Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 16:
Markov Chain Monte Carlo Methods, Metropolis-Hastings Algorithm, Gibbs Sampler

Some slides and figures courtesy Iain Murray's tutorial,
Markov Chain Monte Carlo, MLSS 2009

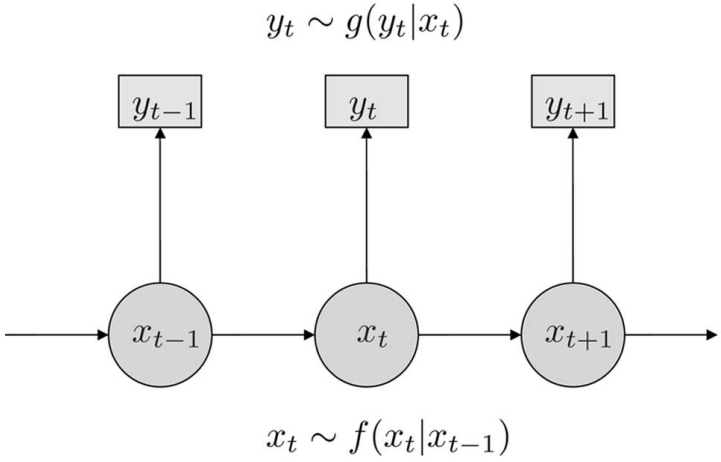
General Sequential Monte Carlo

$$\bar{h} = \int h(x_{0:t})\pi_{0:t|0:t}(x_{0:t}|y_{0:t})dx_{0:t}$$
$$\approx \sum_{i=1}^{N} \frac{\tilde{\omega}_t^{(i)}}{\sum_{j=1}^{N}\tilde{\omega}_t^{(j)}} h\left(\tilde{x}_{0:t}^{(i)}\right).$$

$$\tilde{\omega}_t^{(i)} = \frac{\pi_{0:t|0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)}{q_{0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)}$$

Target Posterior

Proposal Density



- Exploit temporal structure to propose sequences recursively:

$$q_{0:t}(x_{0:t}|y_{0:t}) = \overbrace{q_{0:t-1}(x_{0:t-1}|y_{0:t-1})}^{\text{Keep existing path}}\overbrace{q_t(x_t|x_{t-1},y_t)}^{\text{extend path}}$$

$$\tilde{\omega}_t^{(i)} = \frac{\pi_{0:t|0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)}{q_{0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)} \propto \omega_{t-1}^{(i)} \times \frac{f\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)}\right)g\left(y_t|\tilde{x}_t^{(i)}\right)}{q_t\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)},y_t\right)\ell_{t|0:t-1}(y_t|y_{0:t-1})}$$

- Common choices of proposal distribution:

$$q_t(x_t \mid x_{t-1}, y_t) = f(x_t \mid x_{t-1})$$

Bootstrap filter simulates prior dynamical model

$$q_t(x_t \mid x_{t-1}, y_t) \approx p(x_t \mid x_{t-1}, y_t)$$

If local posterior intractable, use Gaussian approximations

Particle Resampling or Selection

$$p(x_t \mid y_{\bar{t}}) \approx \sum_{\ell=1}^{L} \omega_t^{(\ell)} \delta_{x_t^{(\ell)}}(x_t)$$

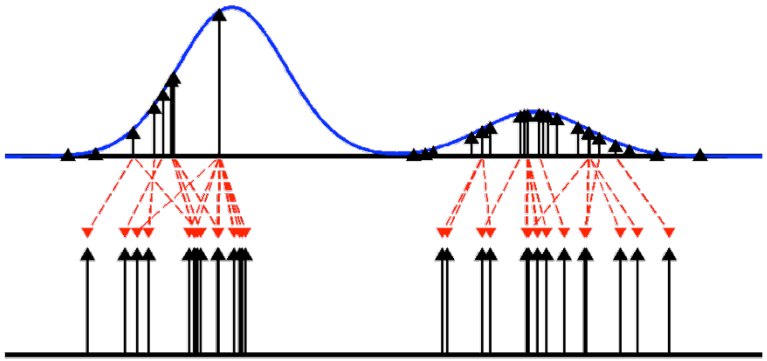
$$p(x_t \mid y_{\bar{t}}) \approx \sum_{\ell=1}^{L} \frac{1}{L} \delta_{\bar{x}_t^{(\ell)}}(x_t)$$

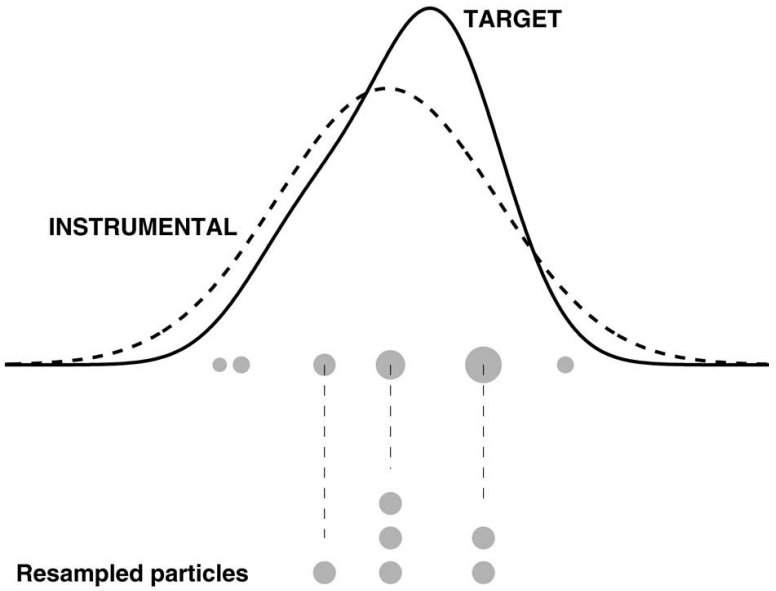
$$\bar{x}_t^{(\ell)} = x_t^{(j_\ell)}$$

$$j_\ell \sim \mathrm{Cat}(\omega_t)$$

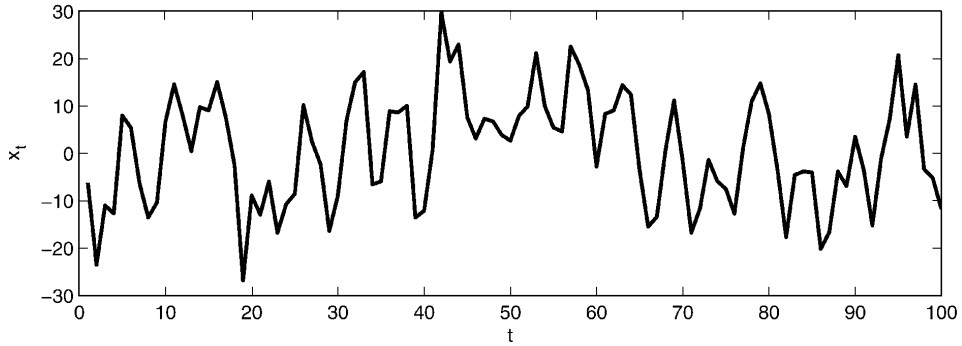
Resampling with replacement produces a random discrete distribution whose mean is the original distribution

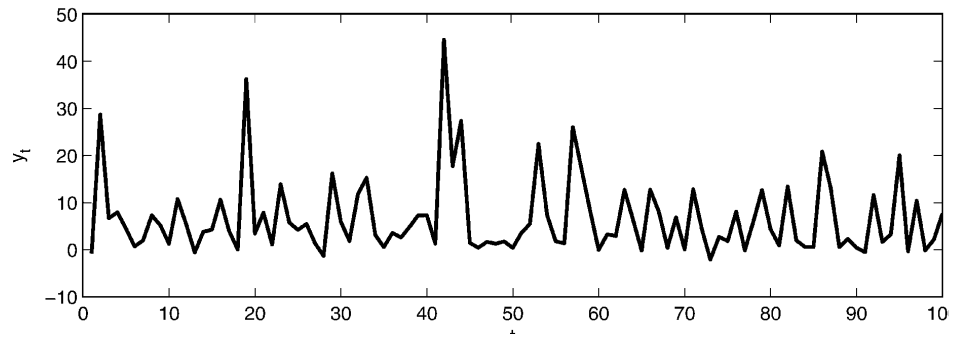
While remaining unbiased, resampling avoids degeneracies in which most weights go to zero

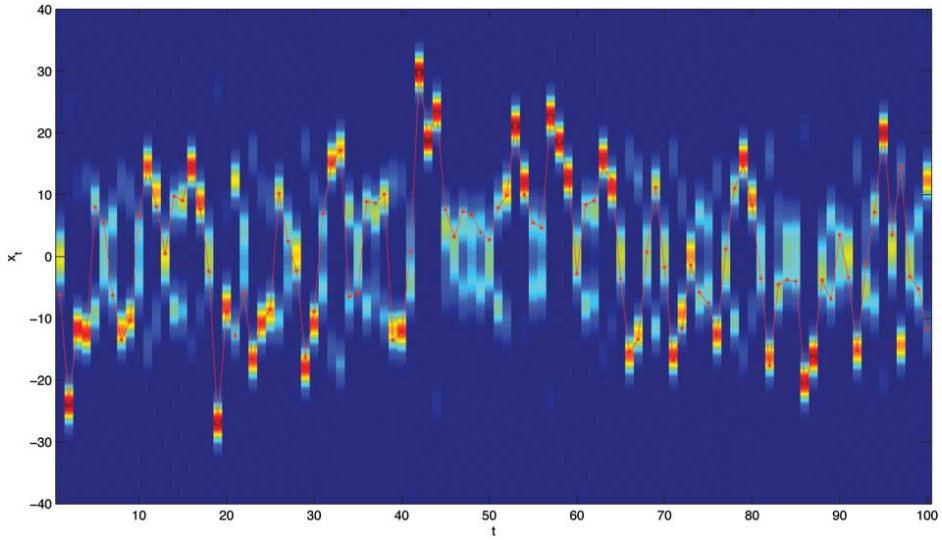


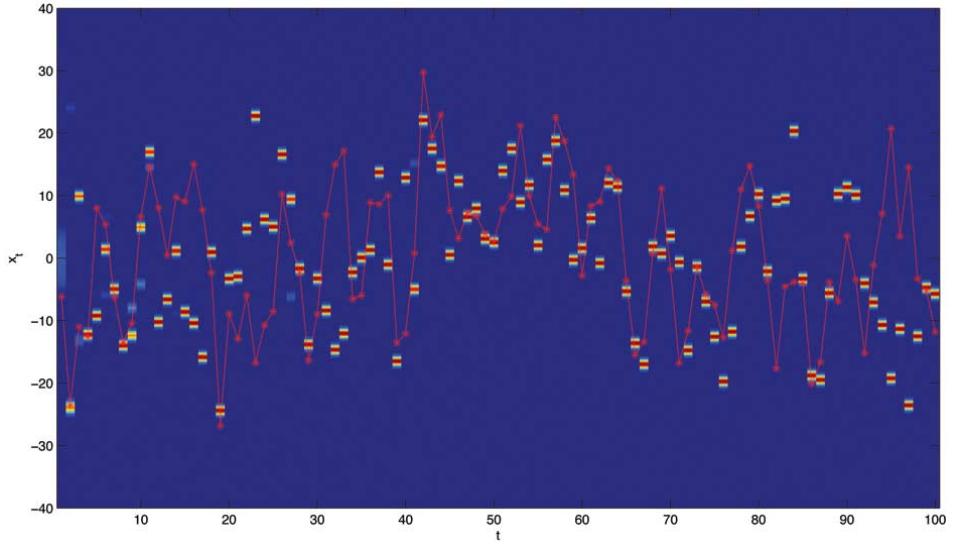


Resampling: The Good







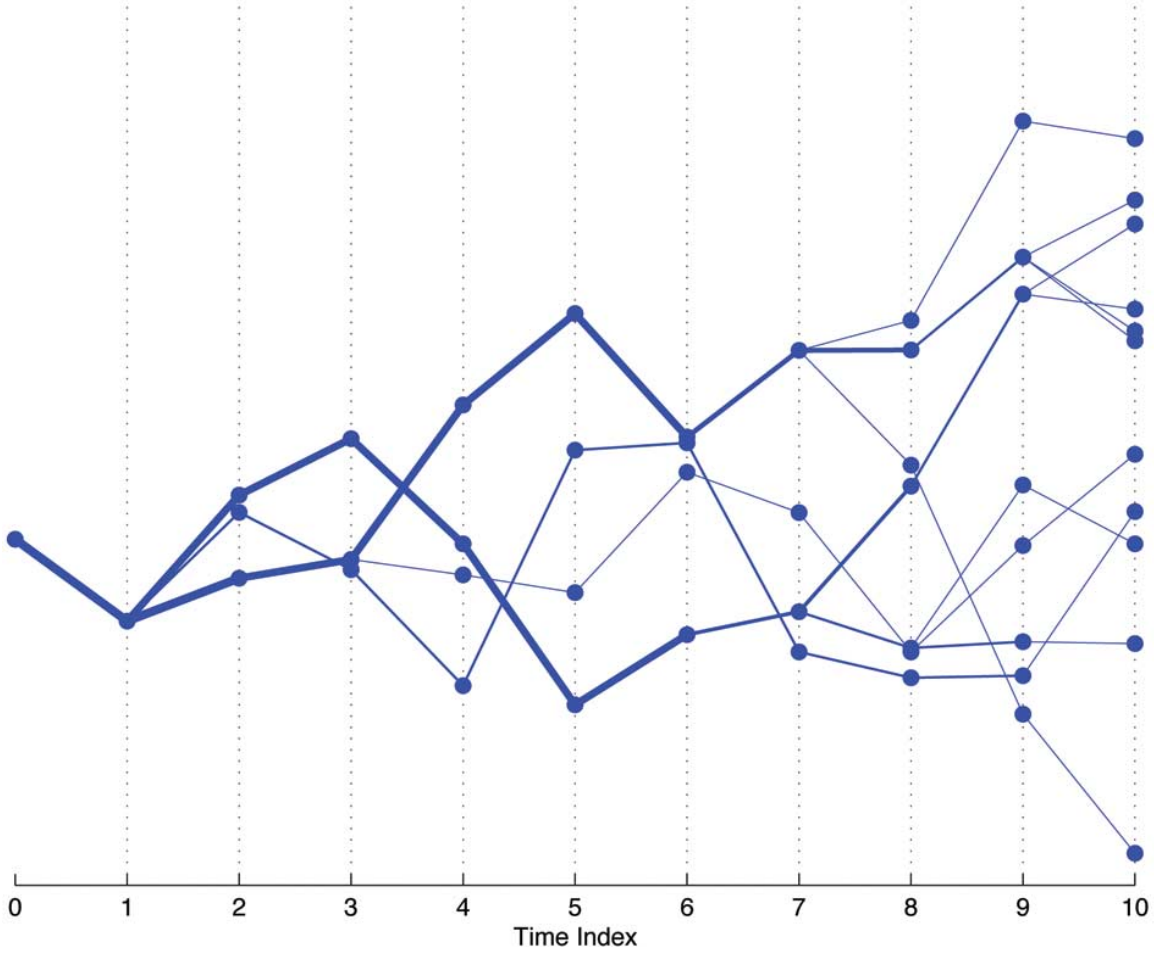


Full Particle Filter KDEs

Full Sequence Importance Sampling

- What is the probability that a state sequence, sampled from the prior model, is consistent with all observations?
- Marginal estimates degenerate on single, mediocre sample

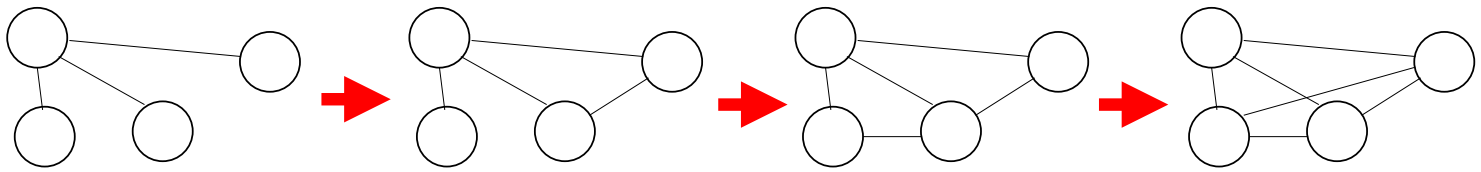
Resampling: The Bad



- After each resampling step, some particles are discarded, and can never be restored in subsequent stages.
- Estimates of *smoothed* marginals/sequences typically poor

Beyond Temporal Particle Filters

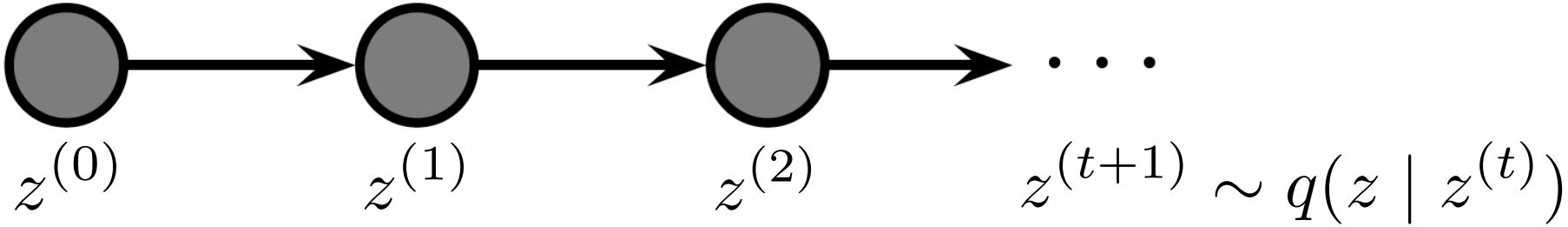
- Can we avoid degeneracies in estimating "smoothed" marginals based on all observations, past and future?
- Can we implement particle-based approximations to BP for tree-structured models, or arbitrary factor graphs?
- Yes! We can apply importance sampling with resampling to any sequence of distributions, such as a sequence of approximations to a complicated graphical model:



Hot Coupling, Hamze & de Freitas, NIPS 2006

- But to do this well, we need to not only be able to reweight and resample particles, we must shift their locations
  - ➢ Markov Chain Monte Carlo (MCMC) methods

Markov Chain Monte Carlo (MCMC)

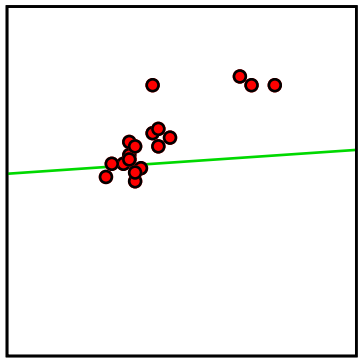


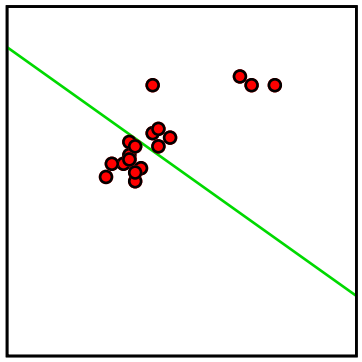
- At each time point, state $z^{(t)}$ is a configuration of *all the variables in the model:* parameters, hidden variables, etc.
- We design the transition distribution $q(z \mid z^{(t)})$ so that the chain is *irreducible* and *ergodic*, with a unique stationary distribution $p^*(z)$

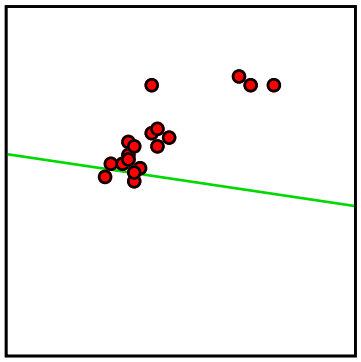
$$p^*(z) = \int_{\mathcal{Z}} q(z \mid z') p^*(z') \, dz'$$

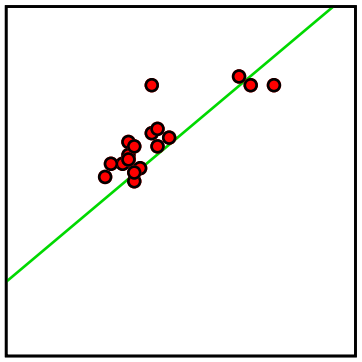
- For learning, the target equilibrium distribution is usually the posterior distribution given data *x*: $p^*(z) = p(z \mid x)$
- Popular recipes: *Metropolis-Hastings and Gibbs samplers*

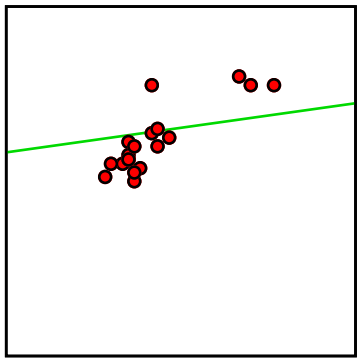
Importance Sampling for Regression

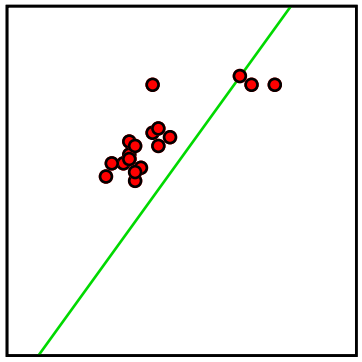

w =0.00548

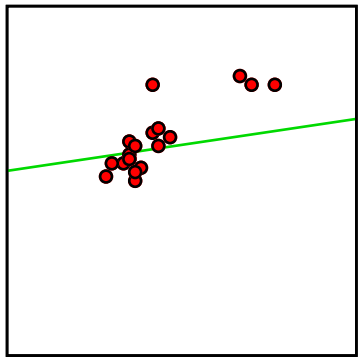

w =1.59e-08

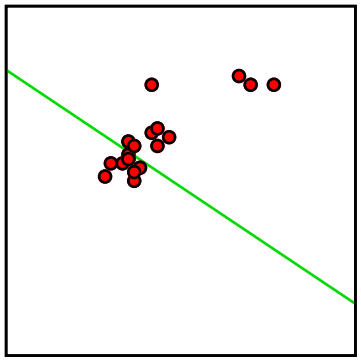

w =9.65e-06

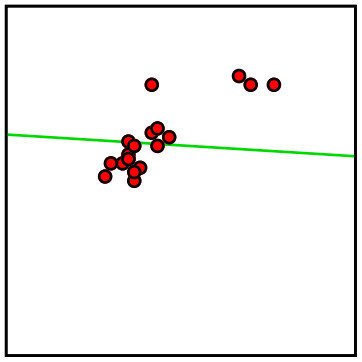

w =0.371

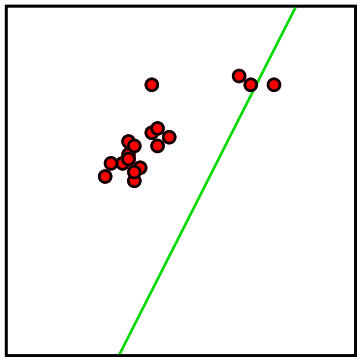

w =0.103


w =1.01e-08


w =0.111

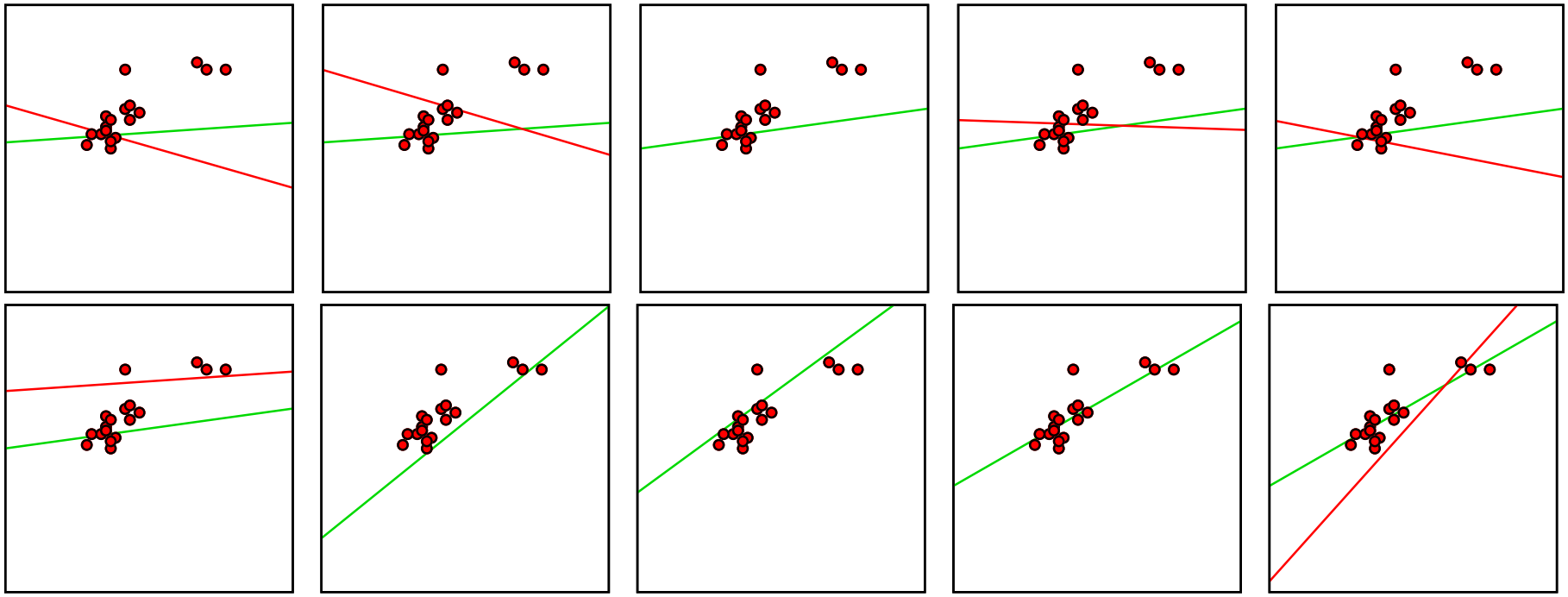

w =1.92e-09


w =0.0126

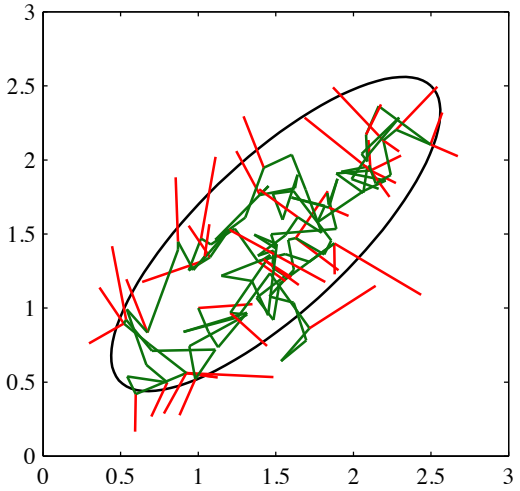

w =1.1e-51

- Model: Gaussian noise around some unknown straight line
- Propose from prior on lines, weight by data likelihood

Metropolis Algorithm for Regression



- 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**



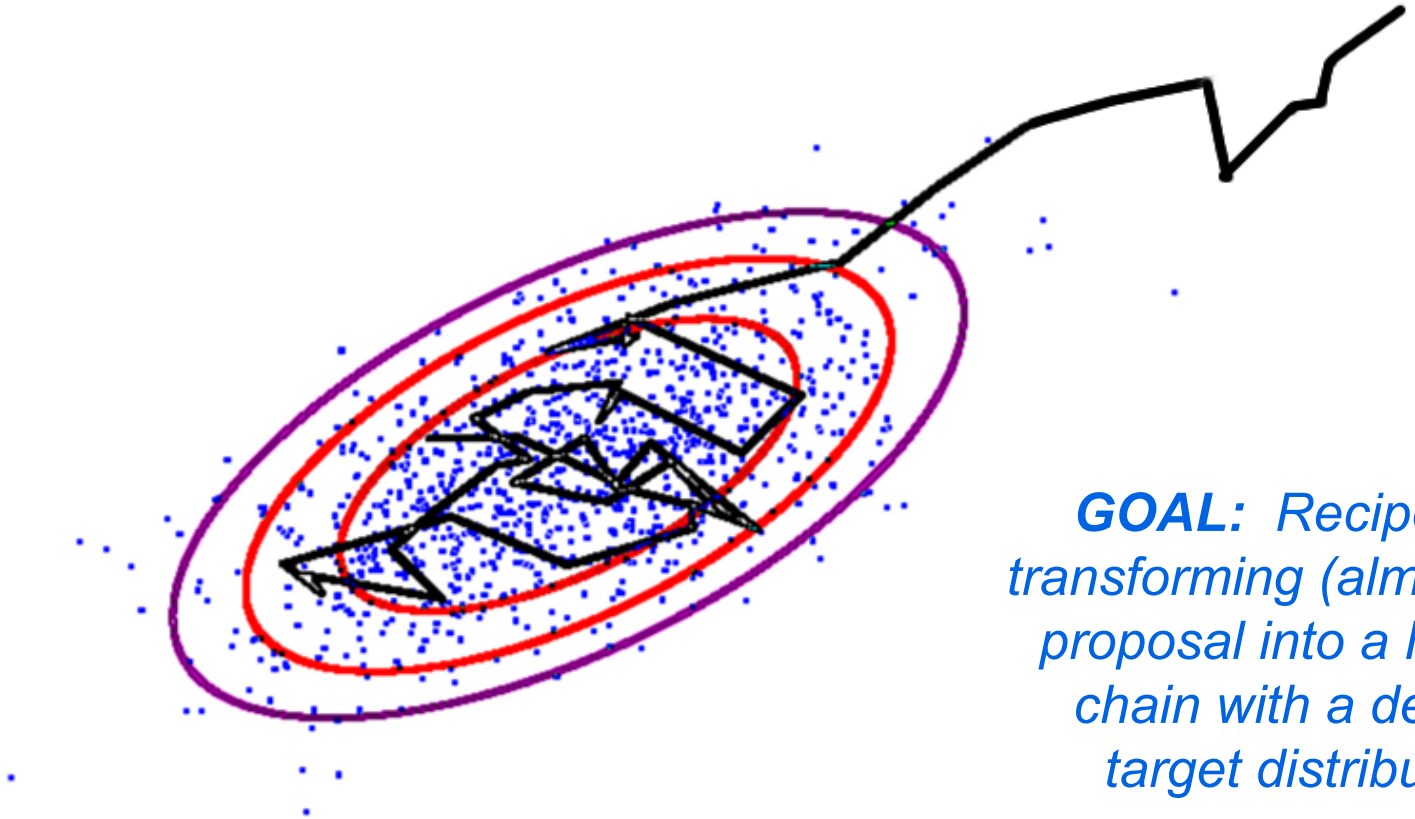
Detail: Metropolis, as stated, requires $Q(\theta';\theta) = Q(\theta;\theta')$

This subfigure from PRML, Bishop (2006)

Markov Chain Monte Carlo

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

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



GOAL: *Recipes for transforming (almost) any proposal into a Markov chain with a desired target distribution*

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

Transition (Proposal) Distributions

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.

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

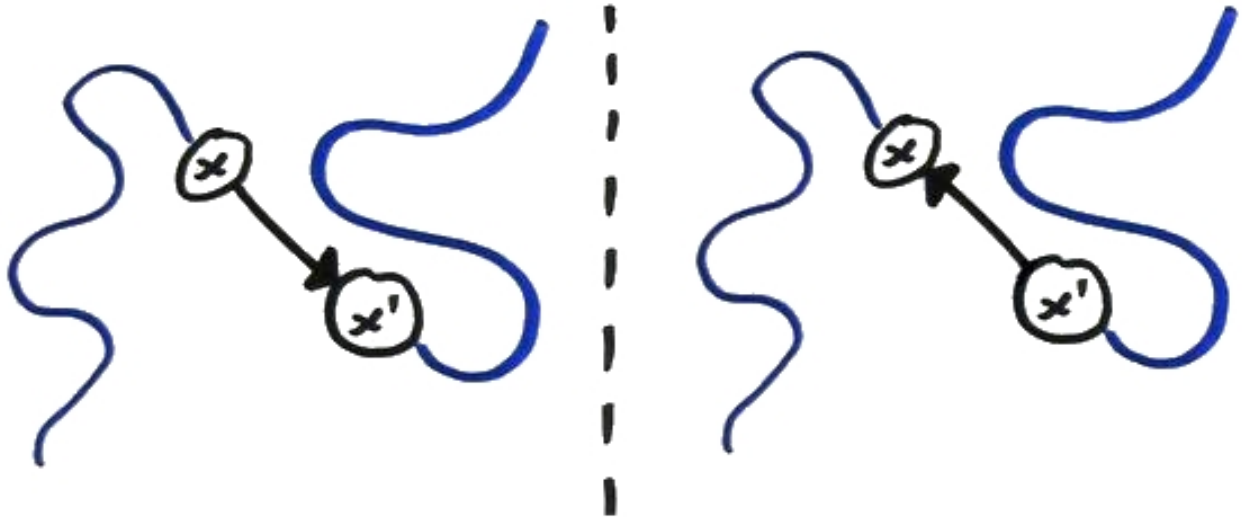
Also $P^\star$ is *the* **equilibrium distribution** of T:

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

Sufficient: Detailed Balance

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



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

Detailed balance implies the invariant condition:

$$\sum_x T(x' \leftarrow x)P^\star(x) \;=\; P^\star(x')\sum_x T(x \leftarrow x') \;\rightarrow\; 1$$

Enforcing detailed balance is easy: it only involves isolated pairs

Necessary: Generalized Balance

If T satisfies stationarity, we can define a *reverse operator*

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

Generalized balance condition:

$$T(x' \leftarrow x) P^\star(x) \;=\; \widetilde{T}(x \leftarrow x') P^\star(x')$$

also implies the invariant condition *and is necessary*.

Operators satisfying detailed balance are their own reverse operator.

Metropolis-Hastings Algorithm

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 to fulfill the other technical requirements

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

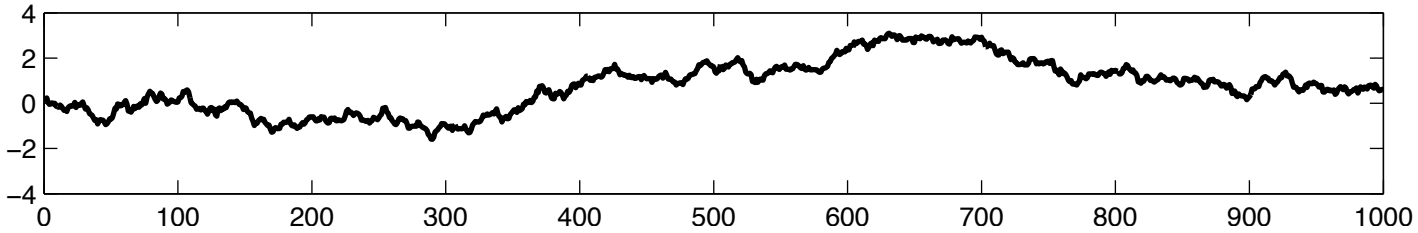
Example: Gaussian Metropolis

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

$$P(x) = \mathcal{N}(x \mid 0, 1) \qquad Q(x'; x) = \mathcal{N}(x' \mid x, \sigma^2)$$

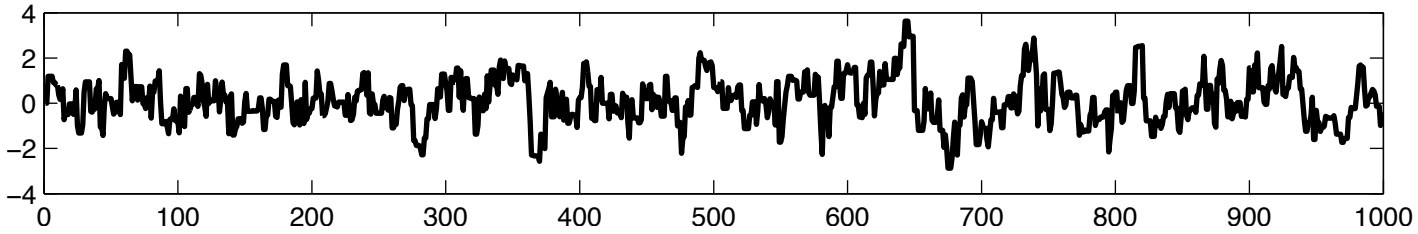
`sigma(0.1)`

99.8% accepts



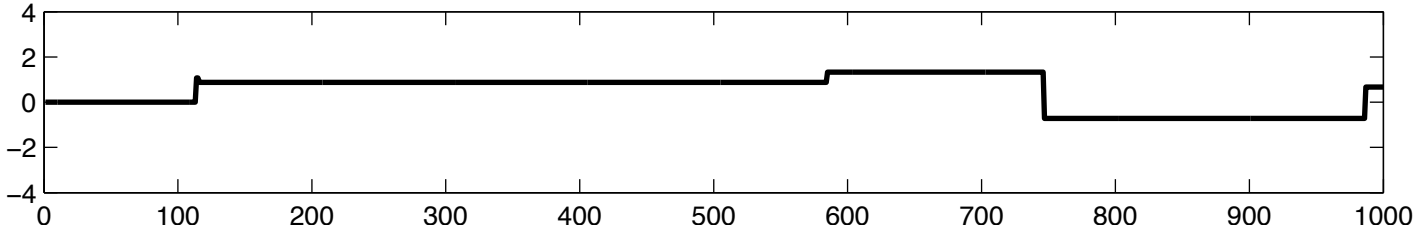
`sigma(1)`

68.4% accepts

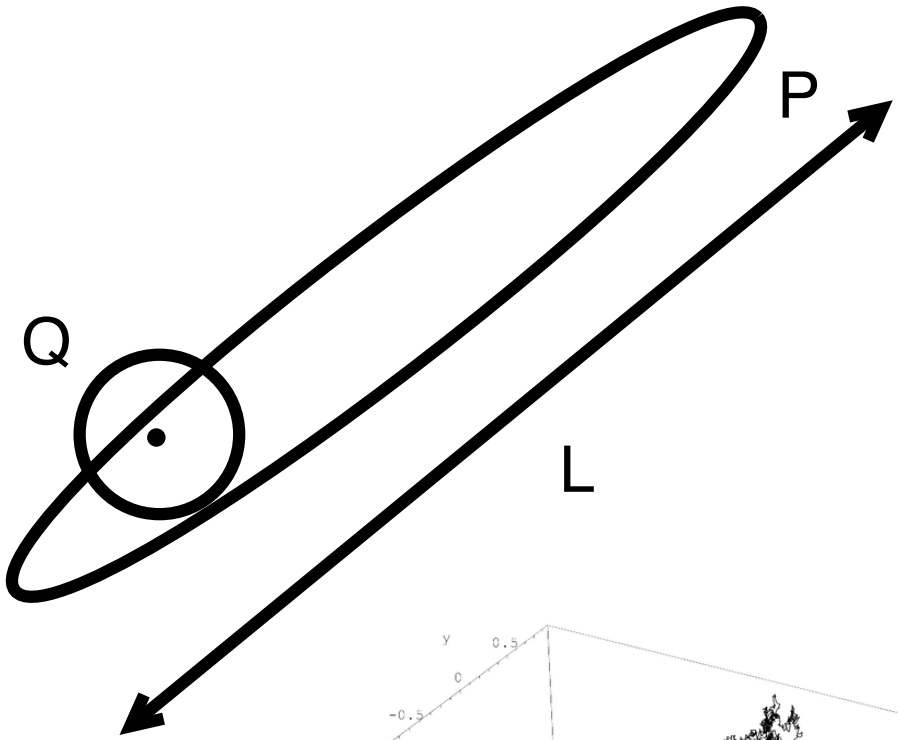


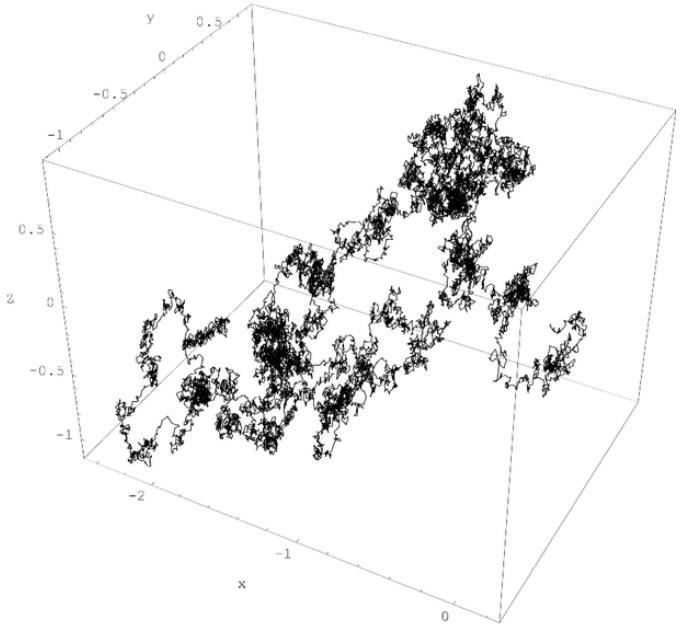
`sigma(100)`

0.5% accepts



Limitations: Metropolis Random Walk





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

σ large $\rightarrow$ many rejections

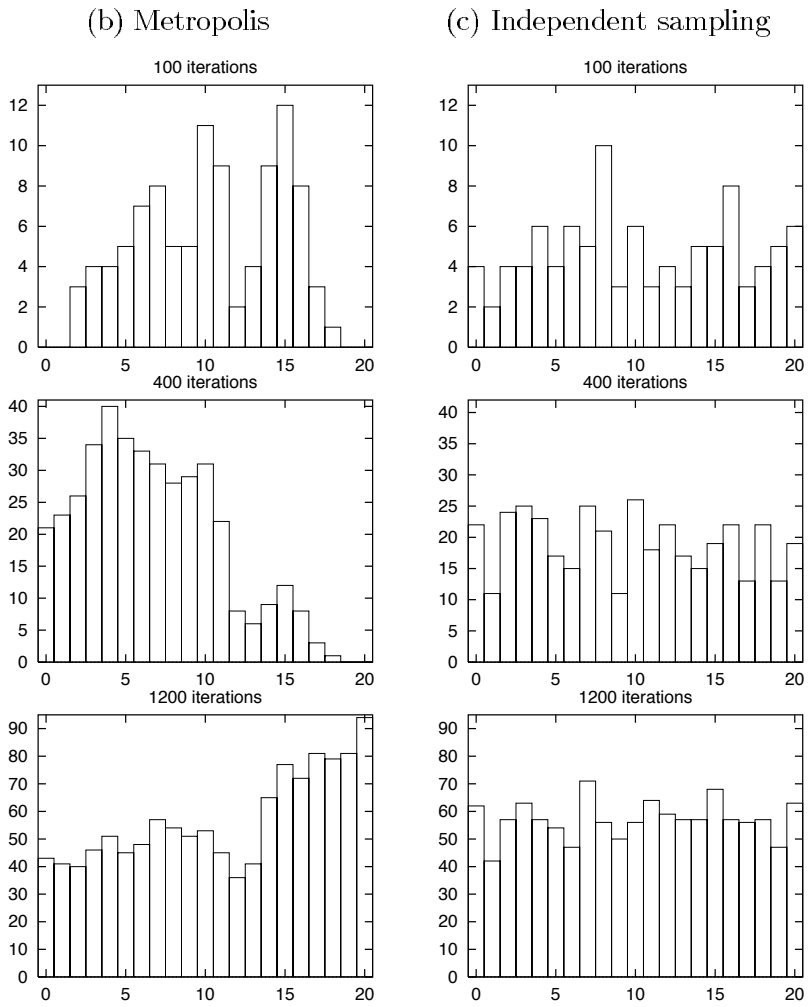
σ small $\rightarrow$ slow diffusion:
$\sim (L/\sigma)^2$ iterations required

Limitations: Metropolis Random Walk

Discrete target distribution is uniform over all states

[redacted]

$$Q(x';x) = \begin{cases} \frac{1}{2} & x' = x \pm 1 \\ 0 & \text{otherwise} \end{cases}$$



Combining MCMC Transition Proposals

A sequence of operators, each with $P^\star$ invariant:

$$
\begin{aligned}
&x_0 \sim P^\star(x) \\
&x_1 \sim T_a(x_1 \leftarrow x_0) \qquad P(x_1) = \textstyle\sum_{x_0} T_a(x_1 \leftarrow x_0) P^\star(x_0) = P^\star(x_1) \\
&x_2 \sim T_b(x_2 \leftarrow x_1) \qquad P(x_2) = \textstyle\sum_{x_1} T_b(x_2 \leftarrow x_1) P^\star(x_1) = P^\star(x_2) \\
&x_3 \sim T_c(x_3 \leftarrow x_2) \qquad P(x_3) = \textstyle\sum_{x_1} T_c(x_3 \leftarrow x_2) P^\star(x_2) = P^\star(x_3) \\
&\quad \ldots \qquad\qquad\qquad\qquad \ldots
\end{aligned}
$$

— Combination $T_c T_b T_a$ leaves $P^\star$ invariant

— If they can reach any x, $T_c T_b T_a$ is a valid MCMC operator

— Individually T_c, T_b and T_a need not be ergodic

Gibbs Samplers

A method with no rejections:

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

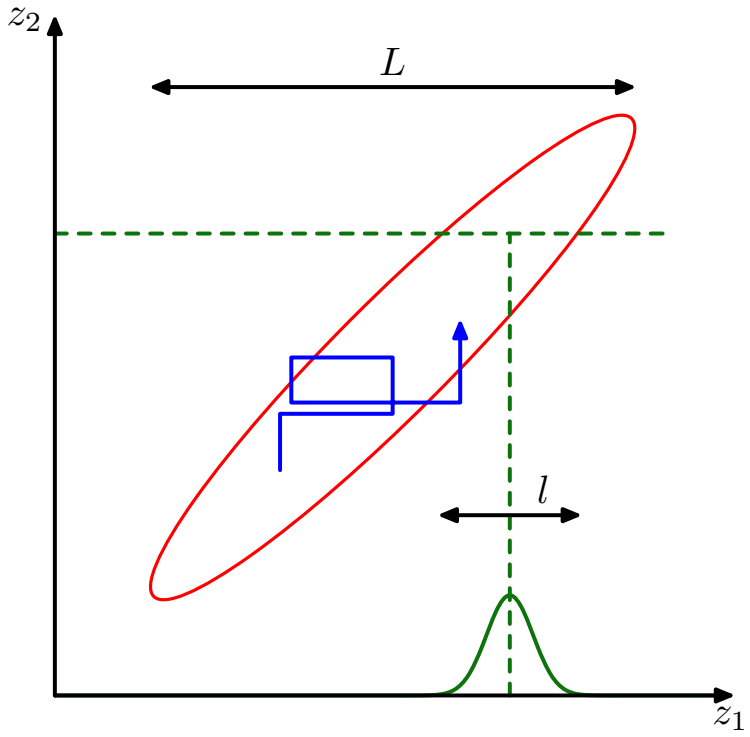


Figure from PRML, Bishop (2006)

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 Samplers

A method with no rejections:

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

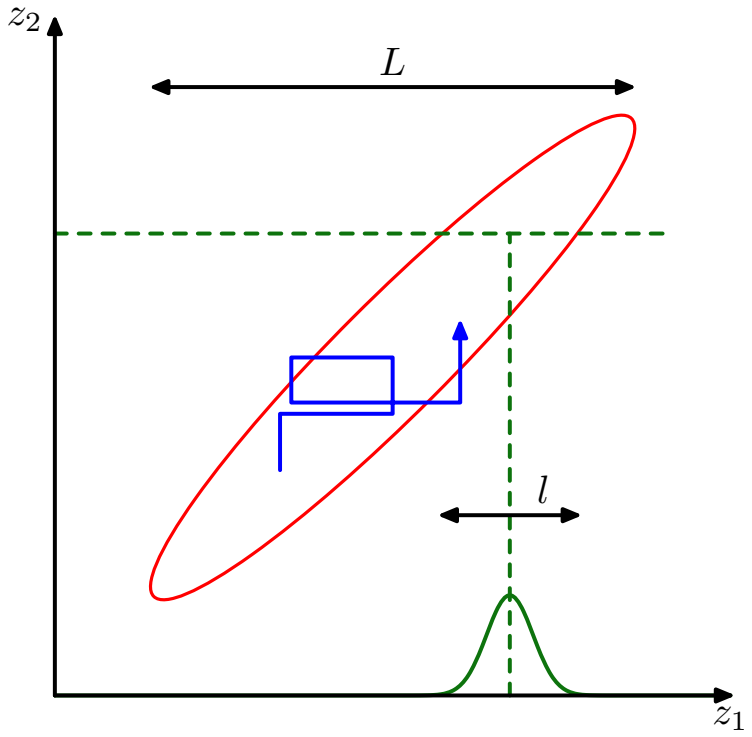


Figure from PRML, Bishop (2006)

Alternative Justification:

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.

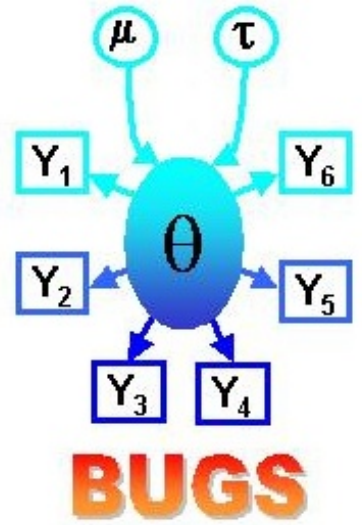
Gibbs Sampling Implementation

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

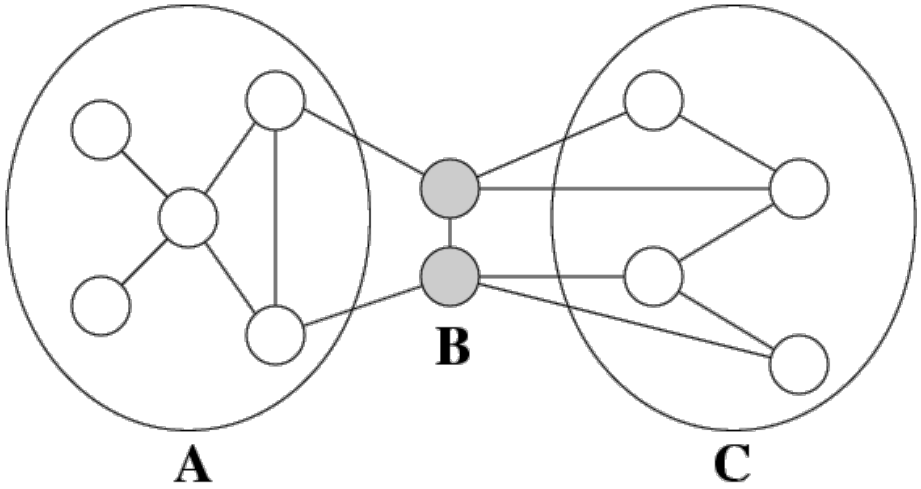
- Conditionals with a few discrete settings can be explicitly normalized:

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

- Continuous conditionals only univariate
  $\Rightarrow$ amenable to standard sampling methods.
  - Inverse CDF sampling
  - Rejection sampling
  - Slice sampling
  - ...



Undirected Graphical Models



Graph Separation

Conditional Independence

$$p(x_A, x_C \mid x_B) = p(x_A \mid x_B)p(x_C \mid x_B)$$

- This global Markov property implies a local Markov property:

$$p\left(x_i \mid x_{\mathcal{V}\setminus i}\right) = p\left(x_i \mid x_{\Gamma(i)}\right)$$

- Practical benefits of Gibbs sampling algorithm:
  - Model and algorithm have same modular structure
  - Conditionals can often be evaluated quickly, because they depend only on the neighboring nodes
  - Exponential families offer further efficiency improvements, by caching and recursively updating sufficient statistics