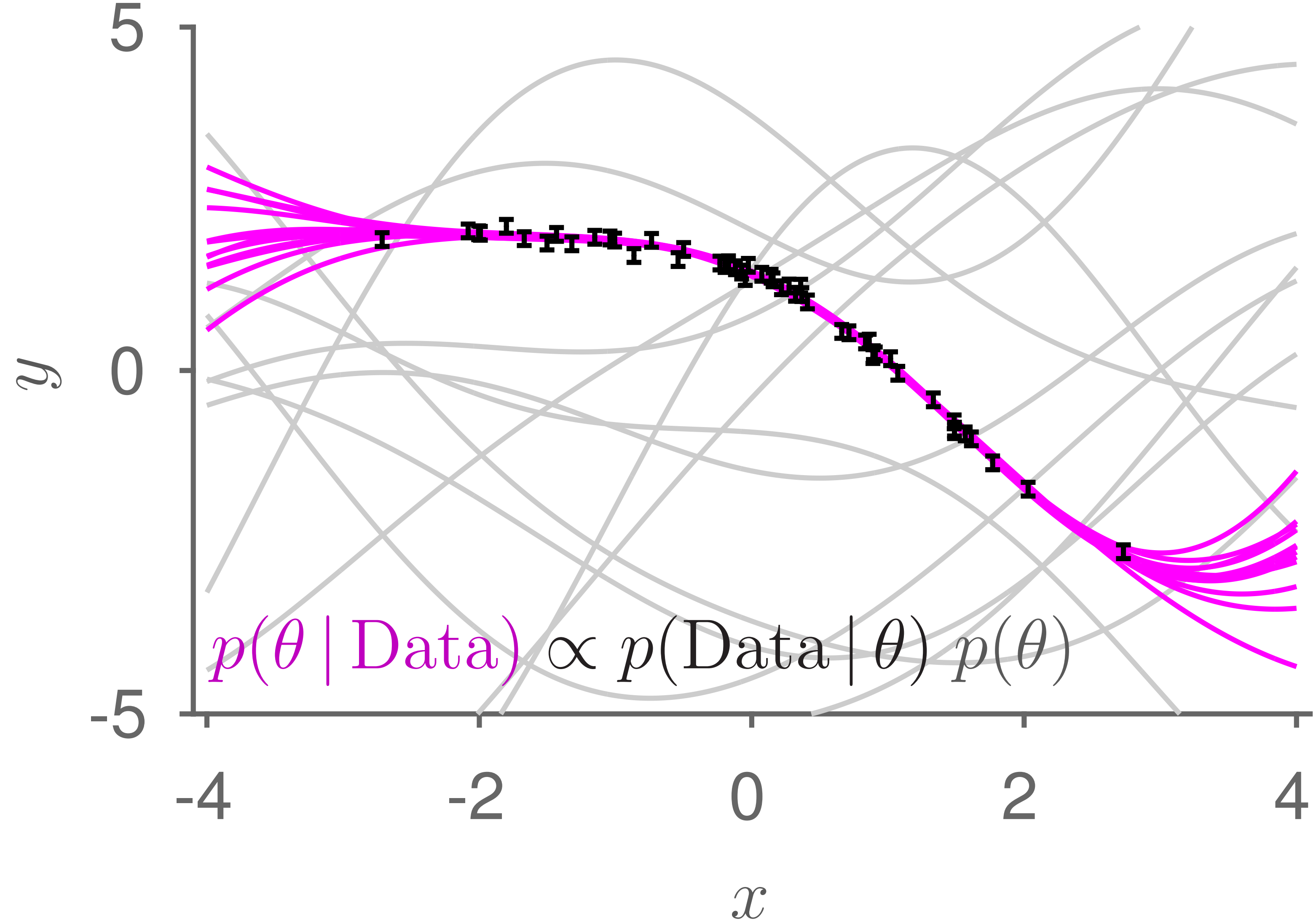
Linear regression

10,000 samples from prior, importance reweighted:



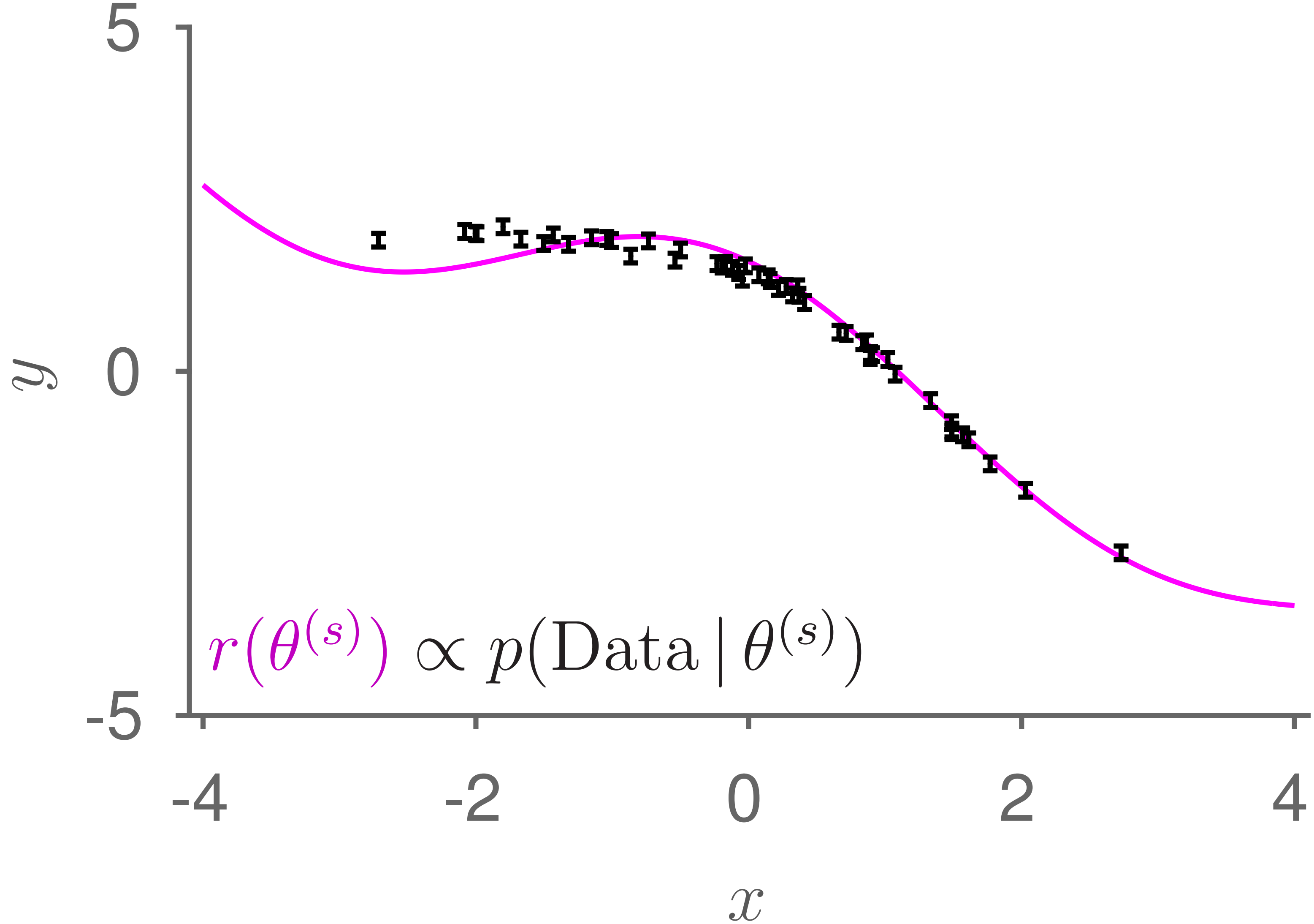
High dimensional θ

12 curves from prior and 12 curves from posterior



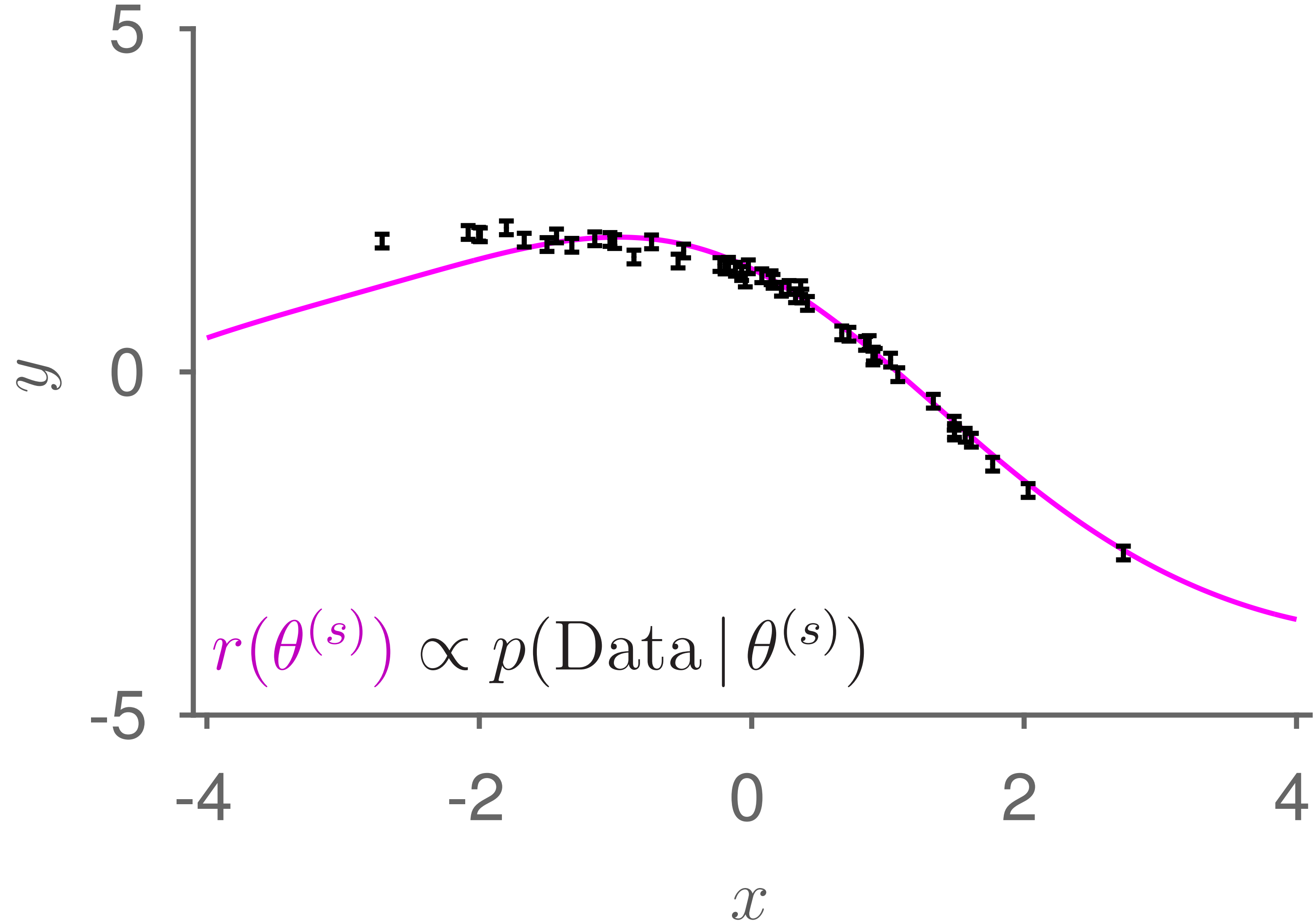
High dimensional θ

10,000 samples from prior, importance reweighted:



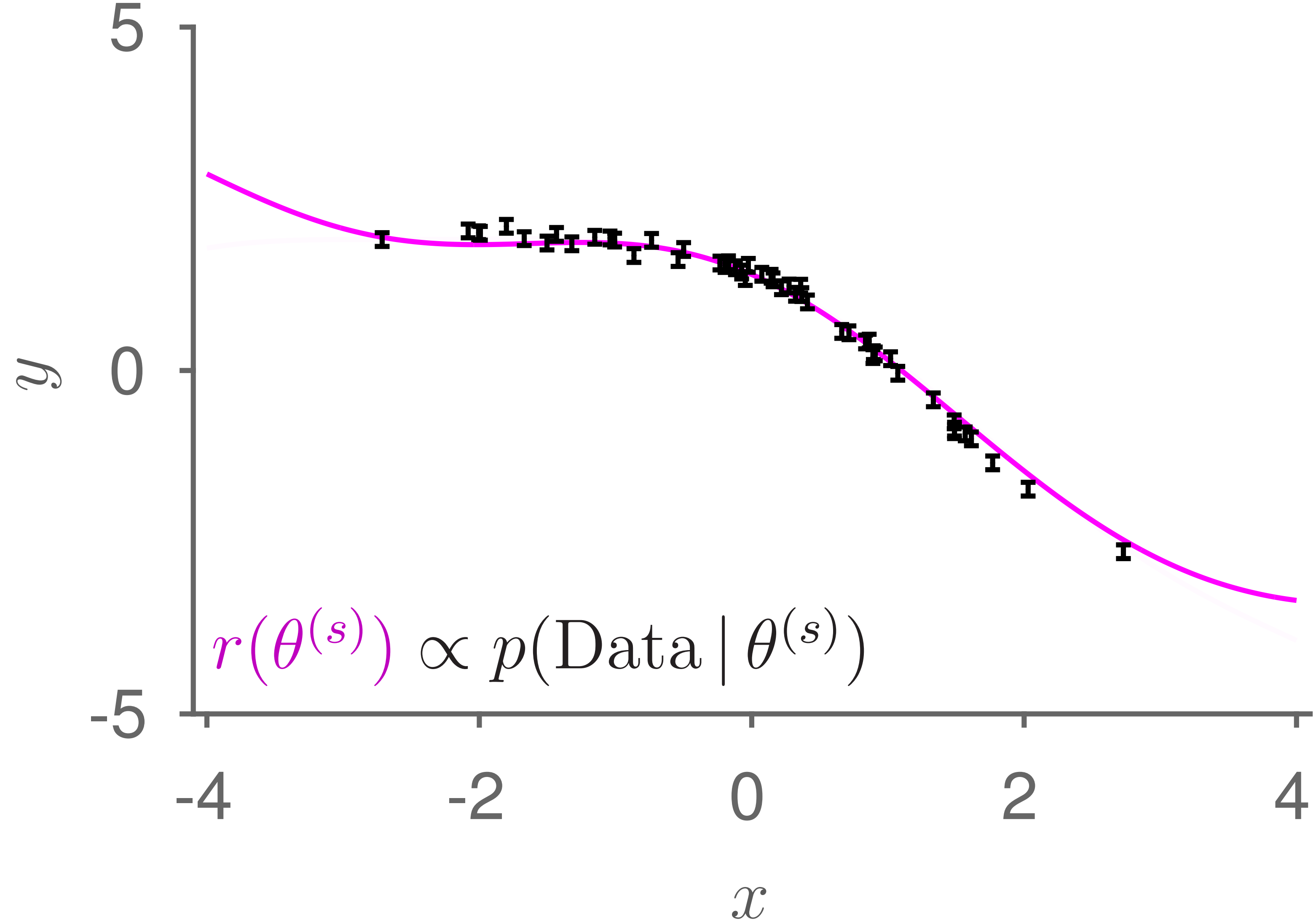
High dimensional θ

100,000 samples from prior, importance reweighted:



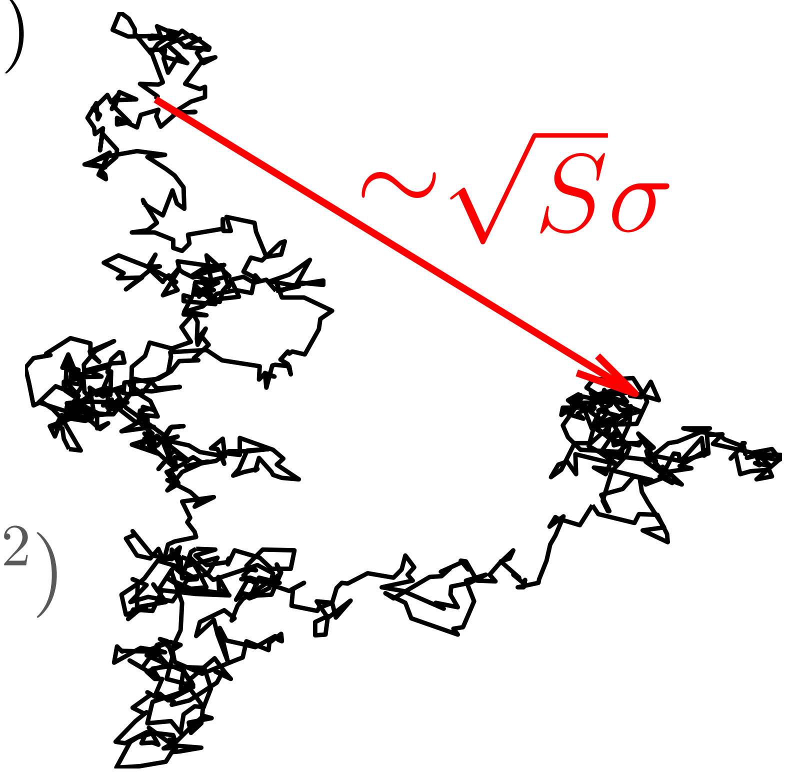
High dimensional θ

1,000,000 samples from prior, importance reweighted:



Markov chains

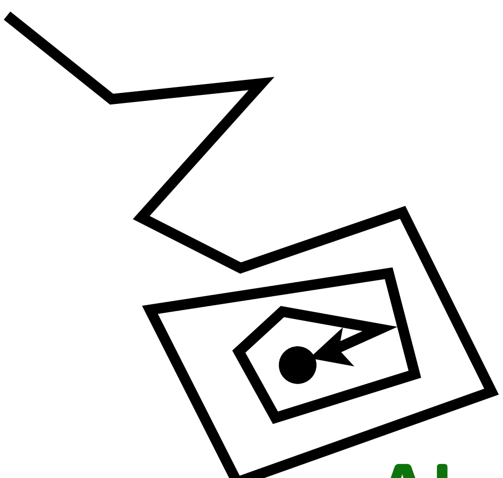
$$p(\theta^{(s+1)} \,|\, \theta^{(1)} \ldots \theta^{(s)}) = T(\theta^{(s+1)} \!\leftarrow\! \theta^{(s)})$$



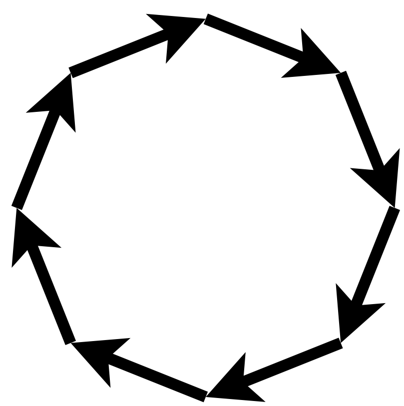
Divergent

e.g., random walk diffusion

$$T(\theta^{(s+1)} \!\leftarrow\! \theta^{(s)}) = \mathcal{N}(\theta^{(s+1)};\, \theta^{(s)}, \sigma^2)$$

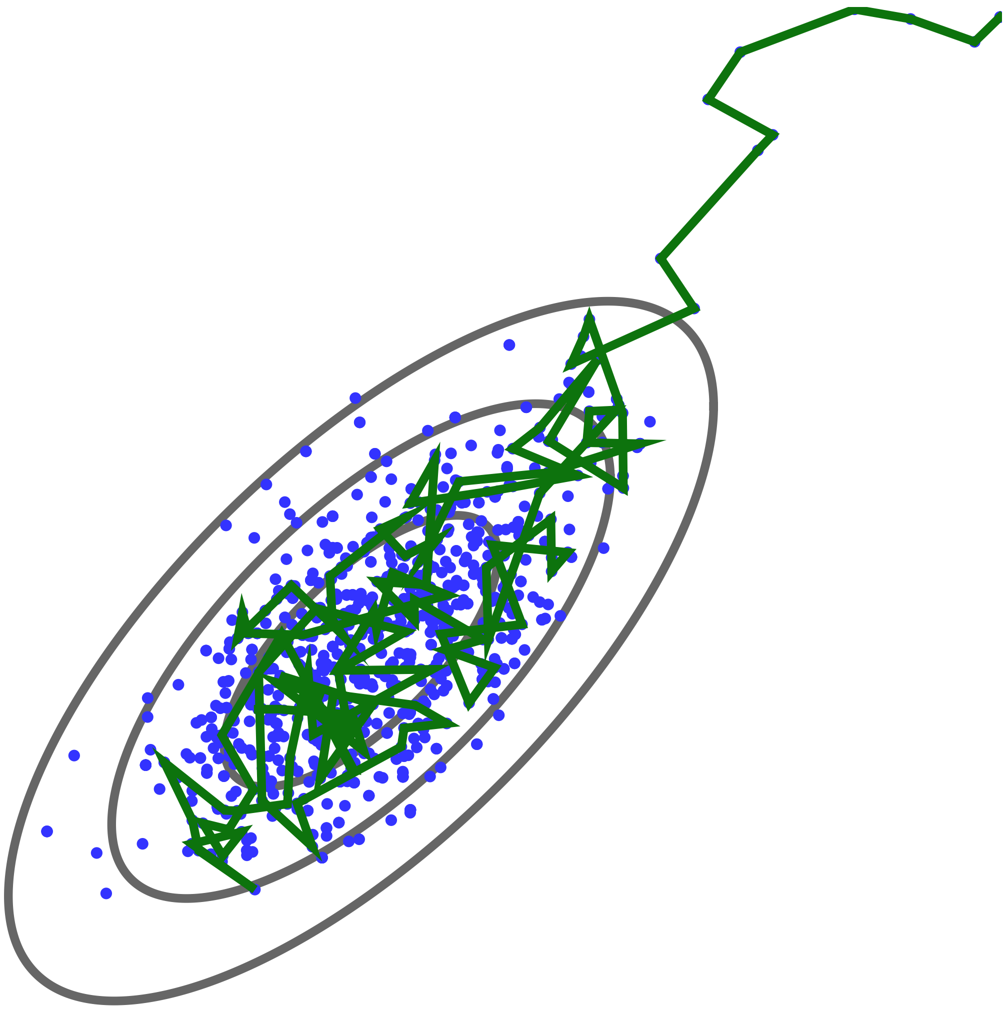


Absorbing states



Cycles

Markov chain equilibrium $\pi(\theta)$



$$\lim_{s\to\infty} p(\theta^{(s)}) = \pi(\theta^{(s)})$$

'Ergodic'

if true for all $\theta^{(s=0)}$

(other definitions of ergodic exist)

Possible to get anywhere in K steps,

($T^K(\theta' \leftarrow \theta) > 0$ for all pairs) $\Rightarrow$ no cycles or islands

Invariant/stationary condition

If $\theta^{(s)}$ is a sample from π,

$\theta^{(s+1)}$ is also a sample from π.

$$p(\theta') = \int T(\theta' \!\leftarrow\! \theta)\, \pi(\theta)\, \mathrm{d}\theta = \pi(\theta')$$

Metropolis–Hastings

$$\theta' \sim q(\theta'; \theta^{(s)})$$

```
if accept:
```

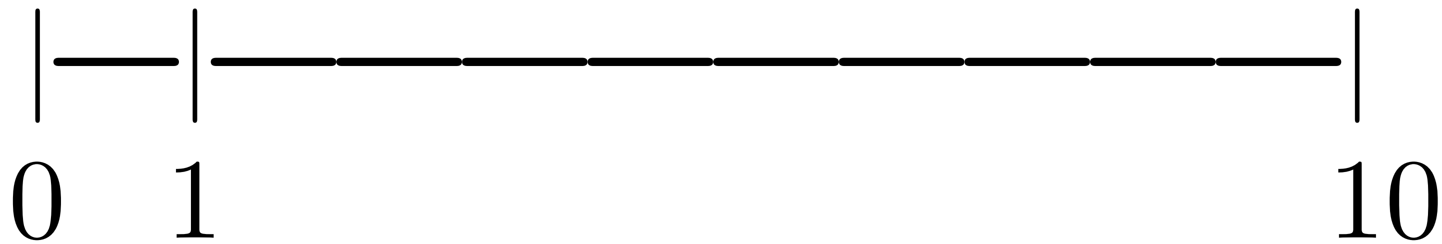
$$\theta^{(s+1)} \leftarrow \theta'$$

```
else:
```

$$\theta^{(s+1)} \leftarrow \theta^{(s)}$$

$$P(\text{accept}) = \min\left(1,\ \frac{\pi^*(\theta')\, q(\theta^{(s)}; \theta')}{\pi^*(\theta^{(s)})\, q(\theta'; \theta^{(s)})}\right)$$

Example / warning



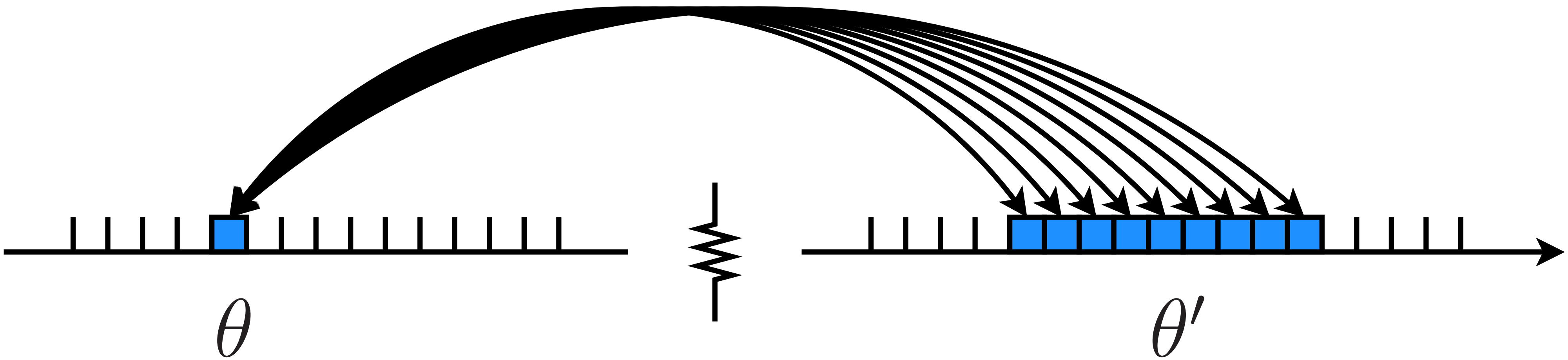
Proposal: $$\begin{cases} \theta^{(s+1)} = 9\theta^{(s)} + 1, & 0 < \theta^{(s)} < 1 \\ \theta^{(s+1)} = (\theta^{(s)} - 1)/9, & 1 < \theta^{(s)} < 10 \end{cases}$$

Accept move with probability:

$$\min\left(1, \frac{\pi^*(\theta')\, q(\theta; \theta')}{\pi^*(\theta)\, q(\theta'; \theta)}\right) = \min\left(1, \frac{\pi^*(\theta')}{\pi^*(\theta)}\right)$$

(Wrong!)

Example / warning



Accept θ' with probability:

$$\min\left(1,\ \frac{q(\theta;\theta')\,\pi^*(\theta')}{q(\theta';\theta)\,\pi^*(\theta)}\right) = \min\left(1,\ \frac{1}{1/9}\,\frac{\pi^*(\theta')}{\pi^*(\theta)}\right)$$

Really, Green (1995):

$$\min\left(1,\ \left|\frac{\partial\theta'}{\partial\theta}\right|\frac{\pi^*(\theta')}{\pi^*(\theta)}\right) = \min\left(1,\ 9\,\frac{\pi^*(\theta')}{\pi^*(\theta)}\right)$$

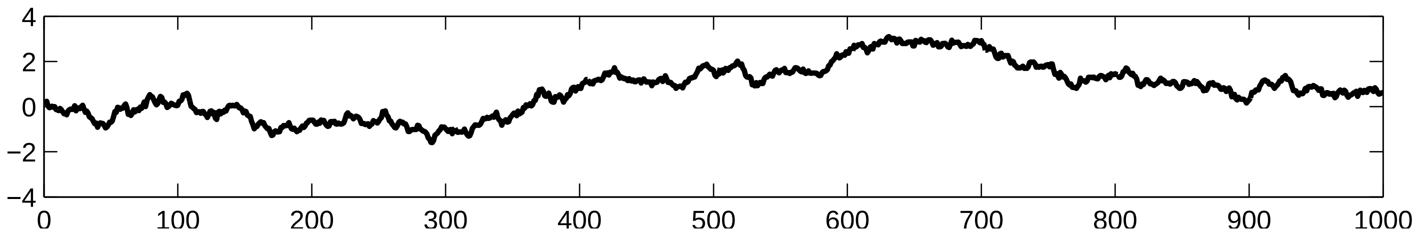
Step-size demo

Explore $\mathcal{N}(0,1)$ with different step sizes σ

```
sigma = @(s) plot(metropolis(0, @(x)-0.5*x*x, 1e3, s));
```

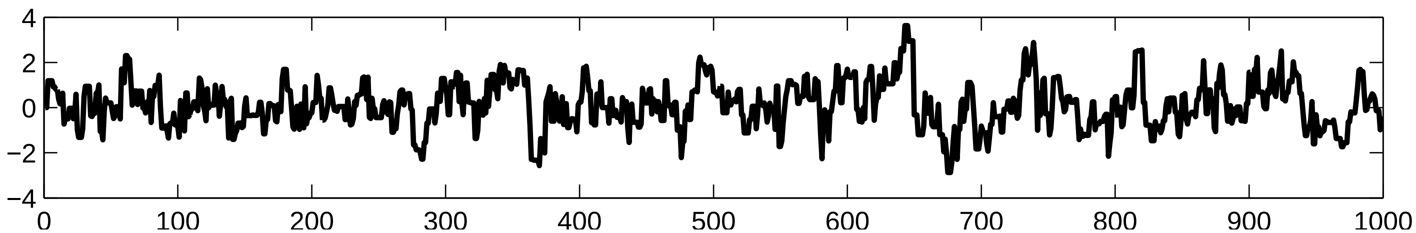
`sigma(0.1)`

99.8% accepts



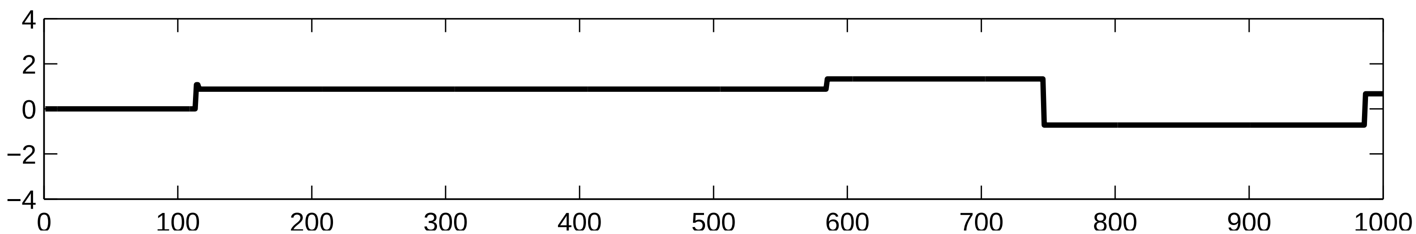
`sigma(1)`

68.4% accepts

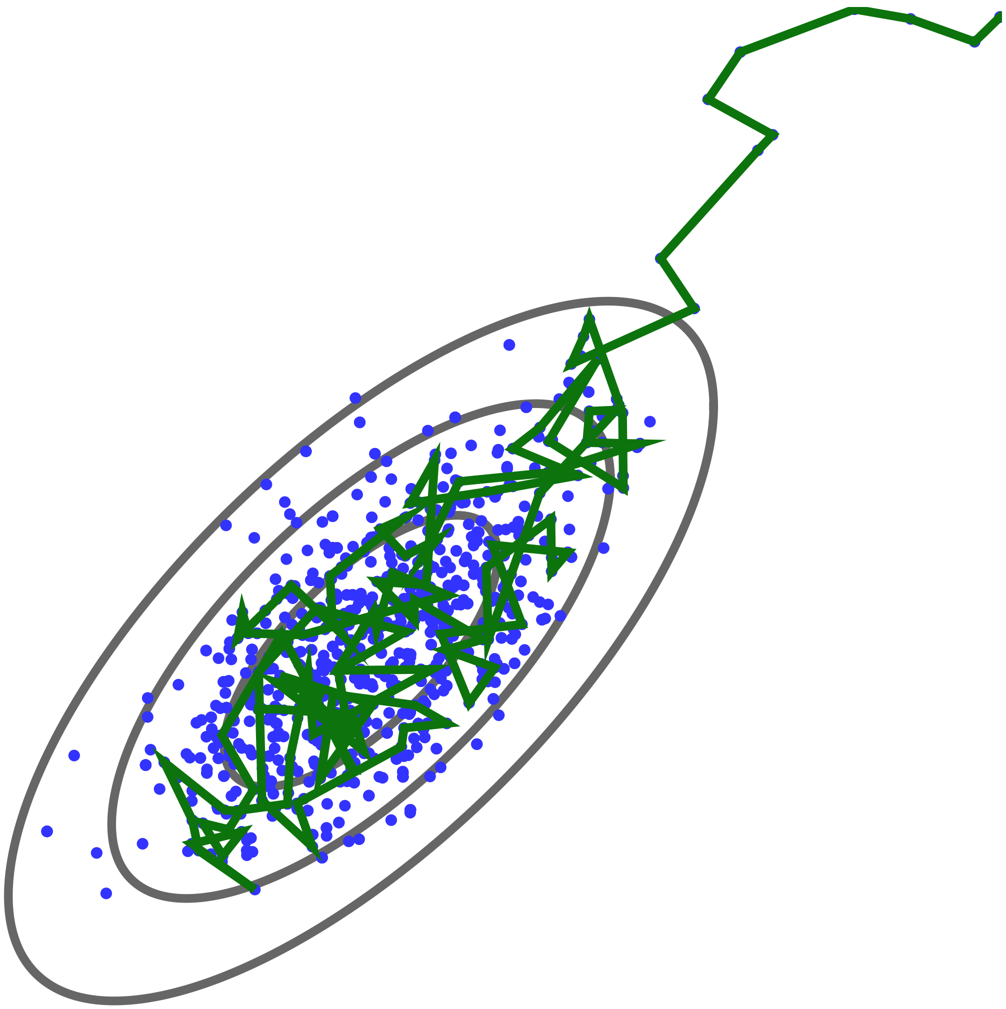


`sigma(100)`

0.5% accepts



Markov chain Monte Carlo (MCMC)



User chooses $\pi(\theta)$

Explore some plausible θ

For large s:

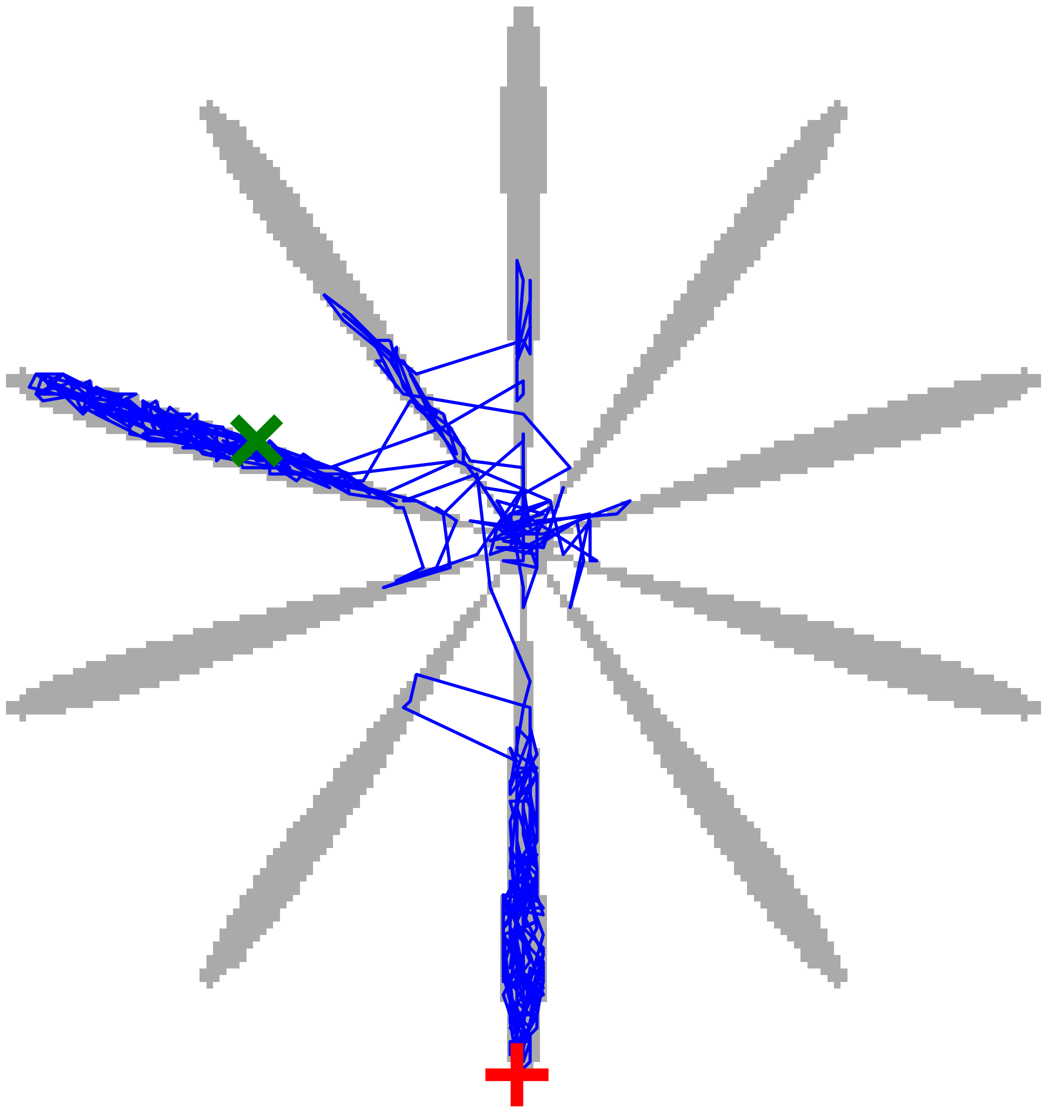
$$p(\theta^{(s)}) = \pi(\theta^{(s)})$$

$$p(\theta^{(s+1)}) = \pi(\theta^{(s+1)})$$

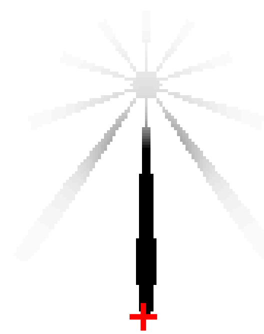
$$\mathbb{E}\left[f(\theta^{(s)})\right] = \mathbb{E}\left[f(\theta^{(s+1)})\right] = \int f(\theta)\pi(\theta)\,\mathrm{d}\theta = \lim_{S\to\infty}\frac{1}{S}\sum_s f(\theta^{(s)})$$

Markov chain mixing

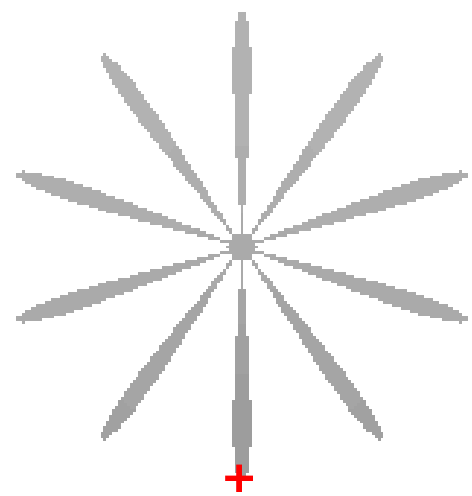
Initialization + $\rightarrow$ 2000 steps $\rightarrow$ **x** 'sample'



$p(\theta^{(s=100)})$



$p(\theta^{(s=2000)}) \approx \pi(\theta)$



Creating an MCMC scheme

M–H gives $T(\theta' \leftarrow \theta)$ with invariant π

Lots of options for $q(\theta'; \theta)$:

— Local diffusions

— Appproximations of π

— Update one variable or all?

— ...

Multiple valid operators T_A, T_B, T_C, ...

Composing operators

If $p(\theta^{(1)}) = \pi(\theta^{(1)})$

$$\theta^{(2)} \sim T_A(\cdot \leftarrow \theta^{(1)}) \quad \Rightarrow \quad p(\theta^{(2)}) = \pi(\theta^{(2)})$$

$$\theta^{(3)} \sim T_B(\cdot \leftarrow \theta^{(2)}) \quad \Rightarrow \quad p(\theta^{(3)}) = \pi(\theta^{(3)})$$

$$\theta^{(4)} \sim T_C(\cdot \leftarrow \theta^{(3)}) \quad \Rightarrow \quad p(\theta^{(4)}) = \pi(\theta^{(4)})$$

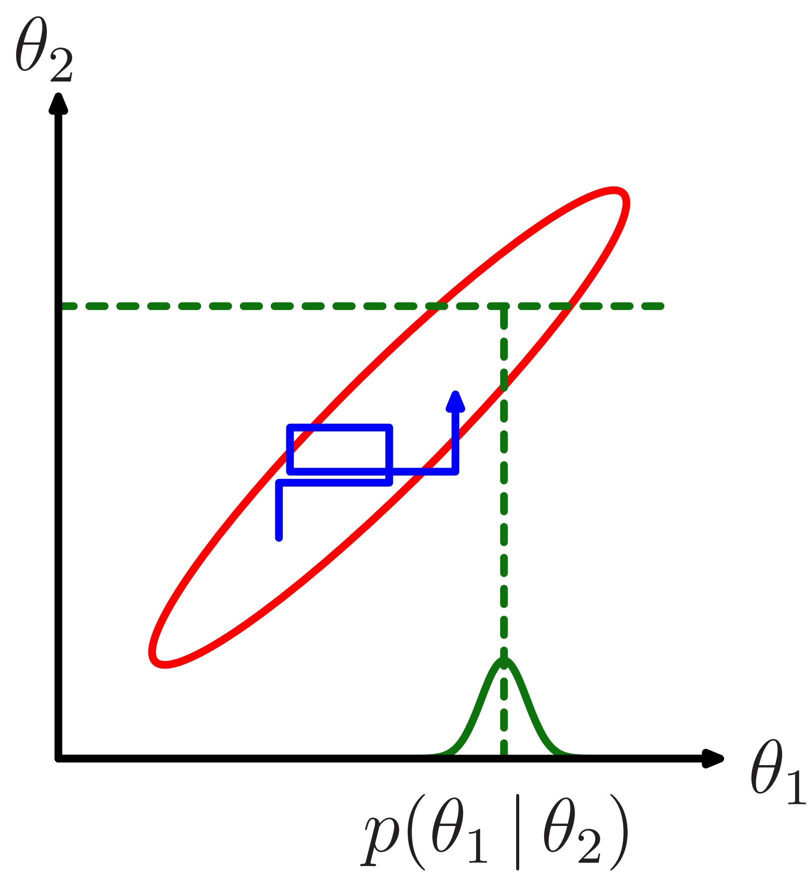
Composition $T_C T_B T_A$ leaves π invariant

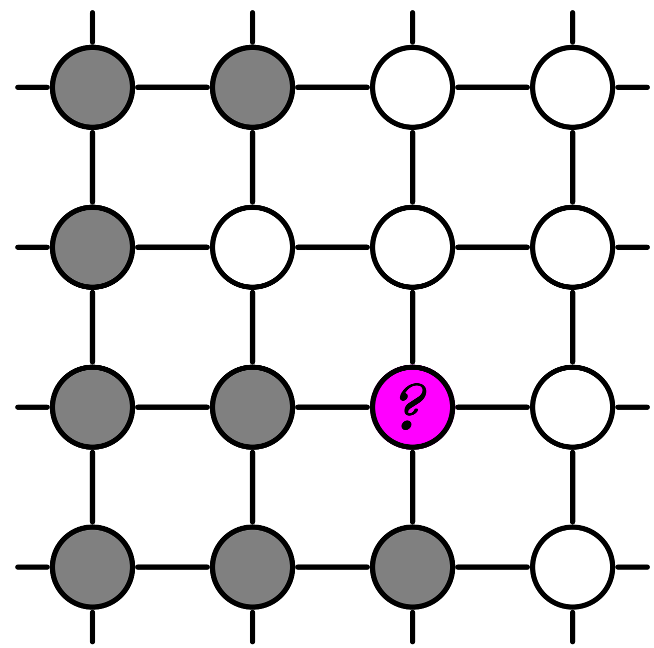
Valid MCMC method if ergodic overall

Gibbs sampling

Pick variables in turn or randomly,

and resample $p(\theta_i \,|\, \theta_{j\neq i})$





$$T_i(\theta' \leftarrow \theta) = p(\theta'_i \,|\, \theta_{j\neq i})\, \delta(\theta'_{j\neq i} - \theta_{j\neq i})$$

LHS adapted from Fig 11.11 Bishop PRML

Gibbs sampling correctness

$$p(\theta) = p(\theta_i \,|\, \theta_{\setminus i})\, p(\theta_{\setminus i})$$

Simulate by drawing $\theta_{\setminus i}$, then $\theta_i \,|\, \theta_{\setminus i}$

Draw $\theta_{\setminus i}$: sample θ, throw initial θ_i away

Blocking / Collapsing

Infer $\theta = (\mathbf{w}, \mathbf{z})$ given $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}$. Model:

$$\mathbf{w} \sim \mathcal{N}(0, I)$$

$$z_n \sim \text{Bernoulli}(0.1)$$

$$y^{(n)} \sim \begin{cases} \mathcal{N}(\mathbf{w}^\top \mathbf{x}^{(n)},\, 0.1^2) & z_n = 0 \\ \mathcal{N}(0,\, 1) & z_n = 1 \end{cases}$$

Block Gibbs: resample $p(\mathbf{w} \,|\, \mathbf{z}, \mathcal{D})$ and $p(\mathbf{z} \,|\, \mathbf{w}, \mathcal{D})$

Collapsing: run MCMC on $p(\mathbf{z} \,|\, \mathcal{D})$ or $p(\mathbf{w} \,|\, \mathcal{D})$

Auxiliary variables

Collapsing: analytically integrate variables out

Auxiliary methods: introduce extra variables;
integrate by MCMC

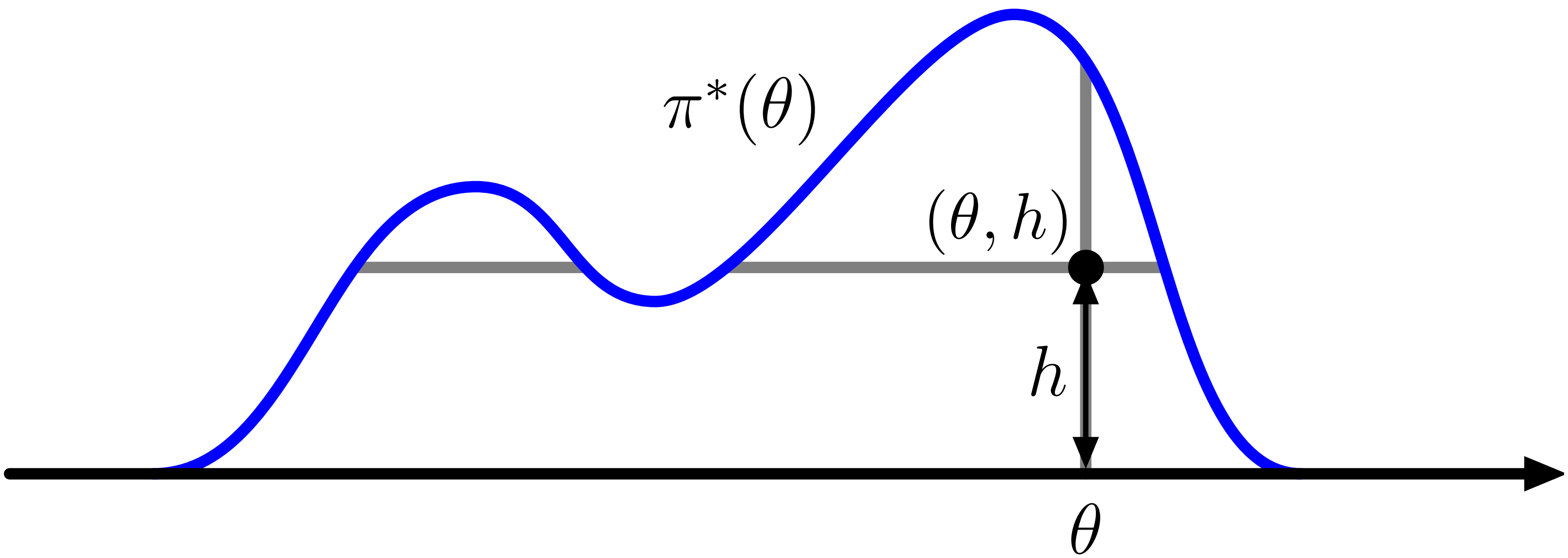
Explore: $\pi(\theta, h)$, where

$$\int \pi(\theta, h)\, \mathrm{d}h = \pi(\theta)$$

Swendsen–Wang, Hamiltonian Monte Carlo (HMC), Slice Sampling, Pseudo-Marginal methods. . .

Slice sampling idea

Sample uniformly under curve $\pi^*(\theta) \propto \pi(\theta)$

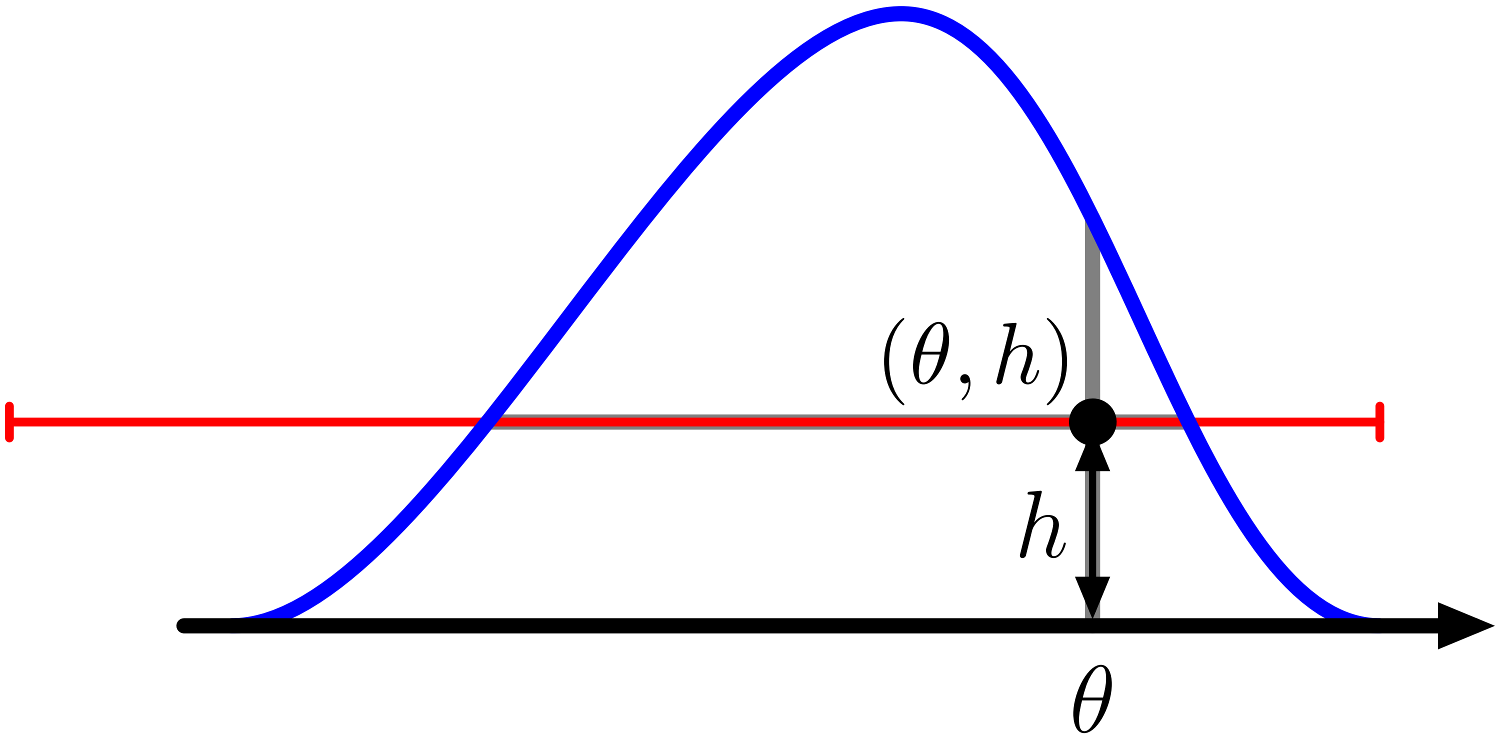


$$p(h \mid \theta) = \text{Uniform}[0, \pi^*(\theta)]$$

$$p(\theta \mid h) \propto \begin{cases} 1 & \pi^*(\theta) \geq h \\ 0 & \text{otherwise} \end{cases} = \text{“Uniform on the slice”}$$

Slice sampling

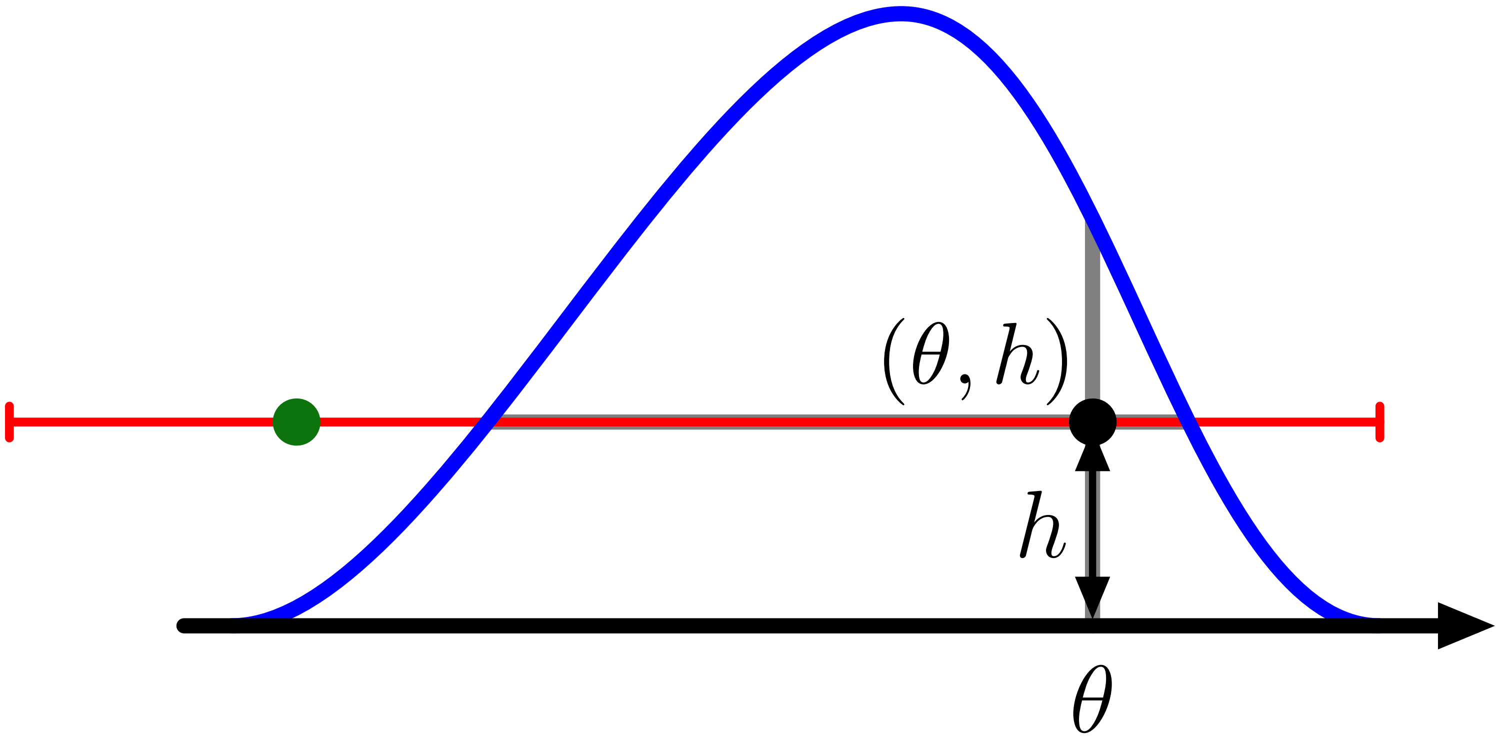
Unimodal conditionals



Rejection sampling $p(\theta \mid h)$ using broader uniform

Slice sampling

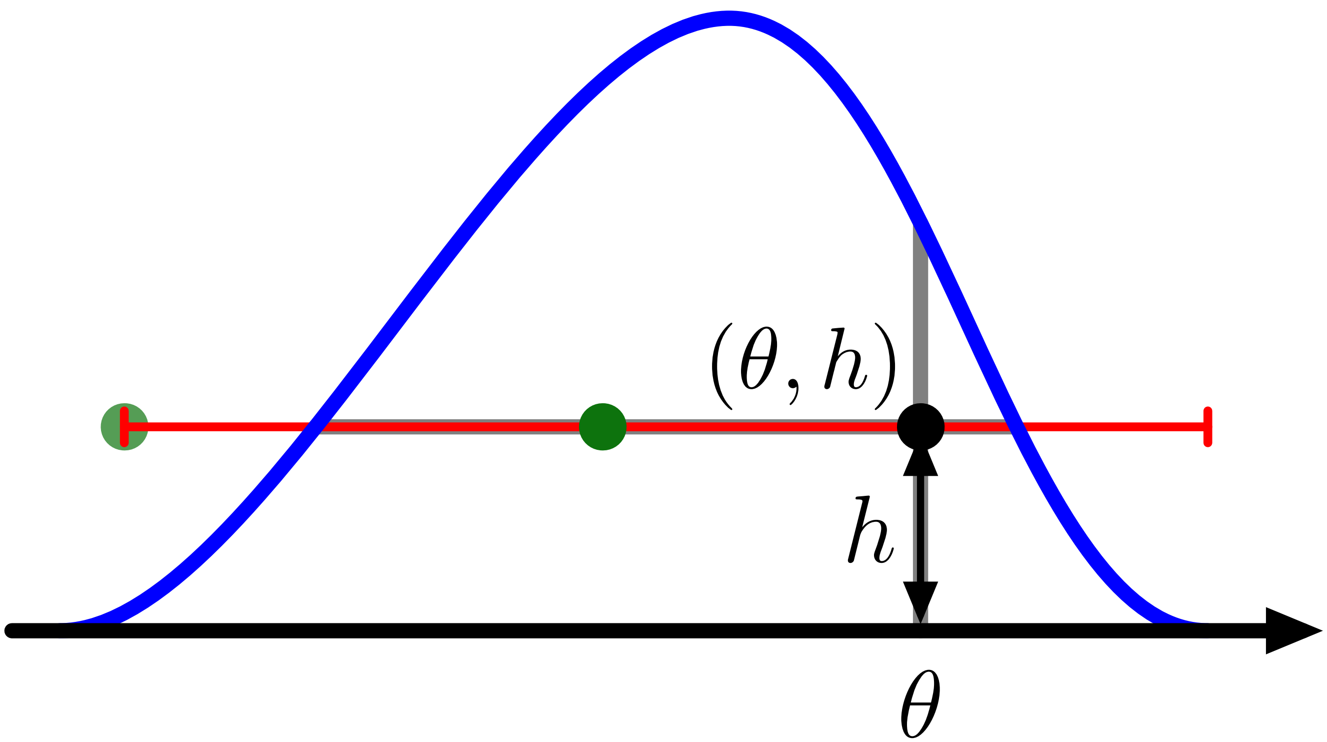
Unimodal conditionals



Adaptive rejection sampling $p(\theta \mid h)$

Slice sampling

Unimodal conditionals

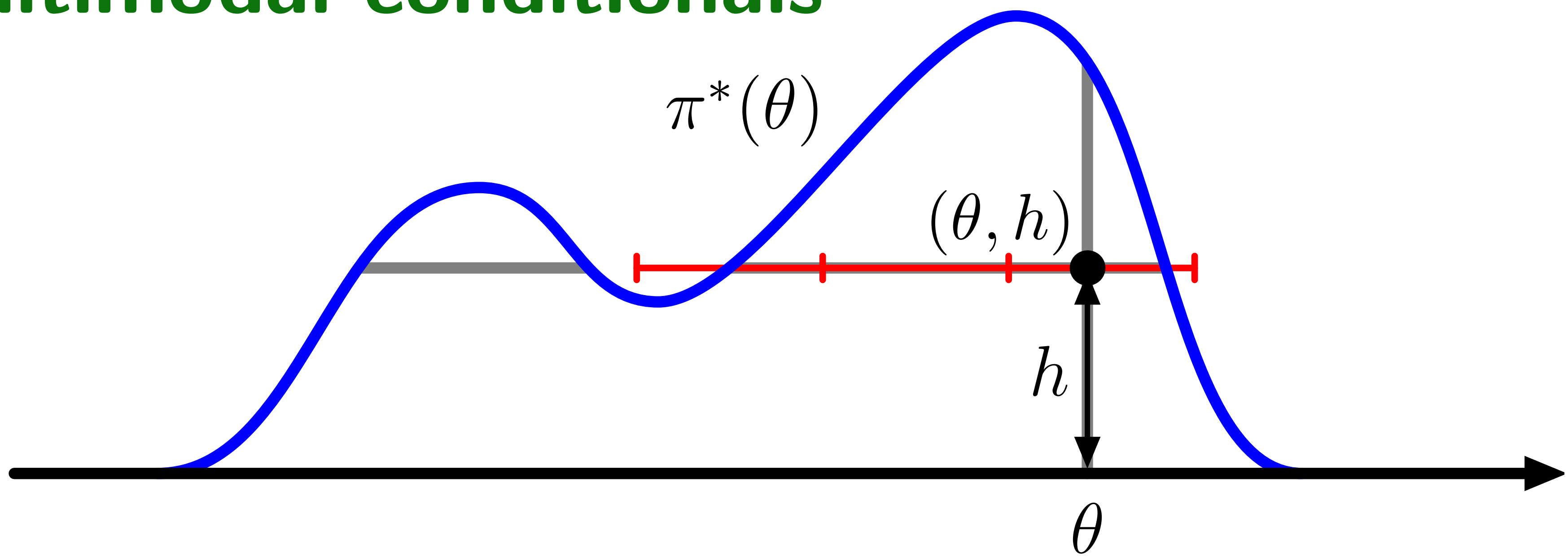


Quickly find new θ

No rejections recorded

Slice sampling

Multimodal conditionals



Use updates that leave $p(\theta \,|\, h)$ invariant:

— place bracket randomly around point
— linearly step out until ends are off slice
— sample on bracket, shrinking as before

Slice sampling

Advantages of slice-sampling:

- Easy — only requires $\pi^*(\theta) \propto \pi(\theta)$
- No rejections
- Step sizes adaptive

Other versions:

Neal (2003): `http://www.cs.toronto.edu/~radford/slice-aos.abstract.html`
Elliptical Slice Sampling: `http://iainmurray.net/pub/10ess/`
Pseudo-Marginal Slice Sampling: `http://arxiv.org/abs/1510.02958`

MCMC: Hybrid Monte Carlo

- Hamiltonian Monte Carlo
  - Probability density : $p(\mathbf{z}) = \frac{1}{Z_p} \exp(-E(\mathbf{z}))$
  - Auxiliary variable r
  - Gradient w.r.t. the variable:

$$\begin{aligned} \widehat{r}_i(\tau + \epsilon/2) &= \widehat{r}_i(\tau) - \frac{\epsilon}{2}\frac{\partial E}{\partial z_i}(\widehat{\mathbf{z}}(\tau)) \\ \widehat{z}_i(\tau + \epsilon) &= \widehat{z}_i(\tau) + \epsilon \widehat{r}_i(\tau + \epsilon/2) \\ \widehat{r}_i(\tau + \epsilon) &= \widehat{r}_i(\tau + \epsilon/2) - \frac{\epsilon}{2}\frac{\partial E}{\partial z_i}(\widehat{\mathbf{z}}(\tau + \epsilon)). \end{aligned}$$

- Metropolis
  - Remove bias associated with the discretization

MCMC: Hamiltonian Monte Carlo

- Hamiltonian Dynamics
  - physicists describe how objects move throughout a system
  - location z and momentum : $r_i = \frac{\mathrm{d}z_i}{\mathrm{d}\tau}$
  - potential energy E(z): $\frac{\mathrm{d}r_i}{\mathrm{d}\tau} = -\frac{\partial E(\mathbf{z})}{\partial z_i}$
  - kinetic energy K(r): $K(\mathbf{r}) = \frac{1}{2}\|\mathbf{r}\|^2 = \frac{1}{2}\sum_i r_i^2$
  - total energy H(z, r): $H(\mathbf{z}, \mathbf{r}) = E(\mathbf{z}) + K(\mathbf{r})$

$$\begin{aligned} \frac{\mathrm{d}z_i}{\mathrm{d}\tau} &= \frac{\partial H}{\partial r_i} \\ \frac{\mathrm{d}r_i}{\mathrm{d}\tau} &= -\frac{\partial H}{\partial z_i} \end{aligned}$$

MCMC: Hamiltonian Monte Carlo

- During the evolution of this dynamical system, the value of the Hamiltonian *H* is constant

$$\begin{aligned}\frac{\mathrm{d}H}{\mathrm{d}\tau} &= \sum_i \left\{ \frac{\partial H}{\partial z_i}\frac{\mathrm{d}z_i}{\mathrm{d}\tau} + \frac{\partial H}{\partial r_i}\frac{\mathrm{d}r_i}{\mathrm{d}\tau} \right\} \\ &= \sum_i \left\{ \frac{\partial H}{\partial z_i}\frac{\partial H}{\partial r_i} - \frac{\partial H}{\partial r_i}\frac{\partial H}{\partial z_i} \right\} = 0\end{aligned}$$

- *Liouville's Theorem:* volume invariant

$$\mathbf{V} = \left(\frac{\mathrm{d}\mathbf{z}}{\mathrm{d}\tau}, \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}\tau}\right)$$

$$\begin{aligned}\operatorname{div}\mathbf{V} &= \sum_i \left\{ \frac{\partial}{\partial z_i}\frac{\mathrm{d}z_i}{\mathrm{d}\tau} + \frac{\partial}{\partial r_i}\frac{\mathrm{d}r_i}{\mathrm{d}\tau} \right\} \\ &= \sum_i \left\{ -\frac{\partial}{\partial z_i}\frac{\partial H}{\partial r_i} + \frac{\partial}{\partial r_i}\frac{\partial H}{\partial z_i} \right\} = 0\end{aligned}$$

MCMC: Hybrid Monte Carlo

- *leapfrog* discretization

$$\frac{\mathrm{d}z_i}{\mathrm{d}\tau} = \frac{\partial H}{\partial r_i} \qquad \frac{\mathrm{d}r_i}{\mathrm{d}\tau} = -\frac{\partial H}{\partial z_i}$$

$$\Rightarrow$$

$$\begin{aligned}
\widehat{r}_i(\tau + \epsilon/2) &= \widehat{r}_i(\tau) - \frac{\epsilon}{2}\frac{\partial E}{\partial z_i}(\widehat{\mathbf{z}}(\tau)) \\
\widehat{z}_i(\tau + \epsilon) &= \widehat{z}_i(\tau) + \epsilon\widehat{r}_i(\tau + \epsilon/2) \\
\widehat{r}_i(\tau + \epsilon) &= \widehat{r}_i(\tau + \epsilon/2) - \frac{\epsilon}{2}\frac{\partial E}{\partial z_i}(\widehat{\mathbf{z}}(\tau + \epsilon))
\end{aligned}$$

- accepted or rejected according to the Metropolis

$$\min\left(1, \exp\{H(\mathbf{z}, \mathbf{r}) - H(\mathbf{z}^\star, \mathbf{r}^\star)\}\right)$$

MCMC: Hybrid Monte Carlo

1. Initialise $x^{(0)}$.
2. For $i = 0$ to $N-1$
   - Sample $v \sim \mathcal{U}_{[0,1]}$ and $u^\star \sim \mathcal{N}(0, I_{n_x})$.
   - Let $x_0 = x^{(i)}$ and $u_0 = u^\star + \rho\Delta(x_0)/2$.
   - For $l = 1, \ldots, L$, take steps

$$x_l = x_{l-1} + \rho u_{l-1}$$
$$u_l = u_{l-1} + \rho_l \Delta(x_l)$$

     where $\rho_l = \rho$ for $l < L$ and $\rho_L = \rho/2$.
   - If $v < \mathcal{A} = \min\left\{1, \frac{p(x_L)}{p(x^{(i)})} \exp\left(-\frac{1}{2}(u_L^{\mathrm{T}} u_L - u^{\star\mathrm{T}} u^\star)\right)\right\}$

$$(x^{(i+1)}, u^{(i+1)}) = (x_L, u_L)$$

     else

$$(x^{(i+1)}, u^{(i+1)}) = (x^{(i)}, u^\star)$$

Stochastic Gradient Hamiltonian Monte Carlo

- Stochastic gradient using mini-batch $\tilde{\mathcal{D}}$

$$\nabla \tilde{U}(\theta) = -\frac{|\mathcal{D}|}{|\tilde{\mathcal{D}}|} \sum_{x \in \tilde{\mathcal{D}}} \nabla \log p(x|\theta) - \nabla \log p(\theta), \ \tilde{\mathcal{D}} \subset \mathcal{D}$$

- Noisy gradient: $\nabla \tilde{U}(\theta) \approx \nabla U(\theta) + \mathcal{N}(0, V(\theta))$

- SDE: $\begin{cases} d\theta = & M^{-1} r \, dt \\ dr = & -\nabla U(\theta) \, dt + \mathcal{N}(0, 2B(\theta) dt) \end{cases}$ $\quad B(\theta) = \frac{1}{2}\epsilon V(\theta)$

- Friction: $\begin{cases} d\theta = M^{-1} r \, dt \\ dr = -\nabla U(\theta) \, dt \boxed{- B M^{-1} r dt} + \mathcal{N}(0, 2B dt) \end{cases}$

Summary

Write down the probability of everything $p(\mathcal{D}, \theta)$

Condition on data, $\mathcal{D}$,

explore unknowns θ by MCMC

Samples give plausible explanations

— Look at them

— Average their predictions

Which method?

Simulate / sample with known distribution:
Exact samples, rejection sampling

Posterior distribution, small, noisy problem
Importance sampling

Posterior distribution, interesting problem
Start with MCMC
Slice sampling, M-H if careful, Gibbs if clever
Hamiltonian methods, HMC, uses gradients

Reference

- https://youtu.be/0l31GVOXflk
- An Introduction to MCMC for Machine Learning, Machine Learning, 2003
- Pattern Recognition and Machine Learning, Chapter 11