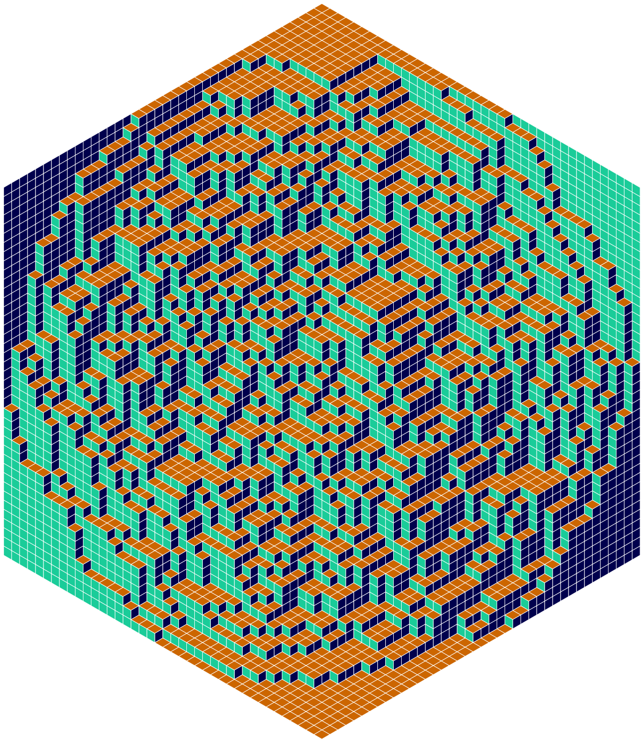
Markov chain Monte Carlo

Probabilistic Models of Cognition, 2011
http://www.ipam.ucla.edu/programs/gss2011/

Roadmap:
— Motivation
— Monte Carlo basics
— What is MCMC?
— Metropolis–Hastings and Gibbs
— …more tomorrow.

Iain Murray
http://homepages.inf.ed.ac.uk/imurray2/

Eye-balling samples

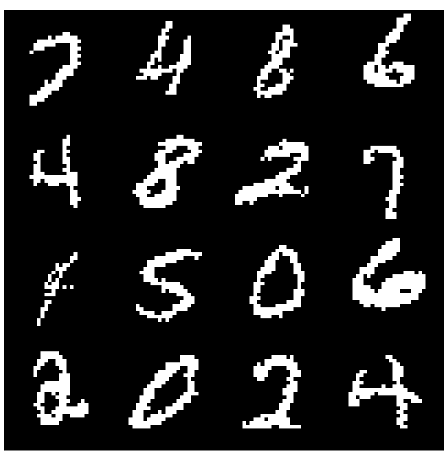


Sometimes samples are pleasing to look at:

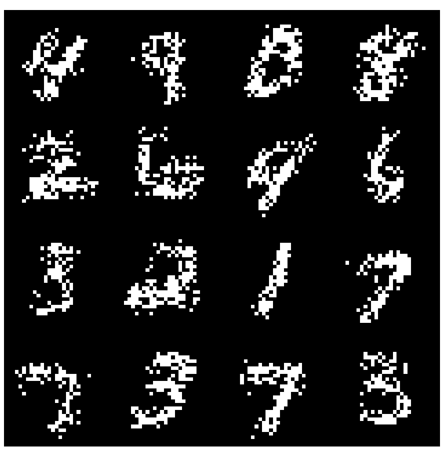
(if you're into geometrical combinatorics)

Figure by Propp and Wilson. Source: MacKay textbook.

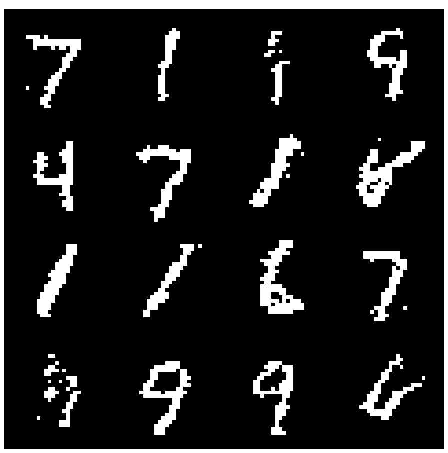
Sanity check probabilistic modeling assumptions:



Data samples



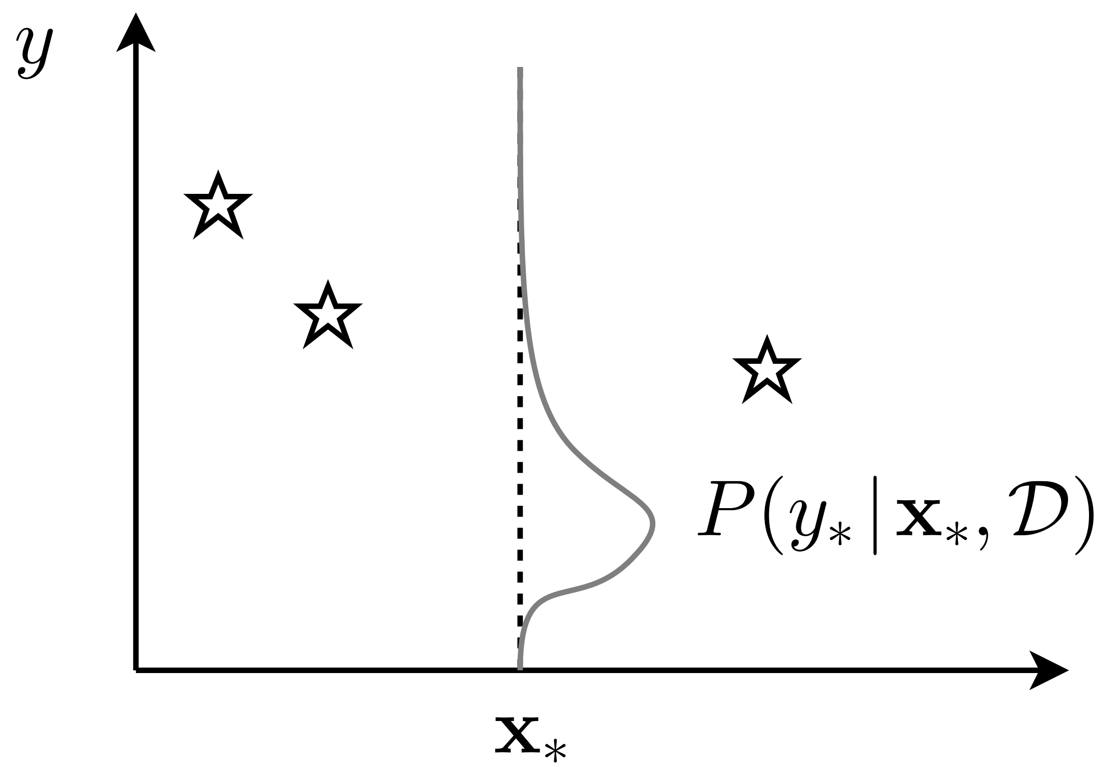
MoB samples



RBM samples

The need for integrals

$$
\begin{aligned}
P(y_* \,|\, \mathbf{x}_*, \mathcal{D}) &= \int \mathrm{d}\theta \; P(y_*, \theta \,|\, \mathbf{x}_*, \mathcal{D}) \\
&= \int \mathrm{d}\theta \; P(y_* \,|\, \theta, \cancel{\mathcal{D}}) \; P(\theta \,|\, \cancel{\mathbf{x}_*}, \mathcal{D})
\end{aligned}
$$



A statistical problem

What is the average height of the GSS2011 lecturers?

Method: measure their heights, add them up and divide by $N \approx 25$.

What is the average height f of people p in California $\mathcal{C}$?

$$E_{p \in \mathcal{C}}[f(p)] \equiv \frac{1}{|\mathcal{C}|} \sum_{p \in \mathcal{C}} f(p), \quad \text{“intractable”?}$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} f\left(p^{(s)}\right), \quad \text{for random survey of } S \text{ people } \{p^{(s)}\} \in \mathcal{C}$$

Surveying works for large and notionally infinite populations.

Simple Monte Carlo

Statistical sampling can be applied to any expectation:

In general:

$$\int f(x) P(x) \, \mathrm{d}x \approx \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)}), \quad x^{(s)} \sim P(x)$$

Example: making predictions

$$\begin{aligned} p(x|\mathcal{D}) &= \int P(x|\theta, \mathcal{D}) \, P(\theta|\mathcal{D}) \, \mathrm{d}\theta \\ &\approx \frac{1}{S} \sum_{s=1}^{S} P(x|\theta^{(s)}, \mathcal{D}), \quad \theta^{(s)} \sim P(\theta|\mathcal{D}) \end{aligned}$$

More examples: E-step statistics in EM, Boltzmann machine learning

Properties of Monte Carlo

Estimator: $$\int f(x)\, P(x)\, \mathrm{d}x \;\approx\; \hat{f} \;\equiv\; \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), \quad x^{(s)} \sim P(x)$$

Estimator is unbiased:

$$\mathbb{E}_{P(\{x^{(s)}\})}\left[\hat{f}\right] \;=\; \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{P(x)}[f(x)] \;=\; \mathbb{E}_{P(x)}[f(x)]$$

Variance shrinks $\propto 1/S$:

$$\mathrm{var}_{P(\{x^{(s)}\})}\left[\hat{f}\right] \;=\; \frac{1}{S^2}\sum_{s=1}^{S}\mathrm{var}_{P(x)}[f(x)] \;=\; \mathrm{var}_{P(x)}[f(x)]\,/S$$

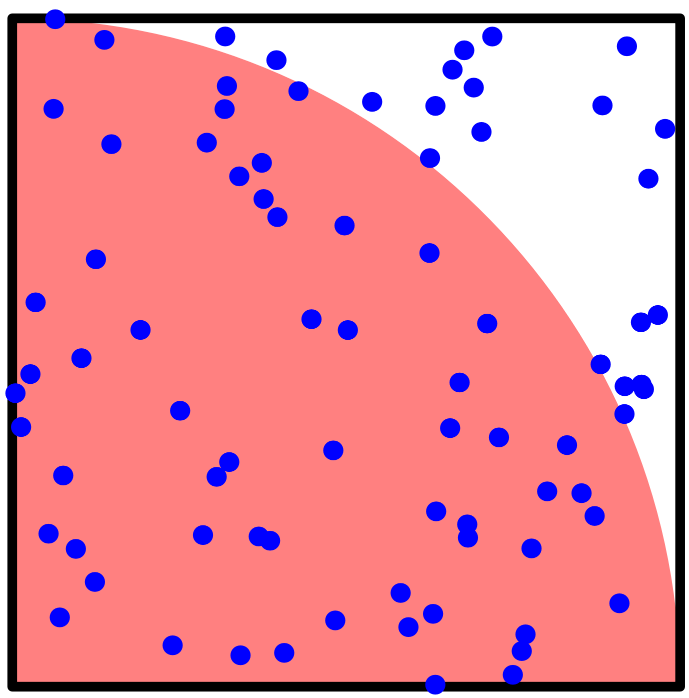
"Error bars" shrink like $\sqrt{S}$

Aside: don't always sample!

"Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse."

— Alan Sokal, 1996

A dumb approximation of π



$$P(x,y) = \begin{cases} 1 & 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4 \iint \mathbb{I}\left((x^2+y^2)<1\right) P(x,y)\, \mathrm{d}x\, \mathrm{d}y$$

```
octave:1> S=12; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.3333
octave:2> S=1e7; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.1418
```

Alternatives to Monte Carlo

There are other methods of numerical integration!

Example: (nice) 1D integrals are easy:

```
octave:1> 4 * quadl(@(x) sqrt(1-x.^2), 0, 1, tolerance)
```

Gives π to 6 dp's in 108 evaluations, machine precision in 2598.

(NB Matlab's `quadl` fails at `tolerance=0`, but Octave works.)

In higher dimensions sometimes determinstic approximations work: Variational Bayes, EP, INLA, . . .

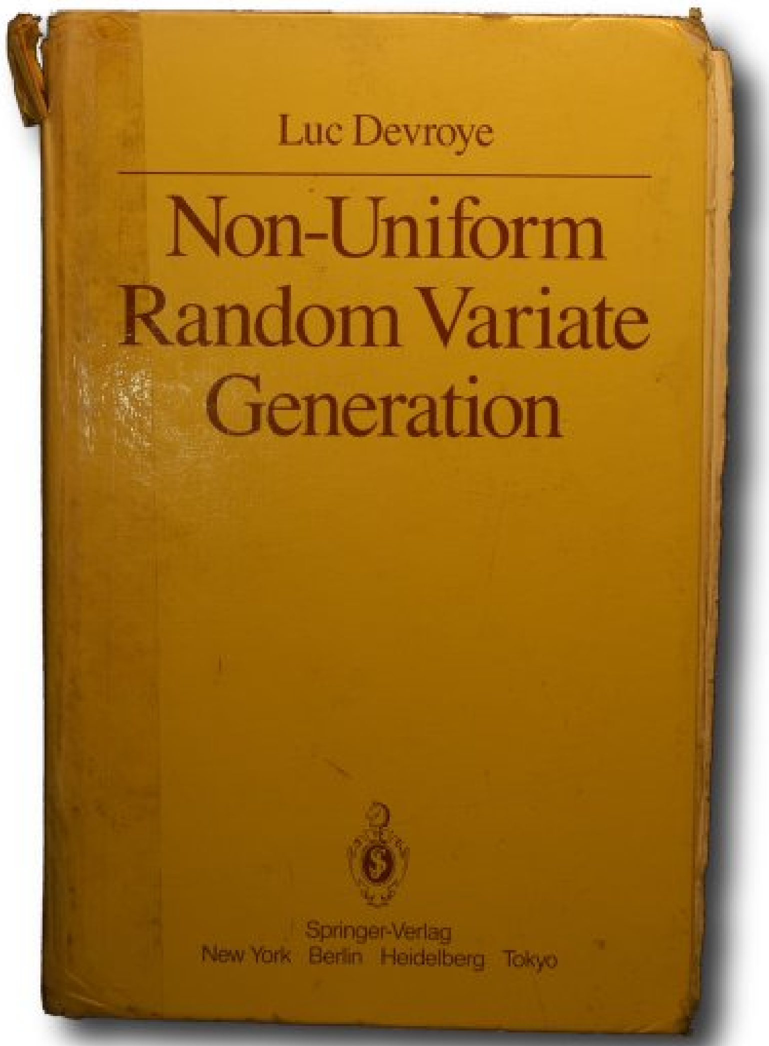
Reminder

Want to sample to approximate expectations:

$$\int f(x)P(x)\,\mathrm{d}x \approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), \quad x^{(s)} \sim P(x)$$

How do we get the samples?

Sampling simple distributions



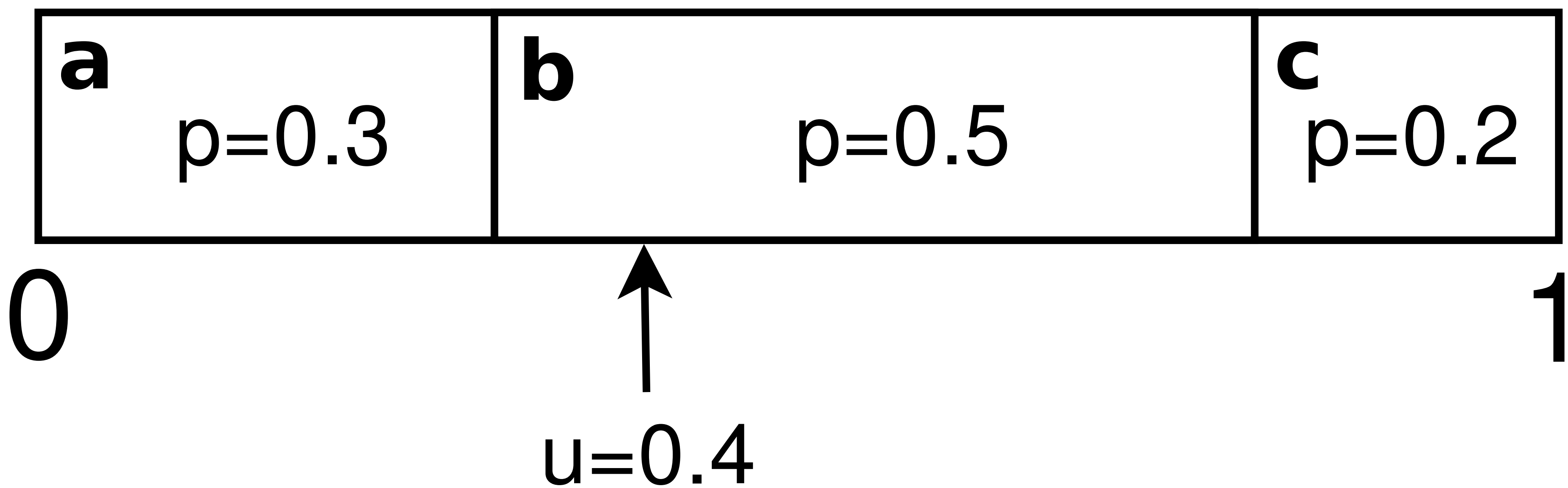
Use library routines for univariate distributions

(and some other special cases)

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

`http://cg.scs.carleton.ca/~luc/rnbookindex.html`

Sampling discrete values



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

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

There are more efficient ways for large numbers of values and samples. See Devroye book.

Sampling from densities

How to convert samples from a Uniform[0,1] generator:

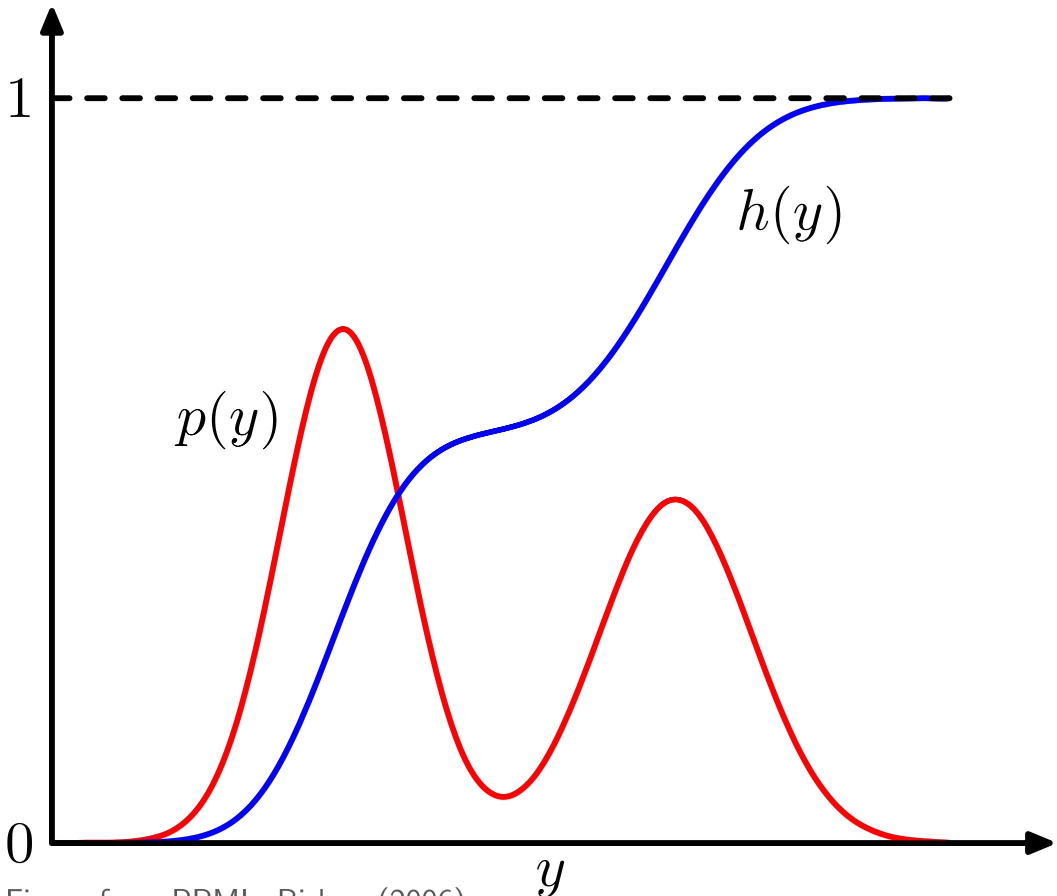


Figure from PRML, Bishop (2006)

$$h(y) = \int_{-\infty}^{y} p(y')\,\mathrm{d}y'$$

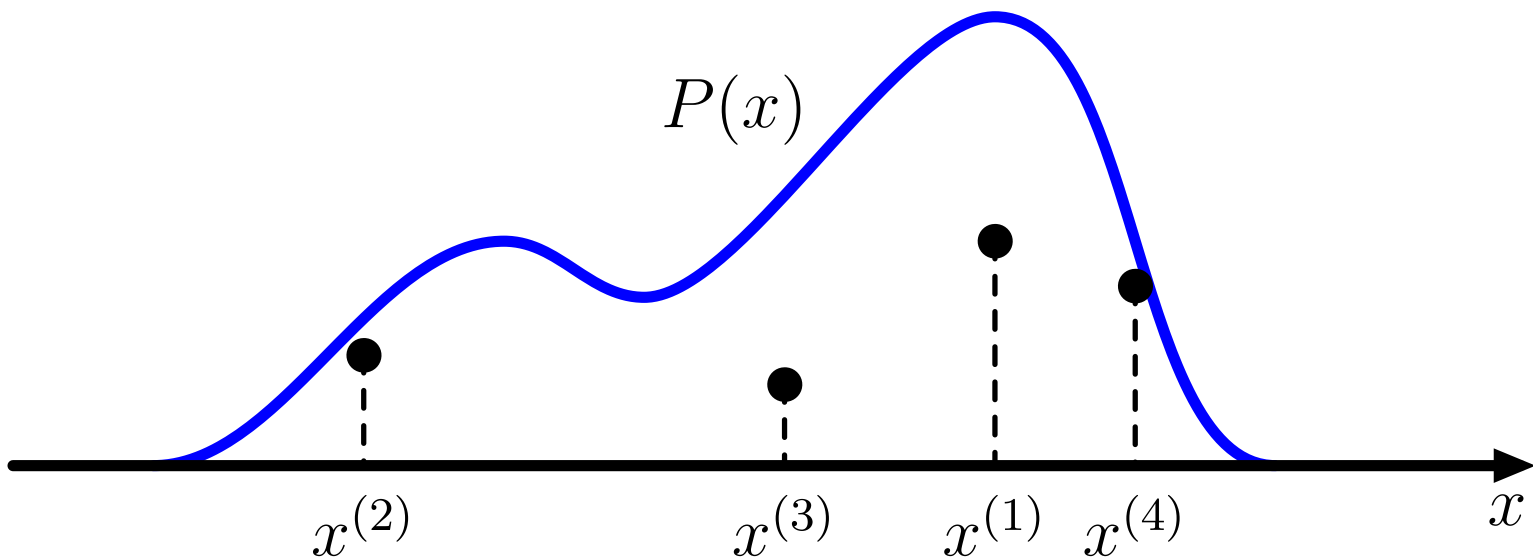
$u \sim$ Uniform[0,1]

Sample, $y(u) = h^{-1}(u)$

Although we can't always compute and invert $h(y)$

Sampling from densities

Draw points uniformly under the curve:



Probability mass to left of point $\sim$ Uniform[0,1]

Rejection sampling

Sampling from $\pi(x)$ using tractable $q(x)$:

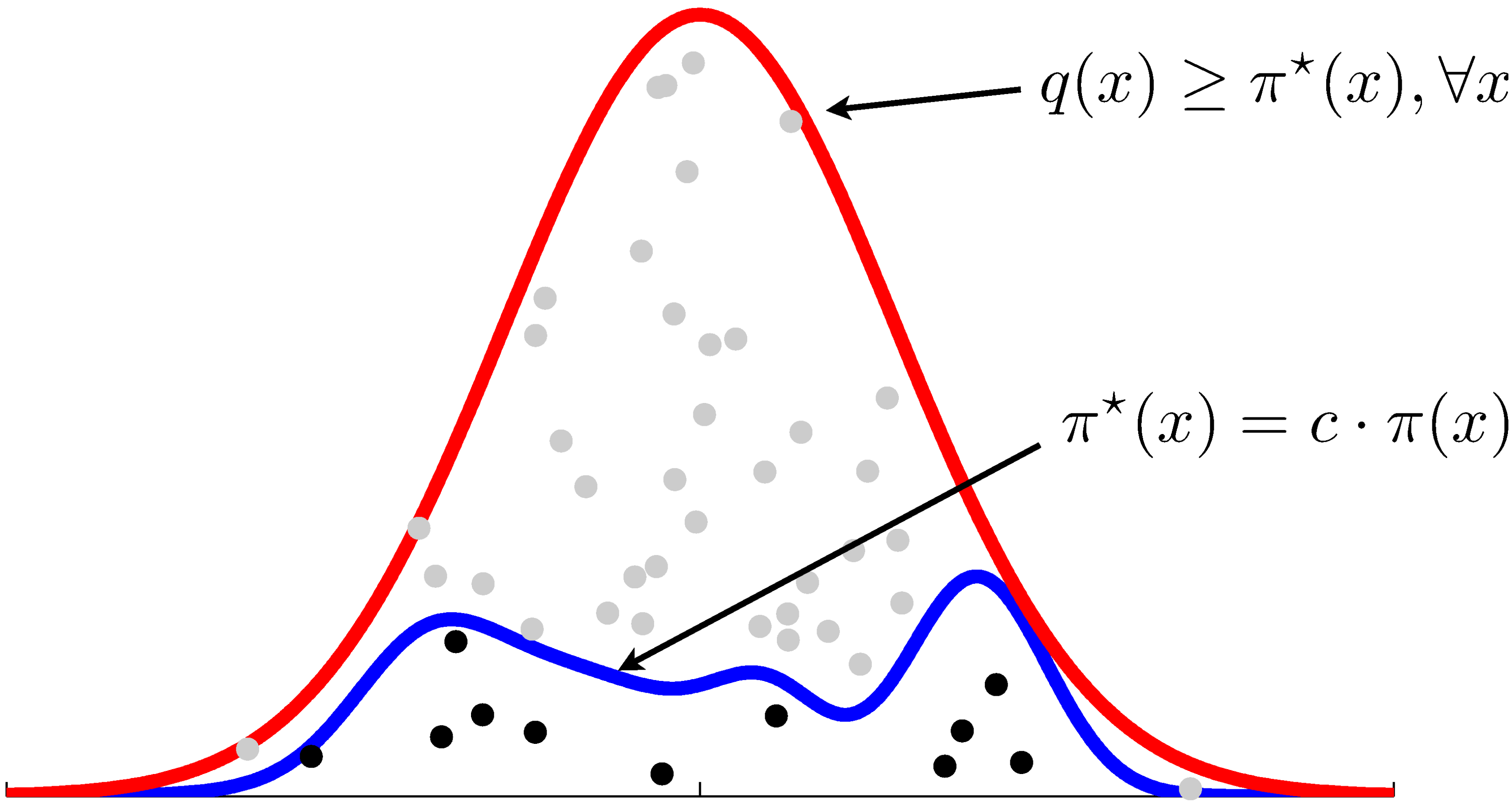


Figure credit: Ryan P. Adams

Importance sampling

Throwing away samples seems wasteful

Instead rewrite the integral as an expectation under Q:

$$\int f(x)\, P(x)\, \mathrm{d}x \;=\; \int f(x)\, \frac{P(x)}{Q(x)} Q(x)\, \mathrm{d}x, \qquad (Q(x) > 0 \text{ if } P(x) > 0)$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)})\, \frac{P(x^{(s)})}{Q(x^{(s)})}, \quad x^{(s)} \sim Q(x)$$

This is just simple Monte Carlo again, so it is unbiased.

Importance sampling applies when the integral is not an expectation.

Divide and multiply any integrand by a convenient distribution.

Importance sampling (2)

Previous slide assumed we could evaluate $P(x) = \tilde{P}(x)/\mathcal{Z}_P$

$$\int f(x)\,P(x)\,\mathrm{d}x \;\approx\; \frac{\mathcal{Z}_Q}{\mathcal{Z}_P}\,\frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\underbrace{\frac{\tilde{P}(x^{(s)})}{\tilde{Q}(x^{(s)})}}_{\tilde{r}^{(s)}}, \quad x^{(s)} \sim Q(x)$$

$$\approx\; \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\,\frac{\tilde{r}^{(s)}}{\frac{1}{S}\sum_{s'}\tilde{r}^{(s')}} \;\equiv\; \sum_{s=1}^{S} f(x^{(s)})\,w^{(s)}$$

This estimator is **consistent** but **biased**

Exercise: Prove that $\mathcal{Z}_P/\mathcal{Z}_Q \approx \frac{1}{S}\sum_s \tilde{r}^{(s)}$

Summary so far

- Sums and integrals, often expectations, occur frequently in statistics
- **Monte Carlo** approximates expectations with a sample average
- **Rejection sampling** draws samples from complex distributions
- **Importance sampling** applies Monte Carlo to 'any' sum/integral

Next: Why are we not done? MCMC, Metropolis–Hastings and Gibbs

Reminder

Need to sample large, non-standard distributions:

$$P(x \,|\, \mathcal{D}) \;\approx\; \frac{1}{S}\sum_{s=1}^{S} P(x \,|\, \theta), \qquad \theta \sim P(\theta \,|\, \mathcal{D})$$

When there are nuisance parameters:

$$P(\theta \,|\, \mathcal{D}) \;=\; \int \mathrm{d}\alpha \; P(\theta, \alpha \,|\, \mathcal{D})$$

$$\theta, \alpha \;\sim\; P(\theta, \alpha \,|\, \mathcal{D}) \;\propto\; P(\alpha)\, P(\theta \,|\, \alpha)\, P(\mathcal{D} \,|\, \theta)$$

and discard α's

Application to large problems

Rejection & importance sampling scale badly with dimensionality

Example:

$$P(x) = \mathcal{N}(0, \mathbb{I}), \quad Q(x) = \mathcal{N}(0, \sigma^2 \mathbb{I})$$

Rejection sampling:

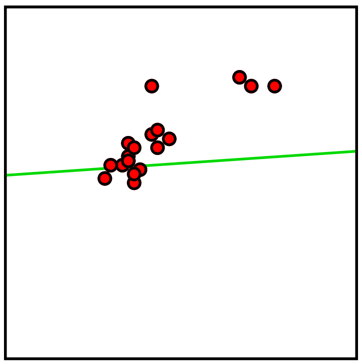
Requires $\sigma \geq 1$. Fraction of proposals accepted $= \sigma^{-D}$

Importance sampling:

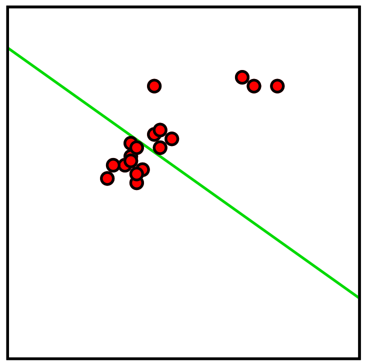
$\text{Var}[P(x)/Q(x)] \;=\; \left(\frac{\sigma^2}{2-1/\sigma^2}\right)^{D/2} - 1$

Infinite / undefined variance if $\sigma \leq 1/\sqrt{2}$

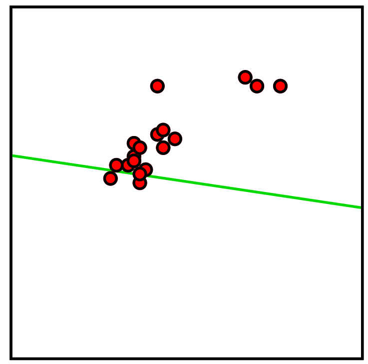
Importance sampling weights



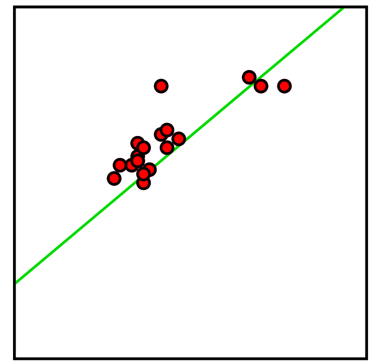
$w = $0.00548



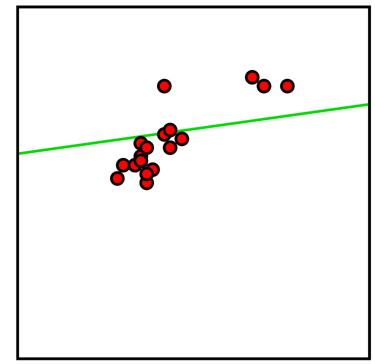
$w = $1.59e-08



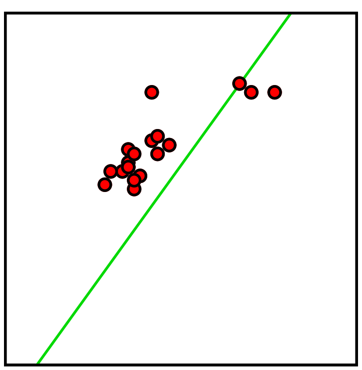
$w = $9.65e-06



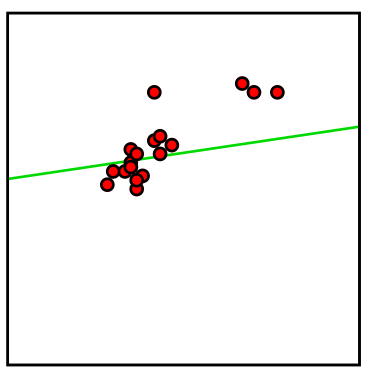
$w = $0.371



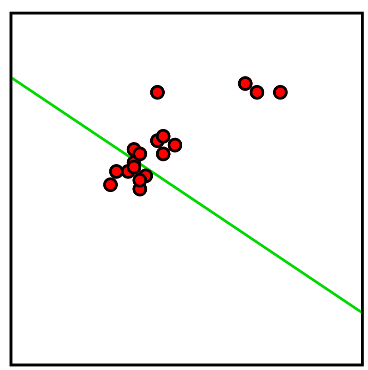
$w = $0.103



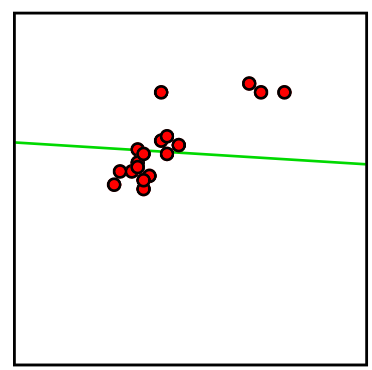
$w = $1.01e-08



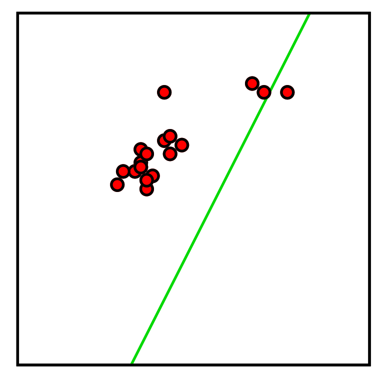
$w = $0.111



$w = $1.92e-09

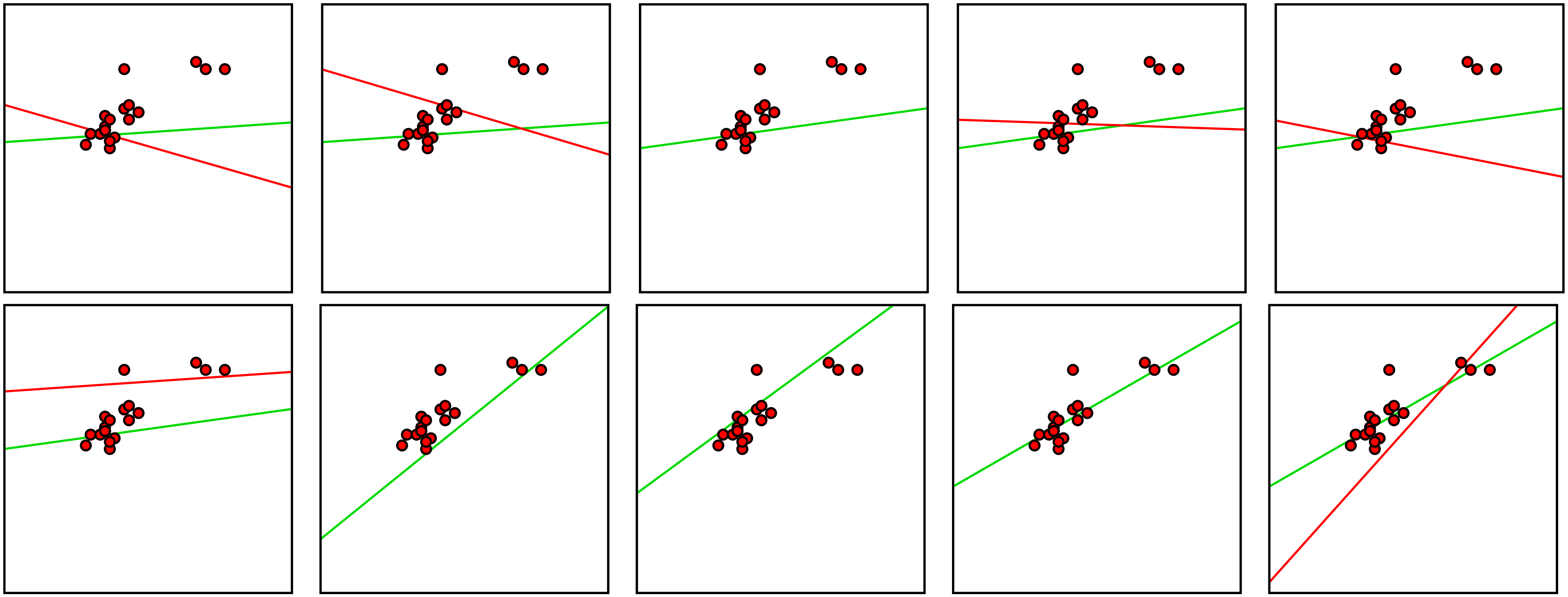


$w = $0.0126

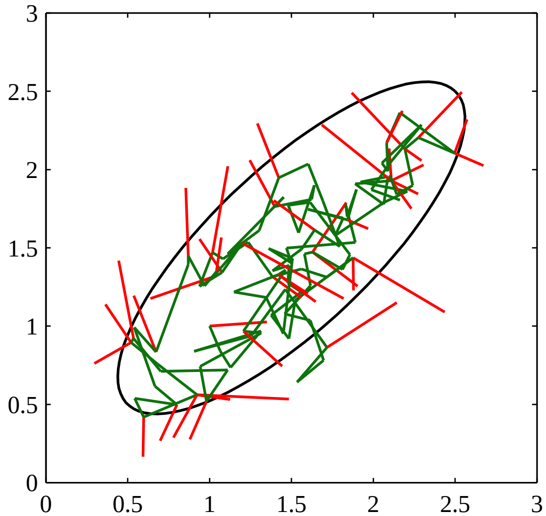


$w = $1.1e-51

Metropolis algorithm



- Perturb parameters: $Q(\theta'; \theta)$, e.g. $\mathcal{N}(\theta, \sigma^2)$
- Accept with probability $\min\left(1, \frac{\tilde{P}(\theta'|\mathcal{D})}{\tilde{P}(\theta|\mathcal{D})}\right)$
- Otherwise **keep old parameters**



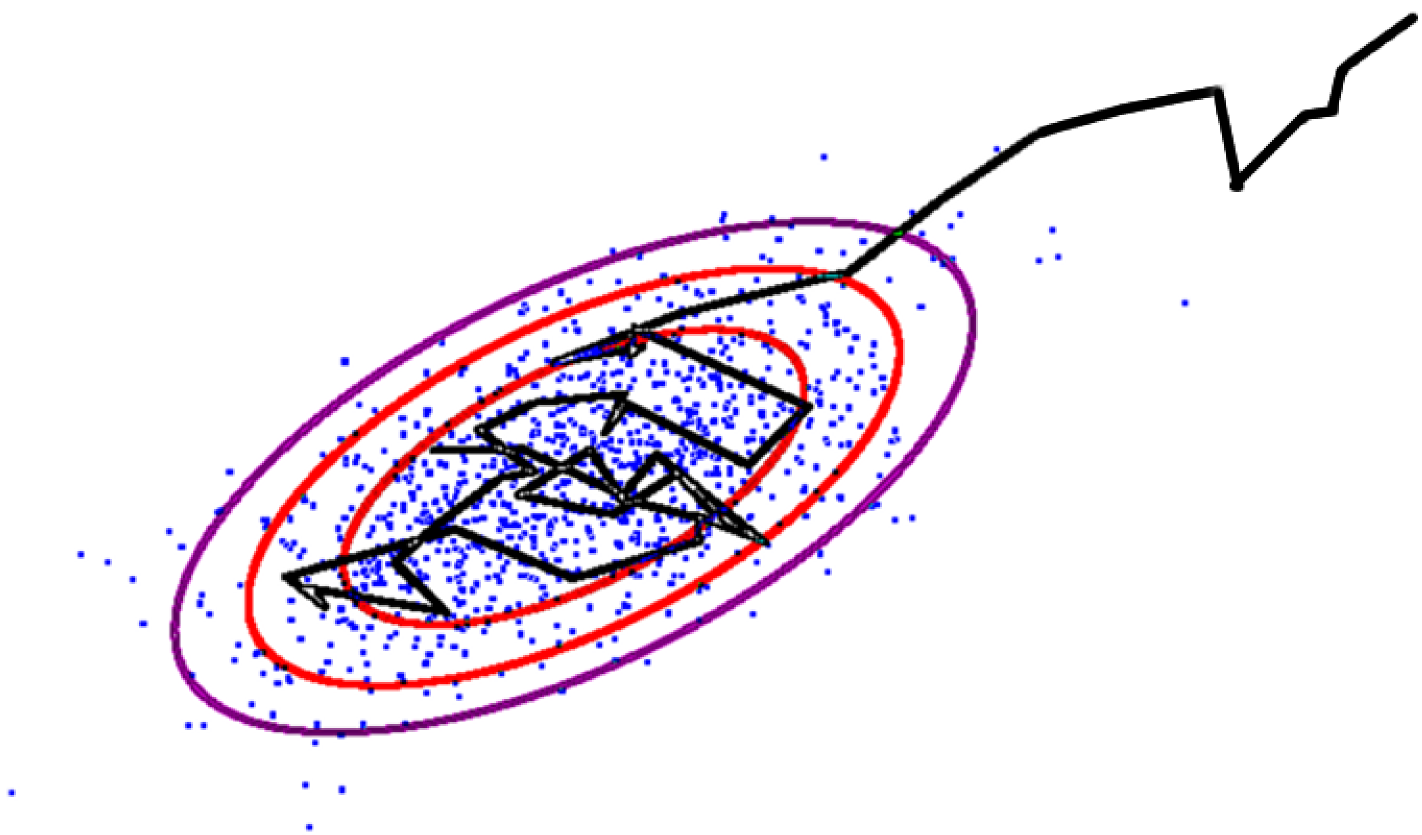
This subfigure from PRML, Bishop (2006)

Detail: Metropolis, as stated, requires $Q(\theta'; \theta) = Q(\theta; \theta')$

Markov chain Monte Carlo

Construct a biased random walk that explores target dist $P^\star(x)$

Markov steps, $x_t \sim T(x_t \leftarrow x_{t-1})$



MCMC gives approximate, correlated samples from $P^\star(x)$

Transition operators

Discrete example

$$P^\star = \begin{pmatrix} 3/5 \\ 1/5 \\ 1/5 \end{pmatrix} \qquad T = \begin{pmatrix} 2/3 & 1/2 & 1/2 \\ 1/6 & 0 & 1/2 \\ 1/6 & 1/2 & 0 \end{pmatrix} \qquad T_{ij} = T(x_i \leftarrow x_j)$$

$P^\star$ is an **invariant distribution** of T because $TP^\star = P^\star$, i.e.

$$\sum_x T(x' \leftarrow x) P^\star(x) = P^\star(x')$$

Also $P^\star$ is *the* **equilibrium distribution** of T:

To machine precision: $T^{100}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 3/5 \\ 1/5 \\ 1/5 \end{pmatrix} = P^\star$

Ergodicity requires: $T^K(x' \leftarrow x) > 0$ for all $x' : P^\star(x') > 0$, for some K

Reverse operators

If T leaves $P^\star(x)$ stationary, we can define a *reverse operator*

$$R(x \leftarrow x') \propto T(x' \leftarrow x)\, P^\star(x) = \frac{T(x' \leftarrow x)\, P^\star(x)}{\sum_x T(x' \leftarrow x)\, P^\star(x)} = \frac{T(x' \leftarrow x)\, P^\star(x)}{P^\star(x')}$$

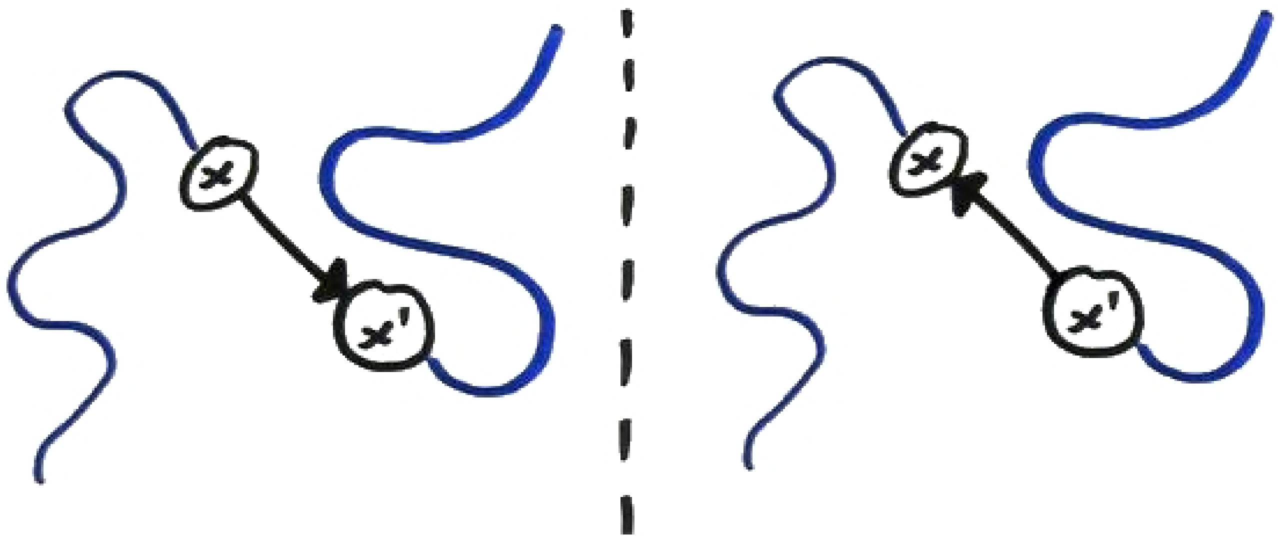
A necessary (and sufficient) condition: there exists R such that:

$$T(x' \leftarrow x)\, P^\star(x) \;=\; R(x \leftarrow x')\, P^\star(x'), \qquad \forall x, x'$$

If $R = T$, operator satisfies **detailed balance** (not necessary)

Balance condition

$\rightarrow x \rightarrow x'$ and $\rightarrow x' \rightarrow x$ are equally probable:



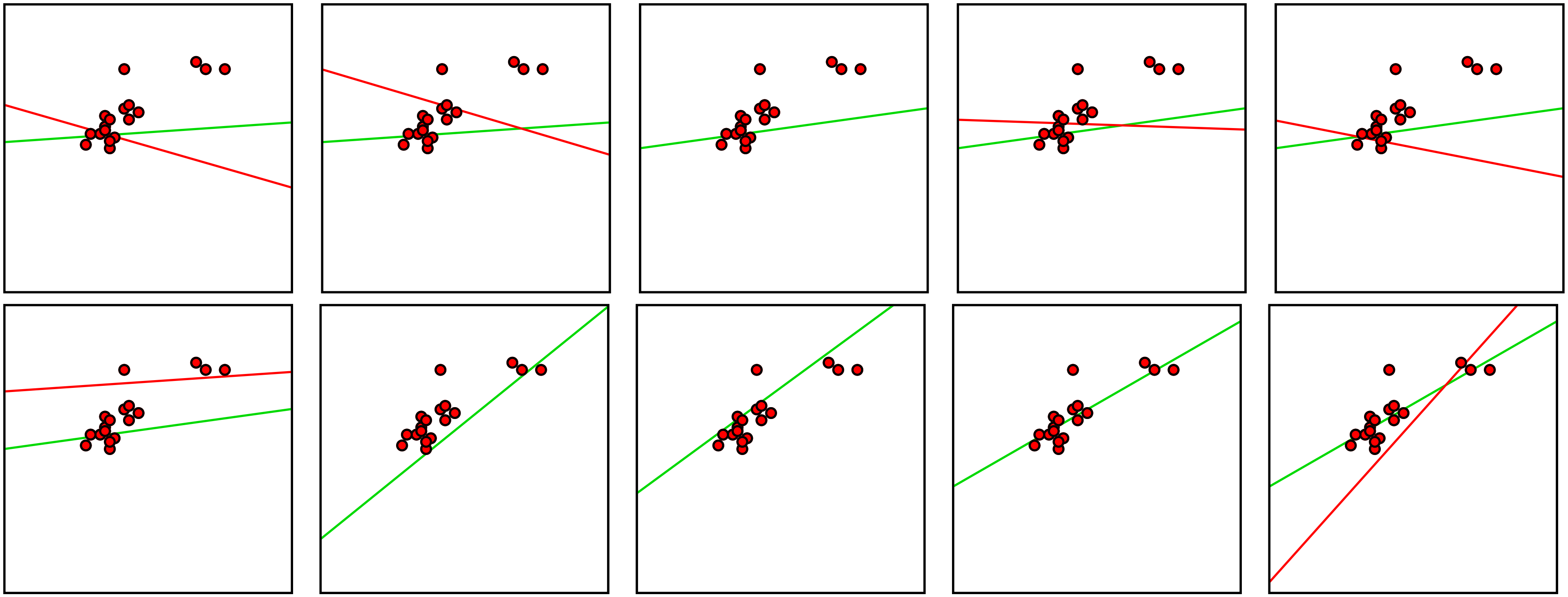
$$T(x' \leftarrow x)\, P^\star(x) \;=\; R(x \leftarrow x')\, P^\star(x')$$

Implies that $P^\star(x)$ is left invariant:

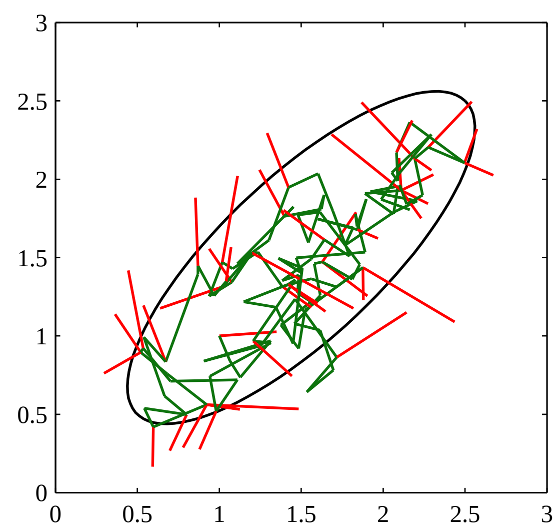
$$\sum_x T(x' \leftarrow x)\, P^\star(x) \;=\; P^\star(x') \sum_x R(x \leftarrow x')^{\nearrow 1}$$

Enforcing the condition is easy: it only involves isolated pairs

Metropolis algorithm



- Perturb parameters: $Q(\theta'; \theta)$, e.g. $\mathcal{N}(\theta, \sigma^2)$
- Accept with probability $\min\left(1, \dfrac{\tilde{P}(\theta'|\mathcal{D})}{\tilde{P}(\theta|\mathcal{D})}\right)$
- Otherwise **keep old parameters**



This subfigure from PRML, Bishop (2006)

Detail: Metropolis, as stated, requires $Q(\theta'; \theta) = Q(\theta; \theta')$

Metropolis–Hastings

Transition operator

- Propose a move from the current state $Q(x'; x)$, e.g. $\mathcal{N}(x, \sigma^2)$
- Accept with probability $\min\!\left(1, \frac{P(x')Q(x;x')}{P(x)Q(x';x)}\right)$
- Otherwise next state in chain is a copy of current state

Notes

- Can use $\tilde{P} \propto P(x)$; normalizer cancels in acceptance ratio
- Satisfies detailed balance (shown below)
- Q must be chosen so chain is ergodic

$$
\begin{aligned}
P(x) \cdot T(x' \leftarrow x) = P(x) \cdot Q(x'; x) \min\left(1, \frac{P(x')Q(x; x')}{P(x)Q(x'; x)}\right) &= \min\Big(P(x)Q(x'; x),\ P(x')Q(x; x')\Big) \\
&= P(x') \cdot Q(x; x') \min\left(1, \frac{P(x)Q(x'; x)}{P(x')Q(x; x')}\right) = P(x') \cdot T(x \leftarrow x')
\end{aligned}
$$

Matlab/Octave code for demo

```
function samples = dumb_metropolis(init, log_ptilde, iters, sigma)

D = numel(init);
samples = zeros(D, iters);

state = init;
Lp_state = log_ptilde(state);
for ss = 1:iters
    % Propose
    prop = state + sigma*randn(size(state));
    Lp_prop = log_ptilde(prop);
    if log(rand) < (Lp_prop - Lp_state)
        % Accept
        state = prop;
        Lp_state = Lp_prop;
    end
    samples(:, ss) = state(:);
end
```

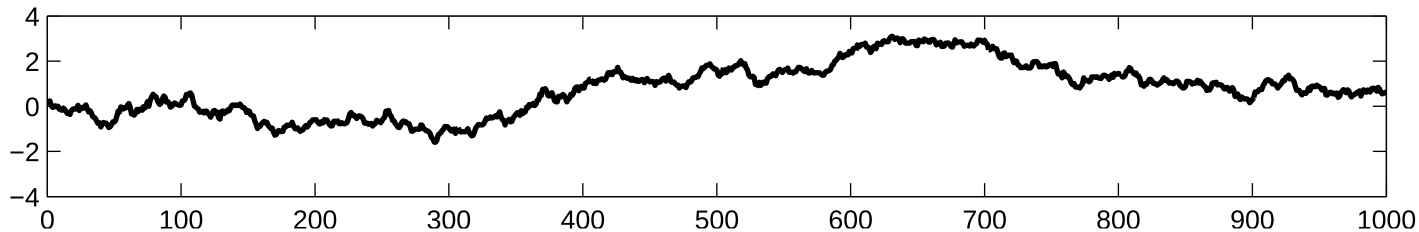
Step-size demo

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

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

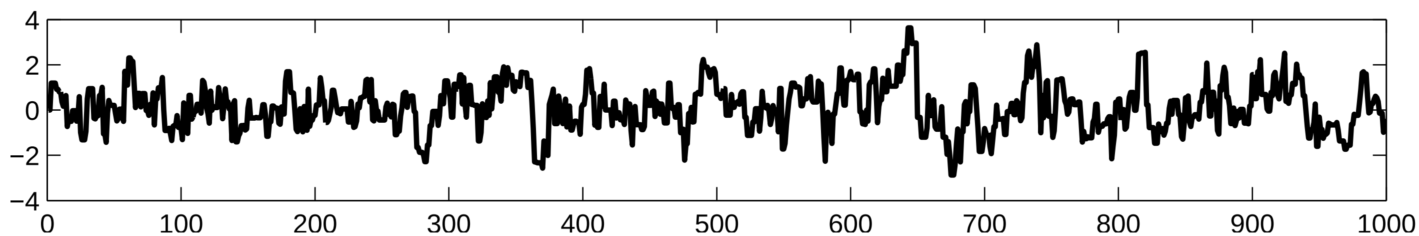
`sigma(0.1)`

99.8% accepts



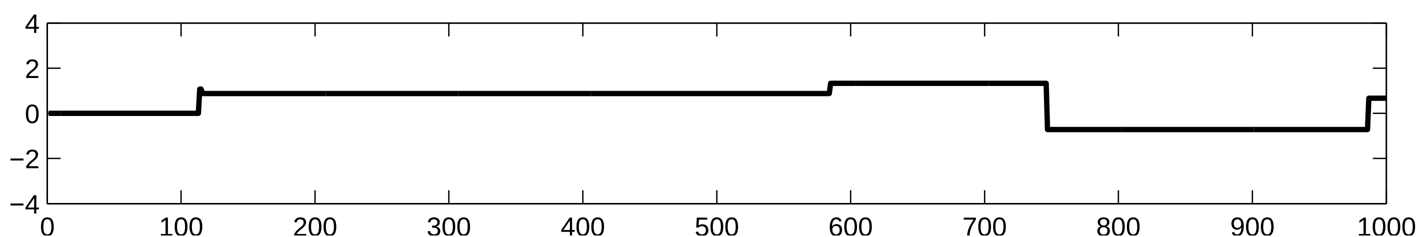
`sigma(1)`

68.4% accepts



`sigma(100)`

0.5% accepts



Gibbs sampling

A method with no rejections:

– Initialize $\mathbf{x}$ to some value
– Pick each variable in turn or randomly and resample $P(x_i|\mathbf{x}_{j\neq i})$

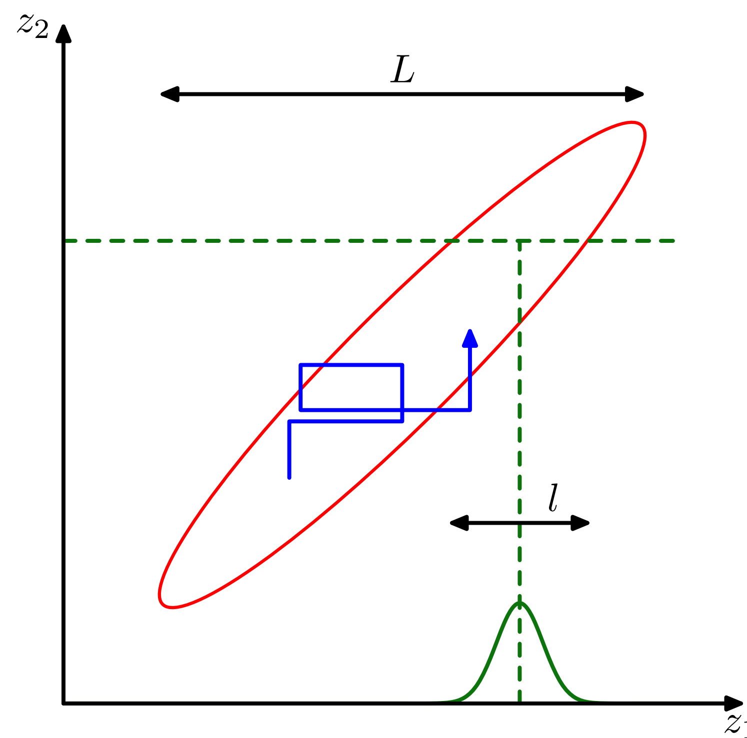


Figure from PRML, Bishop (2006)

Proof of validity: **a)** check detailed balance for component update.
b) Metropolis–Hastings 'proposals' $P(x_i|\mathbf{x}_{j\neq i}) \Rightarrow$ accept with prob. 1
Apply a series of these operators. Don't need to check acceptance.

Gibbs sampling

Alternative explanation:

Chain is currently at $\mathbf{x}$

At equilibrium can assume $\mathbf{x} \sim P(\mathbf{x})$

Consistent with $\mathbf{x}_{j\neq i} \sim P(\mathbf{x}_{j\neq i}), \quad x_i \sim P(x_i \,|\, \mathbf{x}_{j\neq i})$

Pretend x_i was never sampled and do it again.

This view may be useful later for non-parametric applications

Summary so far

- We need approximate methods to solve sums/integrals
- Monte Carlo does not explicitly depend on dimension, although simple methods work only in low dimensions
- Markov chain Monte Carlo (MCMC) can make local moves. By assuming less, it's more applicable to higher dimensions
- simple computations $\Rightarrow$ "easy" to implement (harder to diagnose).