Advanced MCMC methods

http://learning.eng.cam.ac.uk/zoubin/tutorials06.html

Department of Engineering, University of Cambridge
Michaelmas 2006

Iain Murray

i.murray+ta@gatsby.ucl.ac.uk

Probabilistic modelling

Data $\mathcal{D}$, model $\mathcal{M}$; what do we know about x?

Bayesian prediction with unknown parameters θ:

$$P(x|\mathcal{D},\mathcal{M}) = \int P(x,\theta|\mathcal{D},\mathcal{M})\,\mathrm{d}\theta \quad \textbf{Marginalization}$$

$$= \int P(x|\theta,\mathcal{D},\mathcal{M})\underbrace{P(\theta|\mathcal{D},\mathcal{M})}_{\text{from } \textbf{Bayes' rule}}\,\mathrm{d}\theta \quad \textbf{Product rule}$$

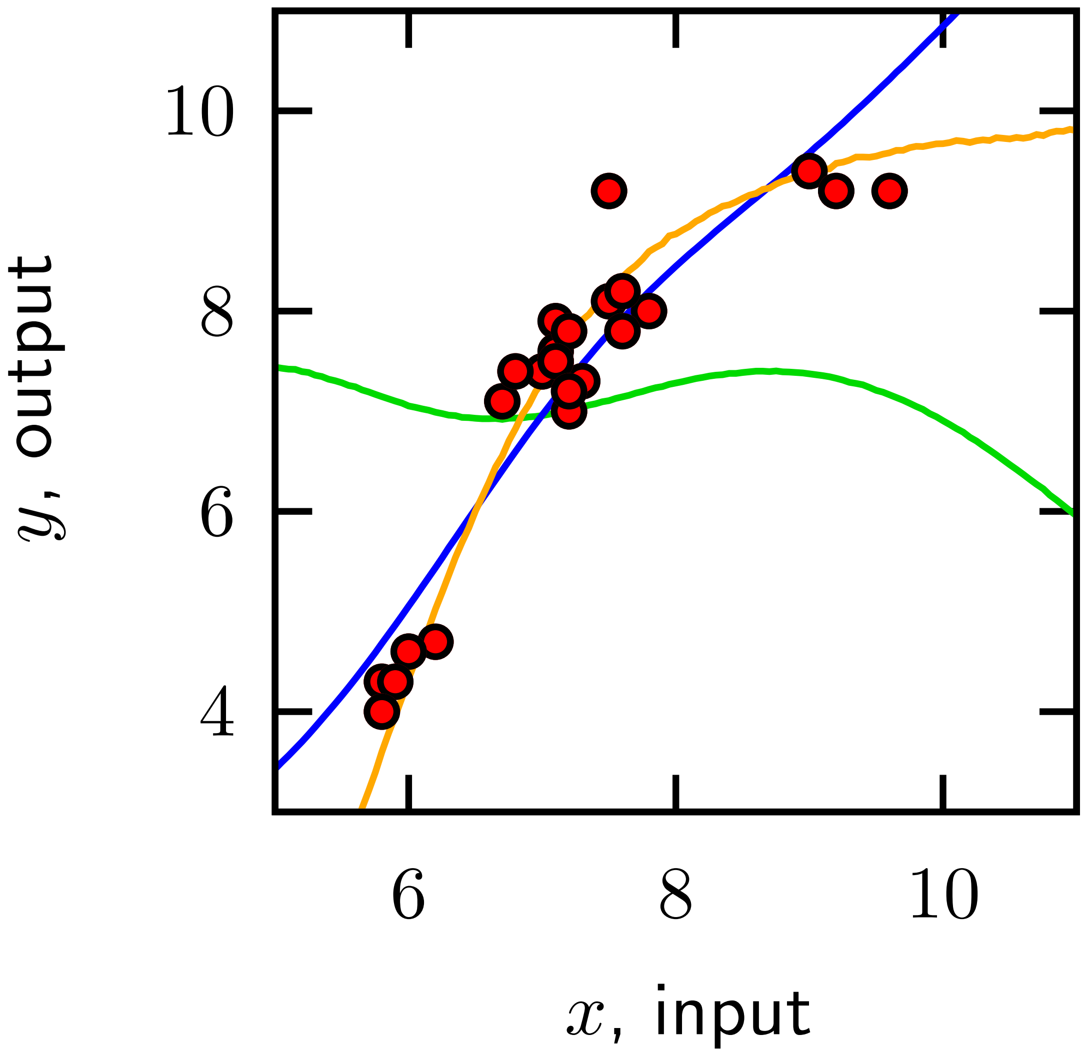
$$= \int \sum_h P(x,h|\theta,\mathcal{D},\mathcal{M}) \sum_H P(\theta,H|\mathcal{D},\mathcal{M})\,\mathrm{d}\theta$$

$$\ldots$$

Inference is the mechanical use of probability theory. . .

. . . provided we can do all the sums and integrals

Example: regression



Non-linear regression

Many parameters:
$\theta = \{\text{curve}, \text{noise}\}$

$$P(y|x, \mathcal{D}, \mathcal{M}) = \int P(y|x, \theta, \mathcal{M}) P(\theta|\mathcal{D}, \mathcal{M}) \, \mathrm{d}\theta$$

Looks tractable? Things are made complicated by
hyper-parameters & complex noise models (logistic $\rightarrow$ classification)

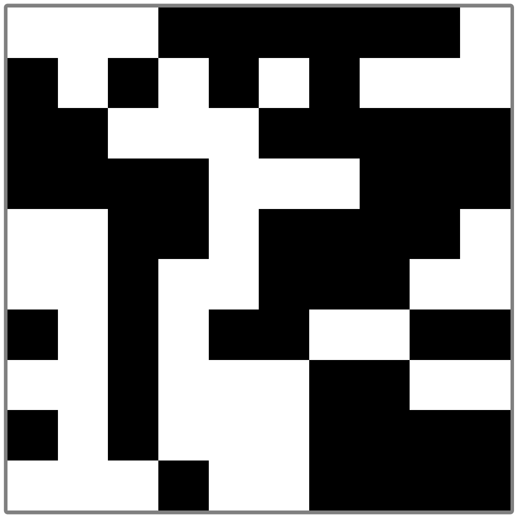
Example: binary latents

100 binary variables $x_i \in \{0, 1\}$, could be:

– a *tiny* patch of pixel labels in computer vision

– assignments to outlier/ordinary of 100 data points

– or a *tiny* patch of idealized magnetic iron



There are 2^{100} possible states

The age of the universe $\approx 2^{98}$ picoseconds

Sum might decompose (e.g. belief propagation)
. . . otherwise must approximate

Even if your 10×10 patch is tractable 100×100 is probably not

Example: topic modelling

Topic 77 word	prob.	Topic 82 word	prob.	Topic 166 word	prob.
MUSIC	.090	LITERATURE	.031	PLAY	.136
DANCE	.034	POEM	.028	BALL	.129
SONG	.033	POETRY	.027	GAME	.065
PLAY	.030	POET	.020	PLAYING	.042
SING	.026	PLAYS	.019	HIT	.032
SINGING	.026	POEMS	.019	PLAYED	.031
BAND	.026	PLAY	.015	BASEBALL	.027
PLAYED	.023	LITERARY	.013	GAMES	.025
SANG	.022	WRITERS	.013	BAT	.019
SONGS	.021	DRAMA	.012	RUN	.019
DANCING	.020	WROTE	.012	THROW	.016
PIANO	.017	POETS	.011	BALLS	.015
PLAYING	.016	WRITER	.011	TENNIS	.011
RHYTHM	.015	SHAKESPEARE	.010	HOME	.010
ALBERT	.013	WRITTEN	.009	CATCH	.010
MUSICAL	.013	STAGE	.009	FIELD	.010

parents^{035} hoped^{268} he might consider^{118} becoming a concert^{077} pianist^{077}. But bix was interested^{268} in another kind^{050} of music^{077}. He wanted^{268} to play^{077} the cornet. And he wanted^{268} to play^{077} jazz^{077}...

playhouses, the playhouses must have the right audiences^{082}. We must remember^{288} that plays^{082} exist^{143} to be performed^{077}, not merely^{050} to be read^{254}. (even when you read^{254} a play^{082} to yourself, try^{288} to perform^{062} it, to put^{174} it on a stage^{078}, as you go along.) as soon^{028} as a play^{082} has to be performed^{082}, then some

game^{166} book^{254}. The boys^{020} see a game^{166} for two. The two boys^{020} play^{166} the game^{166}. The boys^{020} play^{166} the game^{166} for two. The boys^{020} like the game^{166}. Meg^{282} comes^{040} into the house^{282}. Meg^{282} and don^{180} and jim^{296} read^{254} the book^{254}. They see a game^{166} for three. Meg^{282} and don^{180} and jim^{296} play^{166} the game^{166}. They play^{166}...

Adapted from Steyvers and Griffiths (2006)

Simple Monte Carlo

Integration

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

Making predictions

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

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

Unbiased, variance $\sim 1/S$

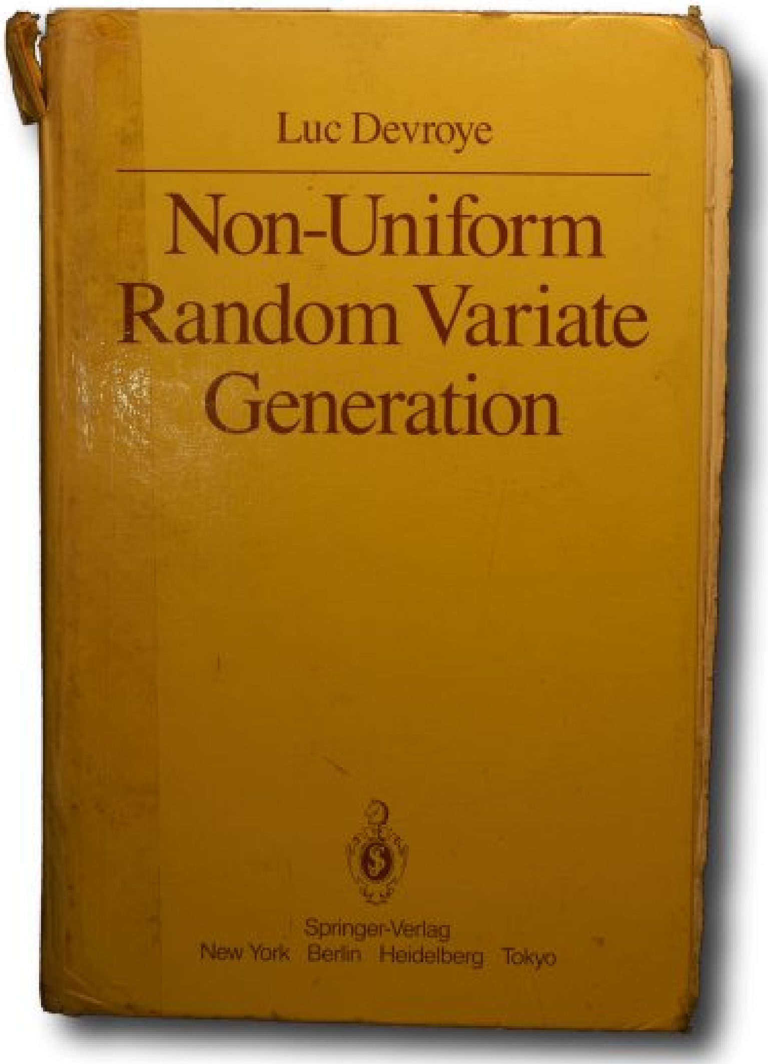
When to sample?

EP versus Monte Carlo

- Monte Carlo is general but expensive
  - A sledgehammer
- EP exploits underlying simplicity of the problem (if it exists)
- Monte Carlo is still needed for complex problems (e.g. large isolated peaks)
- Trick is to know what problem you have

Stolen from Tom Minka's slides from last week

How to sample?



For **univariate distributions** (and some other special cases)

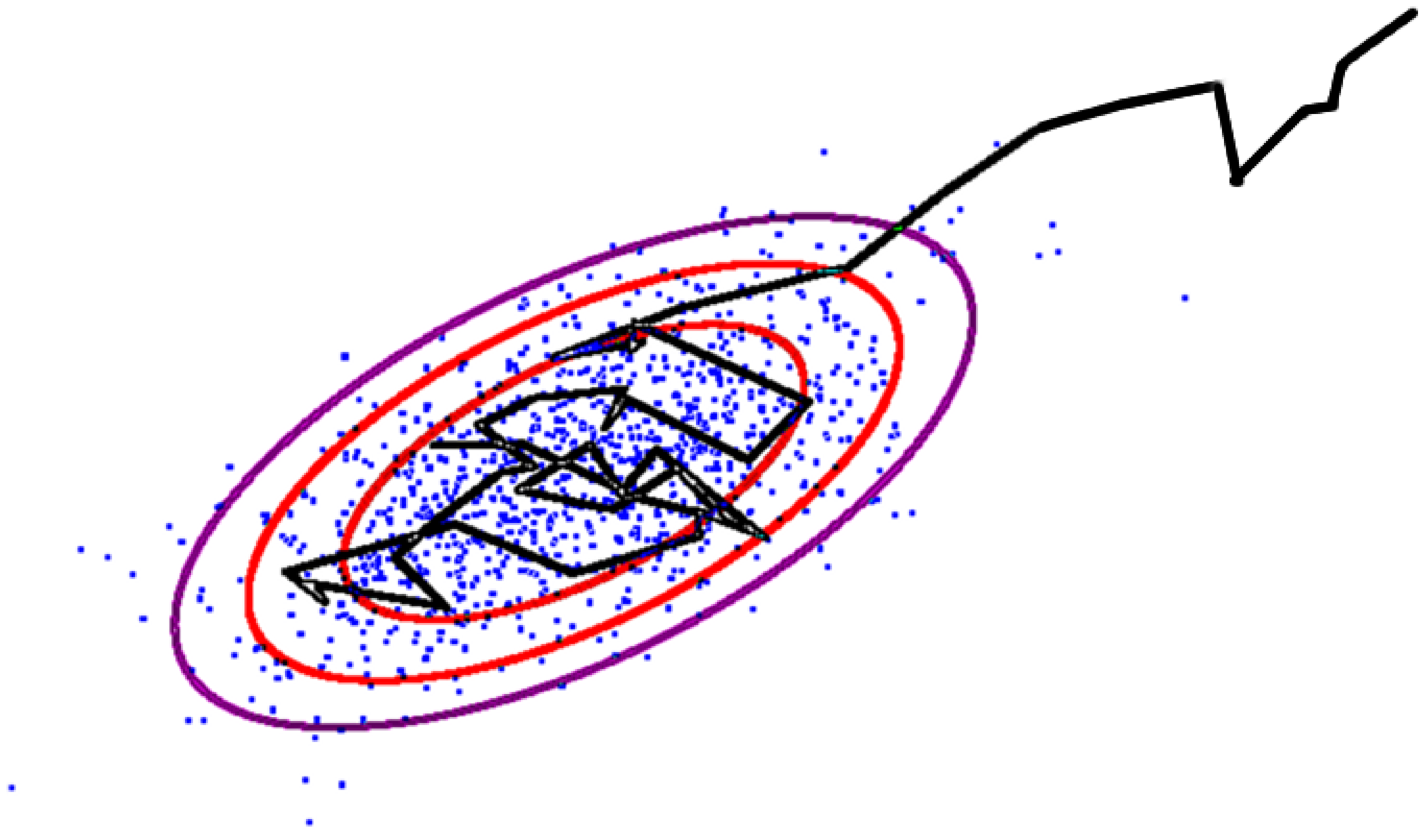
Available free online

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

Markov chain Monte Carlo

Construction a random walk that explores $P(x)$

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



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

Transition operators

Discrete example

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

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

P is a **stationary distribution** of T because $TP = P$, i.e.

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

Also need to explore entire space: $T^K(x' \leftarrow x) > 0$ for all $P(x') > 0$

Detailed balance

Detailed balance means $\rightarrow x \rightarrow x'$ and $\rightarrow x' \rightarrow x$ are equally probable:

$$T(x' \leftarrow x)P(x) = T(x \leftarrow x')P(x')$$

Summing both sides over x:

$$\sum_x T(x' \leftarrow x)P(x) = P(x') \underbrace{\sum_x T(x \leftarrow x')}_{1}$$

detailed balance implies a stationary condition

Enforcing detailed balance is easy: it only involves isolated pairs

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

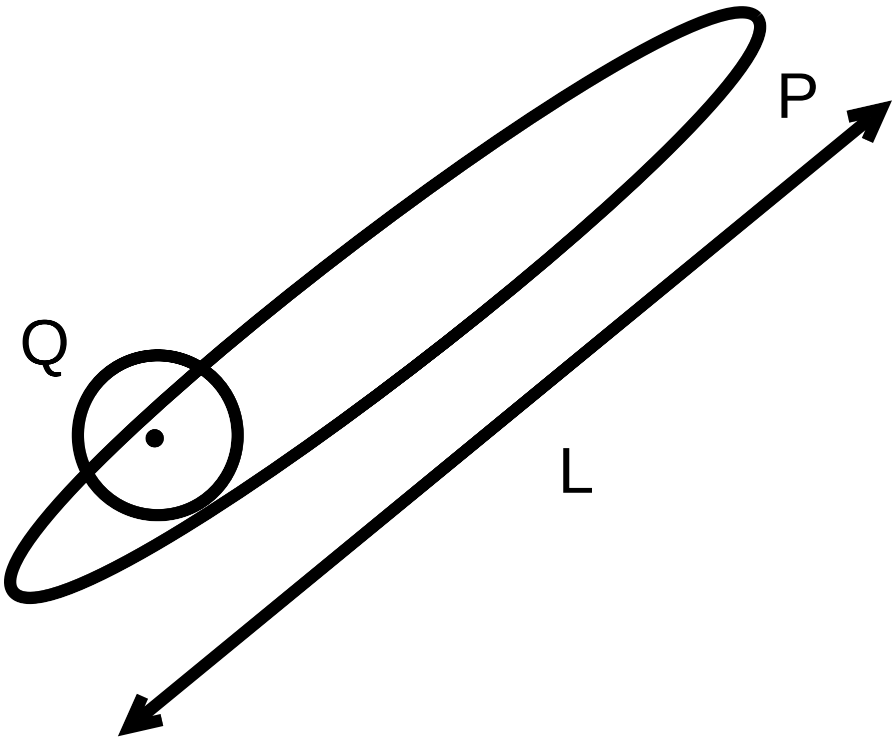
Notes

- Can use $P^* \propto P(x)$; normalizer cancels in acceptance ratio
- Satisfies detailed balance (shown below)
- Q must be chosen to fulfill the other technical requirements

---

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

Metropolis–Hastings



Generic proposals use
$Q(x'; x) = \mathcal{N}(x, \sigma^2)$

σ large $\rightarrow$ many rejections

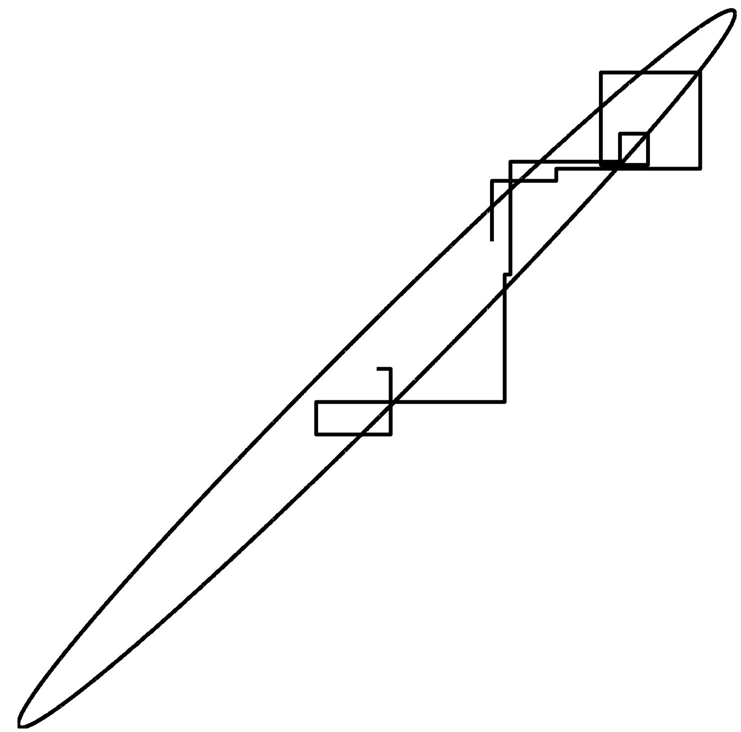
σ small $\rightarrow$ slow diffusion:
$\sim(L/\sigma)^2$ iterations required

Adapted from MacKay (2003)

Gibbs sampling

A method with no rejections:

- Initialize $\mathbf{x}$ to some value
- For each variable in turn successively resample $P(x_i|\mathbf{x}_{j\neq i})$



Proof of validity:

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

Routine Gibbs sampling

Gibbs sampling benefits from few free choices and convenient features of conditional distributions:

- Conditionals with a few discrete settings can be explicitly normalized:

$$P(x_i|\mathbf{x}_{j\neq i}) \propto P(x_i, \mathbf{x}_{j\neq i})$$
$$= \frac{P(x_i, \mathbf{x}_{j\neq i})}{\sum_{x_i'} P(x_i', \mathbf{x}_{j\neq i})} \leftarrow \text{this sum is small and easy}$$

- Continuous conditionals only univariate
  $\Rightarrow$ amenable to standard sampling methods.

`WinBUGS` and `OpenBUGS` sample graphical models using these tricks

<messages>

MCMC

- tackles high-dimensional integrals
- good proposals may require ingenuity
- sometimes simple and routine
- but can be very slow

</messages>

Finding $P(x_i = 1)$

Method 1: fraction of time $x_i = 1$

$$P(x_i = 1) = \sum_{x_i} \mathbb{I}(x_i = 1) P(x_i) \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}(x_i^{(s)}), \quad x_i^{(s)} \sim P(x_i)$$

Method 2: average of $P(x_i = 1 | \mathbf{x}_{\setminus i})$

$$P(x_i = 1) = \sum_{\mathbf{x}_{\setminus i}} P(x_i = 1 | \mathbf{x}_{\setminus i}) P(\mathbf{x}_{\setminus i})$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} P(x_i = 1 | \mathbf{x}_{\setminus i}^{(s)}), \quad \mathbf{x}_{\setminus i}^{(s)} \sim P(\mathbf{x}_{\setminus i})$$

Processing samples

This is easy

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

But we can do better

$$I = \sum_{\mathbf{x}} f(x_i) P(x_i | \mathbf{x}_{\setminus i}) P(\mathbf{x}_{\setminus i}) = \sum_{\mathbf{x}_{\setminus i}} \left(\sum_{x_i} f(x_i) P(x_i | \mathbf{x}_{\setminus i}) \right) P(\mathbf{x}_{\setminus i})$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} \left(\sum_{x_i} f(x_i) P(x_i | \mathbf{x}_{\setminus i}^{(s)}) \right), \quad \mathbf{x}_{\setminus i}^{(s)} \sim P(\mathbf{x}_{\setminus i})$$

A “Rao-Blackwellization”. See also “waste recycling”

<messages>

Using samples effectively

- Monte Carlo is inherently noisy
- conditioned on some samples many integrals become tractable
- **There is a choice of estimators. . .**
  . . . did we remember to consider it?

</messages>

Auxiliary variables

The point of MCMC is to marginalize out variables, but one can introduce more variables:

$$\int f(x)P(x)\,\mathrm{d}x = \int f(x)P(x,v)\,\mathrm{d}x\,\mathrm{d}v$$

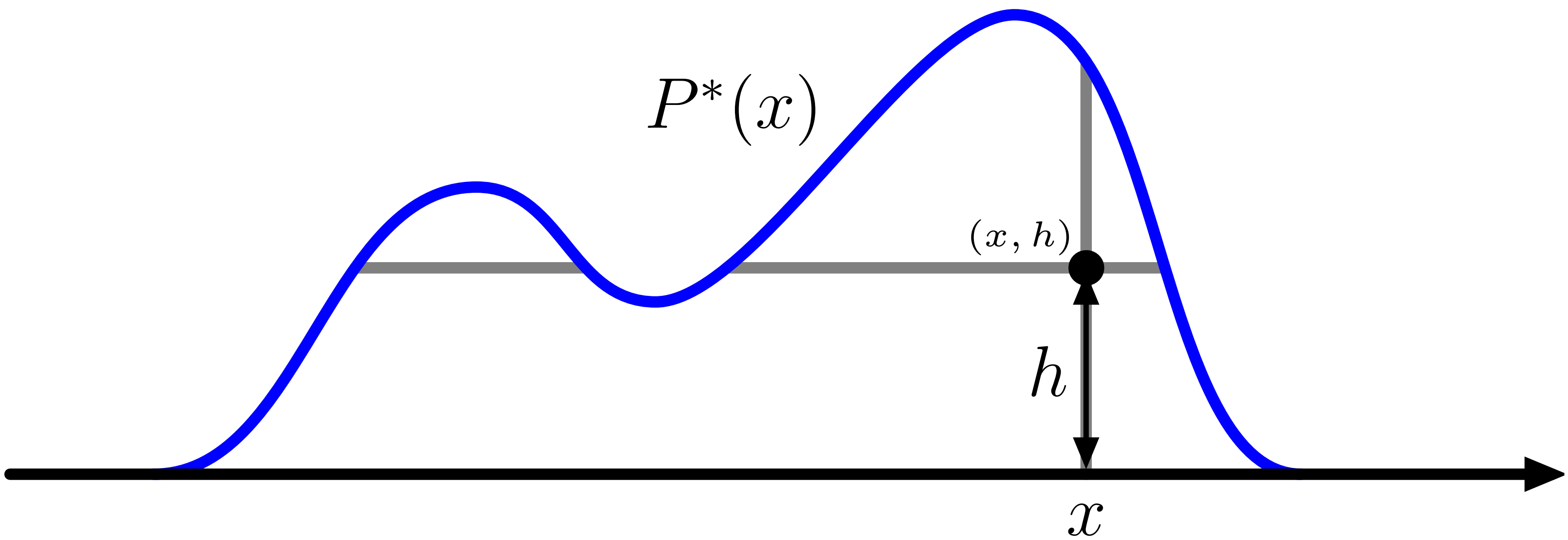
$$\approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), \quad x,v \sim P(x,v)$$

We might want to do this if

- $P(x|v)$ and $P(v|x)$ are simple
- $P(x,v)$ is otherwise easier to navigate

Slice sampling idea

Sample point uniformly under curve $P^*(x) \propto P(x)$



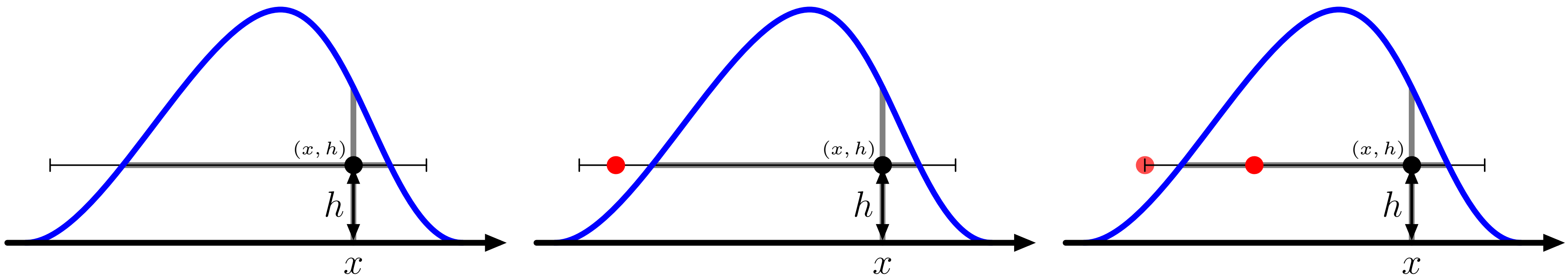
Height h is an auxiliary variable:

$$p(h|x) = \text{Uniform}[0, P^*(x)]$$

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

Slice sampling

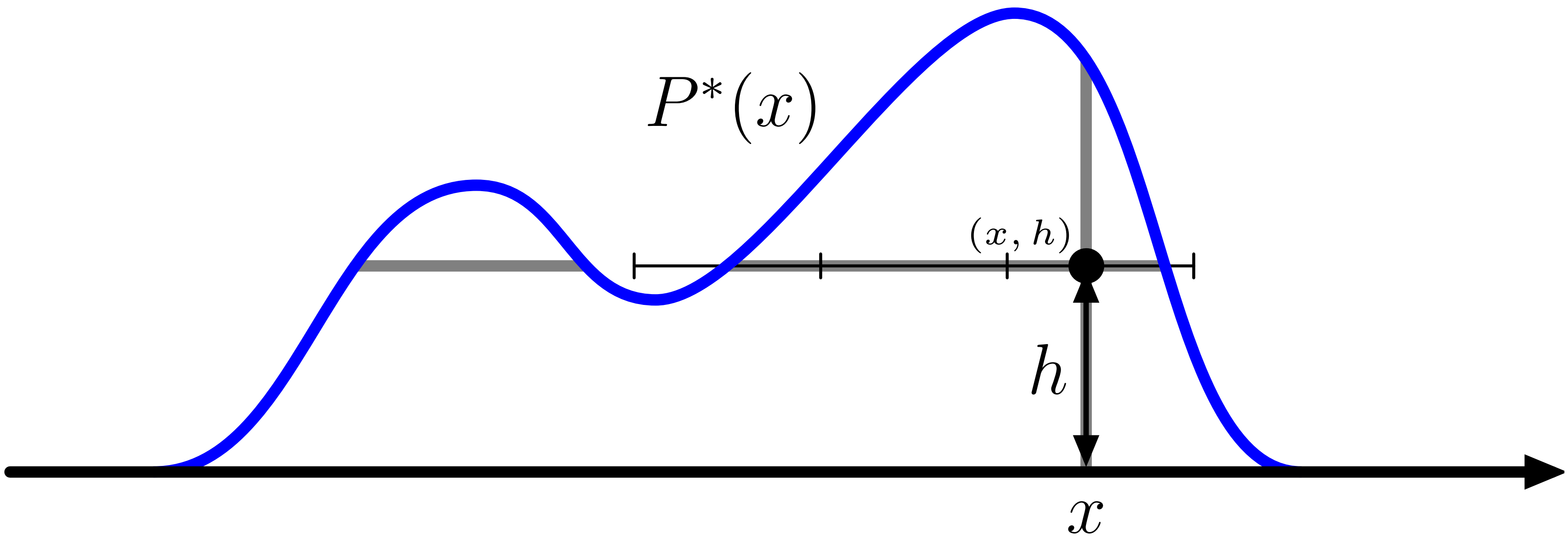
Unimodal conditionals



- bracket slice
- sample uniformly within bracket
- shrink bracket if $P^*(x) < h$ (off slice)
- accept first point on the slice

Slice sampling

Multimodal conditionals



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

Satisfies detailed balance, leaves $p(x|h)$ invariant

Slice sampling

The many nice features of slice-sampling:

- Easy — only require $P^*(x) \propto P(x)$ pointwise
- No rejections
- Step-size parameters less important than Metropolis
- Linear bracketing one of several operators on slice
- Also provides frameworks for:
  - adaptation
  - random walk reduction

Hamiltonian dynamics

Construct a landscape with gravitational potential energy, E(x):

$$P(x) \propto e^{-E(x)}, \qquad E(x) = -\log P^*(x)$$

Introduce velocity v carrying kinetic energy $K(v) = v^\top v/2$

Some physics:

- Total energy or Hamiltonian, $H = E(x) + K(v)$
- Frictionless ball rolling $(x, v) \rightarrow (x', v')$ satisfies $H(x', v') = H(x, v)$
- Ideal Hamiltonian dynamics are time reversible:
  - reverse v and the ball will return to its start point

Hamiltonian Monte Carlo

Define a joint distribution:

- $P(x, v) \propto e^{-E(x)} e^{-K(v)} = e^{-E(x)-K(v)} = e^{-H(x,v)}$
- Velocity independent of position and Gaussian distributed

Markov chain operators

- Gibbs sample velocity
- Simulate Hamiltonian dynamics then flip sign of velocity
  - Hamiltonian 'proposal' is deterministic and reversible
    $q(x', v'; x, v) = q(x, v; x', v') = 1$
  - Conservation of energy means $P(x, v) = P(x', v')$
  - Metropolis acceptance probability is 1

Except we can't simulate Hamiltonian dynamics exactly

Leap-frog dynamics

a discrete approximation to Hamiltonian dynamics:

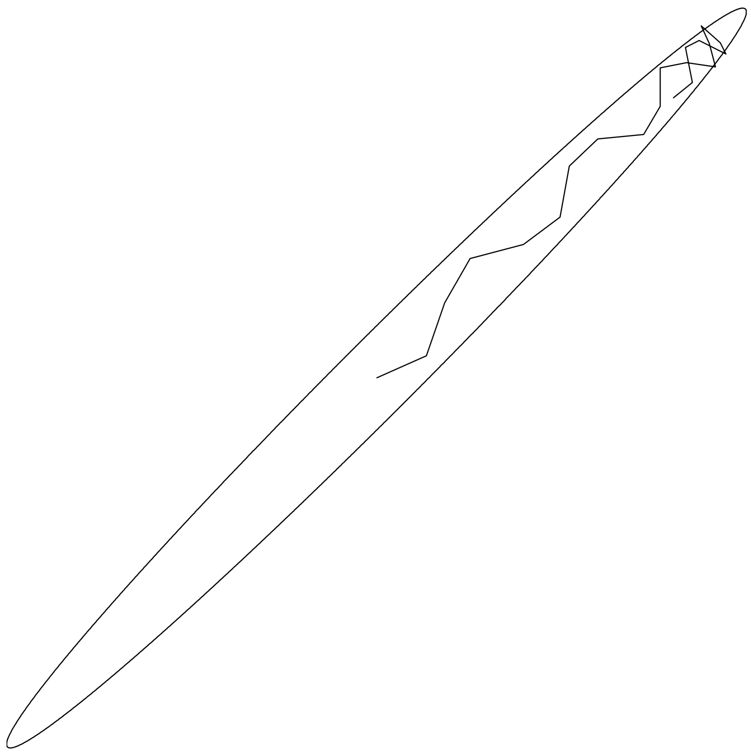
$$\begin{aligned} v_i(t+\tfrac{\epsilon}{2}) &= v_i(t) - \frac{\epsilon}{2}\,\frac{\partial E(x(t))}{\partial x_i} \\ x_i(t+\epsilon) &= x_i(t) + \epsilon\, v_i(t+\tfrac{\epsilon}{2}) \\ p_i(t+\epsilon) &= v_i(t+\tfrac{\epsilon}{2}) - \frac{\epsilon}{2}\,\frac{\partial E(x(t+\epsilon))}{\partial x_i} \end{aligned}$$

- H is not conserved
- dynamics are still deterministic and reversible
- Acceptance probability becomes $\min[1, \exp(H(v,x) - H(v',x'))]$

Hamiltonian Monte Carlo

The algorithm:

- Gibbs sample velocity $\sim \mathcal{N}(0,1)$
- Simulate Leapfrog dynamics for L steps
- Accept new position with probability $\min[1, \exp(H(v,x) - H(v',x'))]$



The original name is **Hybrid Monte Carlo**, a *hybrid* of traditional dynamical simulation and the Metropolis algorithm.

<messages>

Auxiliary variables

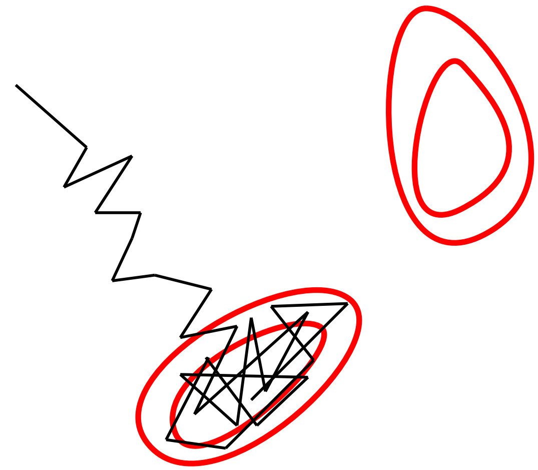
– potentially a computational burden
– can make using MCMC simpler:
Slice sampling robust to step-sizes
– can help navigation:
HMC uses gradient information

</messages>

Three problems

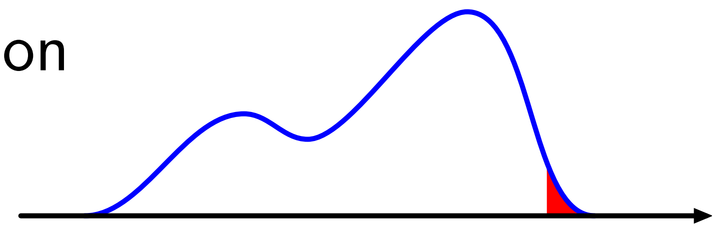
Mixing:

Efficient burn-in and mode exploration can be a problem



Rare events:

Need many samples from a distribution to estimate its tail



Normalizing constants

$$p(x) = \frac{p^*(x)}{\mathcal{Z}},$$ MCMC doesn't need $\mathcal{Z}$. . . or find it either

Finding normalizers is hard

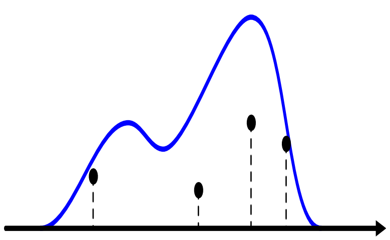
Prior sampling: **like finding fraction of needles in a hay-stack**

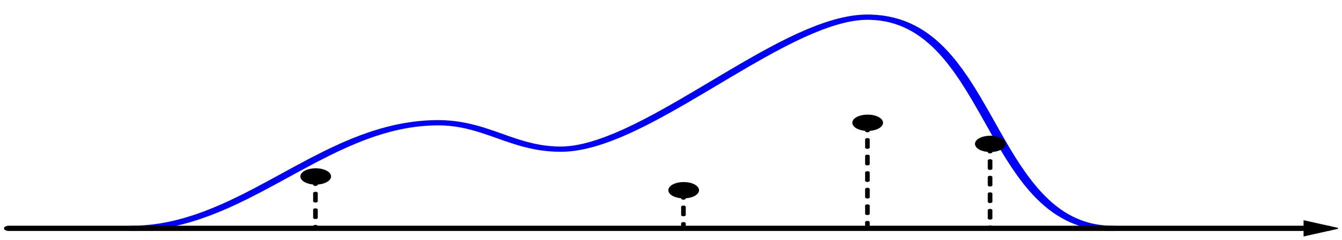
$$P(\mathcal{D}|\mathcal{M}) = \int P(\mathcal{D}|\theta, \mathcal{M}) P(\theta|\mathcal{M}) \, \mathrm{d}\theta$$

$$= \frac{1}{S} \sum_{s=1}^{S} P(\mathcal{D}|\theta^{(s)}, \mathcal{M}), \quad \theta^{(s)} \sim P(\theta|\mathcal{M})$$

. . . can have huge or infinite variance

Posterior sampling: returns values at large *points*, but not spacing

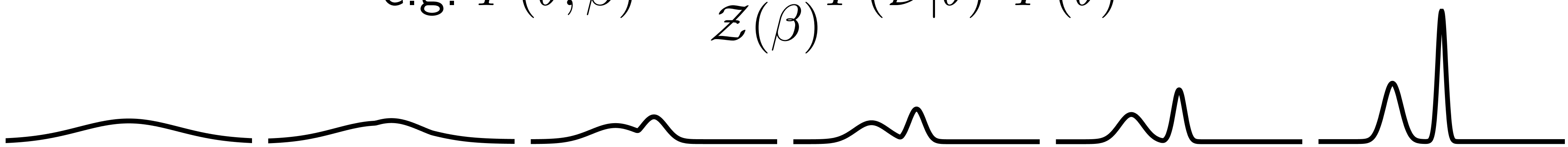




Using other distributions

Bridge between posterior and prior:

$$\text{e.g. } P(\theta;\beta) = \frac{1}{\mathcal{Z}(\beta)} P(\mathcal{D}|\theta)^{\beta} P(\theta)$$



$\beta = 0$ $\beta = 0.01$ $\beta = 0.1$ $\beta = 0.25$ $\beta = 0.5$ $\beta = 1$

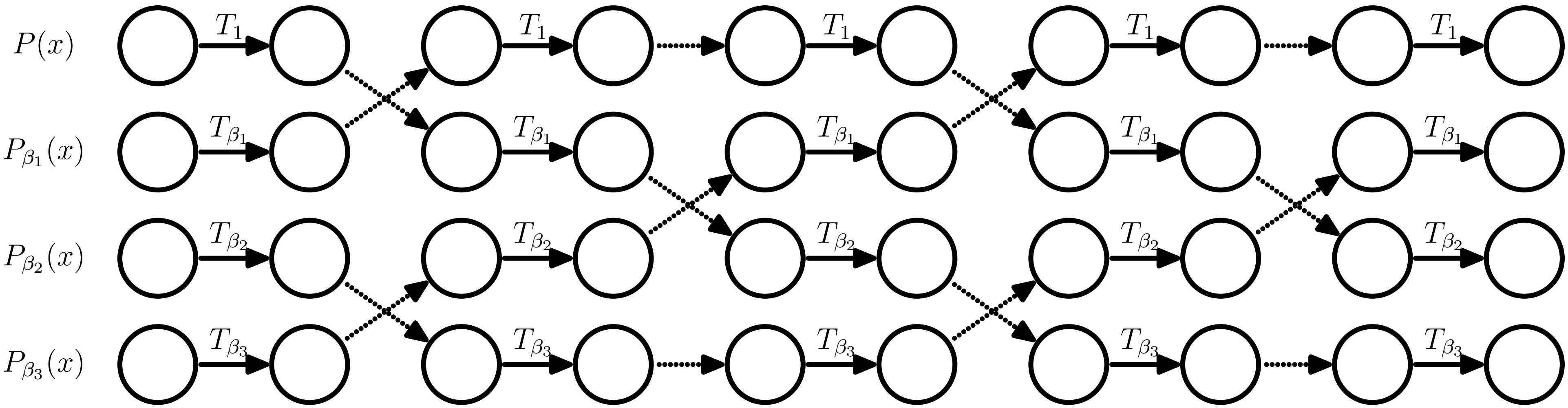
Advantages:

- mixing easier at low β, good initialization for higher β?
- $\displaystyle \frac{\mathcal{Z}(1)}{\mathcal{Z}(0)} = \frac{\mathcal{Z}(\beta_1)}{\mathcal{Z}(0)} \cdot \frac{\mathcal{Z}(\beta_2)}{\mathcal{Z}(\beta_1)} \cdot \frac{\mathcal{Z}(\beta_3)}{\mathcal{Z}(\beta_2)} \cdot \frac{\mathcal{Z}(\beta_4)}{\mathcal{Z}(\beta_3)} \cdot \frac{\mathcal{Z}(1)}{\mathcal{Z}(\beta_4)}$

Related to *annealing* or *tempering*, $1/\beta =$ “temperature”

Parallel tempering

Normal MCMC transitions + swap proposals on $P(X) = \prod_{\beta} P(X;\beta)$

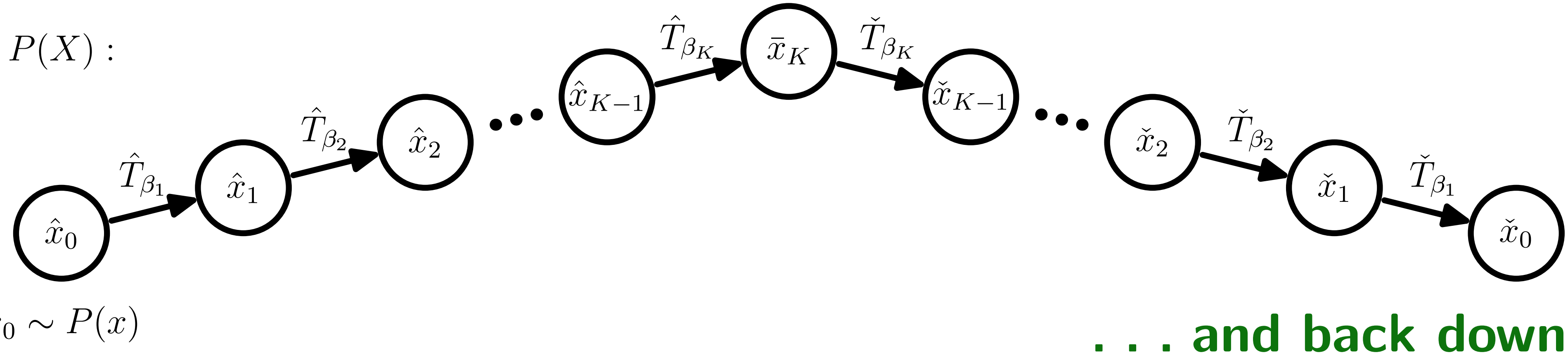


Problems / trade-offs:

- obvious space cost
- need to equilibriate larger system
- information from low β diffuses up by slow random walk

Tempered transitions

Drive temperature up. . .



. . . and back down

Proposal: swap order of points so final point $\check{x}_0$ putatively $\sim P(x)$

Acceptance probability:

$$\min\left[1,\ \frac{P_{\beta_1}(\hat{x}_0)}{P(\hat{x}_0)} \cdots \frac{P_{\beta_K}(\hat{x}_{K-1})}{P_{\beta_{K-1}}(\hat{x}_0)} \frac{P_{\beta_{K-1}}(\check{x}_{K-1})}{P_{\beta_K}(\check{x}_{K-1})} \cdots \frac{P(\check{x}_0)}{P_{\beta_1}(\check{x}_0)}\right]$$

<messages>

Bridging between distributions

- can help mixing
- can usually use your existing code
- gives extra information, e.g. normalizers

</messages>

Key points

- **MCMC** — a powerful tool for high dimensional integrals
- **Use samples effectively:**
  remember to consider alternative estimators
- **Auxiliary variables** — Slice sampling and HMC:
  potentially much faster than simple Metropolis
- **Consider different distributions:**
  helps mixing, answers new questions

[The End]

Further reading (1/2)

General references:

Probabilistic inference using Markov chain Monte Carlo methods, Radford M. Neal, Technical report: CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993. `http://www.cs.toronto.edu/~radford/review.abstract.html`

Information theory, inference, and learning algorithms. David MacKay, 2003. `http://www.inference.phy.cam.ac.uk/mackay/itila/`

Specific points:

The topic modelling figure was adapted from:
Probabilistic topic models, Mark Steyvers and Tom Griffiths, *Latent Semantic Analysis: A Road to Meaning*, T. Landauer, D. McNamara, S. Dennis and W. Kintsch (editors), Laurence Erlbaum, 2006. `http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf`

If you do Gibbs sampling with continuous distributions you should know about this method, which I omitted for material-overload reasons:
Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation, Radford M. Neal, *Learning in graphical models*, M. I. Jordan (editor), 205–228, Kluwer Academic Publishers, 1998. `http://www.cs.toronto.edu/~radford/overk.abstract.html`

An example of picking estimators carefully:
Speed-up of Monte Carlo simulations by sampling of rejected states, Frenkel, D, *Proceedings of the National Academy of Sciences*, 101(51):17571–17575, The National Academy of Sciences, 2004. `http://www.pnas.org/cgi/content/abstract/101/51/17571`

A key reference for auxiliary variable methods is:
Generalizations of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm, Robert G. Edwards and A. D. Sokal, *Physical Review*, 38:2009–2012, 1988.

Slice sampling, Radford M. Neal, *Annals of Statistics*, 31(3):705–767, 2003. `http://www.cs.toronto.edu/~radford/slice-aos.abstract.html`

Bayesian training of backpropagation networks by the hybrid Monte Carlo method, Radford M. Neal, Technical report: CRG-TR-92-1, Connectionist Research Group, University of Toronto, 1992. `http://www.cs.toronto.edu/~radford/bbp.abstract.`

An early reference for parallel tempering:
Markov chain Monte Carlo maximum likelihood, Geyer, C. J, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 156–163, 1991.

Sampling from multimodal distributions using tempered transitions, Radford M. Neal, *Statistics and Computing*, 6(4):353–366, 1996.

Further reading (2/2)

Software:

Gibbs sampling for graphical models: `http://mathstat.helsinki.fi/openbugs/`

Neural networks and other flexible models: `http://www.cs.utoronto.ca/~radford/fbm.software.html`

Other Monte Carlo methods:

Nested sampling is a new Monte Carlo method that challenges the traditional approach to Bayesian computation. Highly recommended reading; I would have needed the entire tutorial to give it justice:
Nested sampling for general Bayesian computation, John Skilling, *Bayesian Analysis*, 2006.
(to appear, posted online June 5). `http://ba.stat.cmu.edu/journal/forthcoming/skilling.pdf`

Approaches based on the "multi-canonicle ensemble" also solve some of the problems with traditional bridging methods:
Multicanonical ensemble: a new approach to simulate first-order phase transitions, Bernd A. Berg and Thomas Neuhaus, *Phys. Rev. Lett*, 68(1):9–12, 1992. `http://prola.aps.org/abstract/PRL/v68/i1/p9_1`

Extended Ensemble Monte Carlo. Y Iba. Int J Mod Phys C [Computational Physics and Physical Computation] 12(5):623-656. 2001.

Particle filters / Sequential Monte Carlo are famously successful in time series modelling, but are more generally applicable.
This may be a good place to start: `http://www.cs.ubc.ca/~arnaud/journals.html`

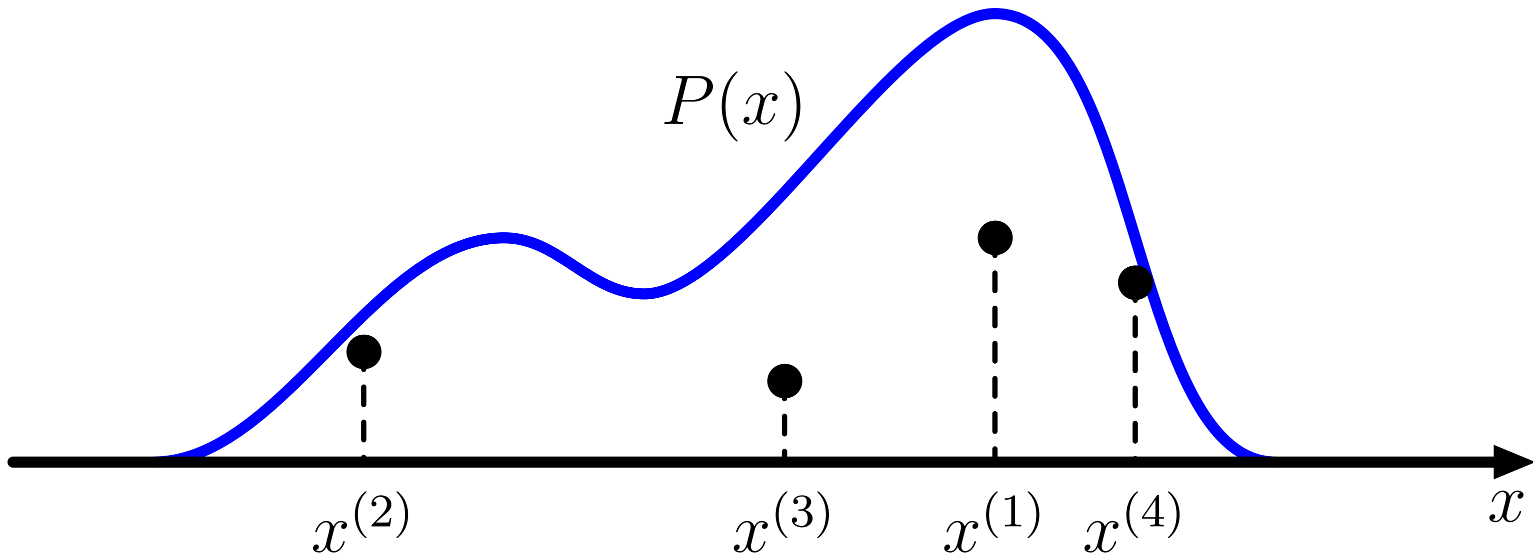
Exact or perfect sampling uses Markov chain simulation but suffers no initialization bias. An amazing feat when it can be performed:
Annotated bibliography of perfectly random sampling with Markov chains, David B. Wilson
`http://dbwilson.com/exact/`

MCMC does not apply to doubly-intractable distributions. For what that even means and possible solutions see:
An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants, J. Møller, A. N. Pettitt, R. Reeves and K. K. Berthelsen, *Biometrika*, 93(2):451–458, 2006.
MCMC for doubly-intractable distributions, Iain Murray, Zoubin Ghahramani and David J. C. MacKay, *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Rina Dechter and Thomas S. Richardson (editors), 359–366, AUAI Press, 2006.
`http://www.gatsby.ucl.ac.uk/~iam23/pub/06doubly_intractable/doubly_intractable.pdf`

Assorted spare slides

Sampling from distributions

Draw points from the unit area under the curve



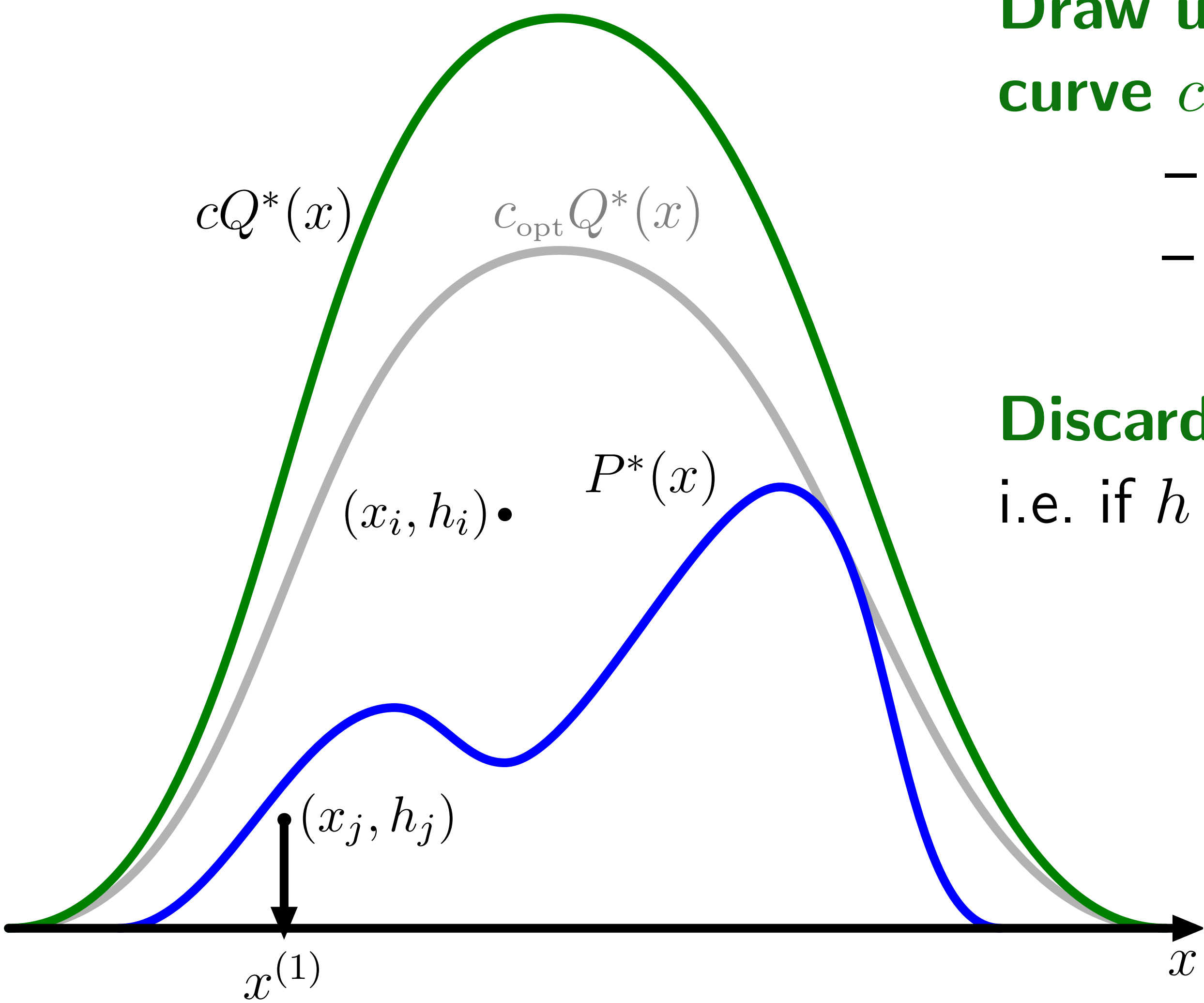
Draw probability mass to left of point, $u \sim$ Uniform[0,1]

Sample $x(u) = c^{-1}(u)$, where $c(x) = \int_{-\infty}^{x} P(x')\,\mathrm{d}x'$

Problem: often can't even normalize P, eg $P(\theta|\mathcal{D}) \propto P(D|\theta)P(\theta)$

Rejection sampling

Sampling underneath a $P^*(x) \propto P(x)$ curve is also valid



Draw underneath a simple curve $cQ^*(x) \geq P^*(x)$:

- Draw $x \sim Q(x)$
- height $h \sim$ Uniform$[0, cQ^*(x)]$

Discard points above P^*,
i.e. if $h > P^*(x)$

Importance sampling

Computing $P^*(x)$ and $Q^*(x)$, then *throwing* x *away* seems wasteful
Instead rewrite the integral as an expectation under Q:

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

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

Unbiased; but light-tailed $Q(x)$ can give the estimator infinite variance

. . . and you might not notice.

Importance sampling applies when the integral is not an expectation.

Importance sampling (2)

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

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

$$\approx \frac{\cancel{1}}{\cancel{S}}\sum_{s=1}^{S} f(x^{(s)})\frac{w^{(s)}}{\frac{\cancel{1}}{\cancel{S}}\sum_{s'} w^{(s')}}$$

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

Note that $\mathcal{Z}_P/\mathcal{Z}_Q \approx \frac{1}{S}\sum_s w^{(s)}$

Doubly-intractable problems

MCMC can sample most distributions

- MRFs / Undirected graphical models: $p(x|\theta) = \frac{1}{\mathcal{Z}(\theta)} e^{-E(x;\theta)}$

- parameter posteriors: $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$

Some distributions are much harder

- MRF parameter posterior: $p(\theta|x) = \frac{\frac{1}{\mathcal{Z}(\theta)} e^{-E(x;\theta)} p(\theta)}{p(x)}$

See Møller et al. (2004, 2006) and Murray et al. (2006) for partial solutions