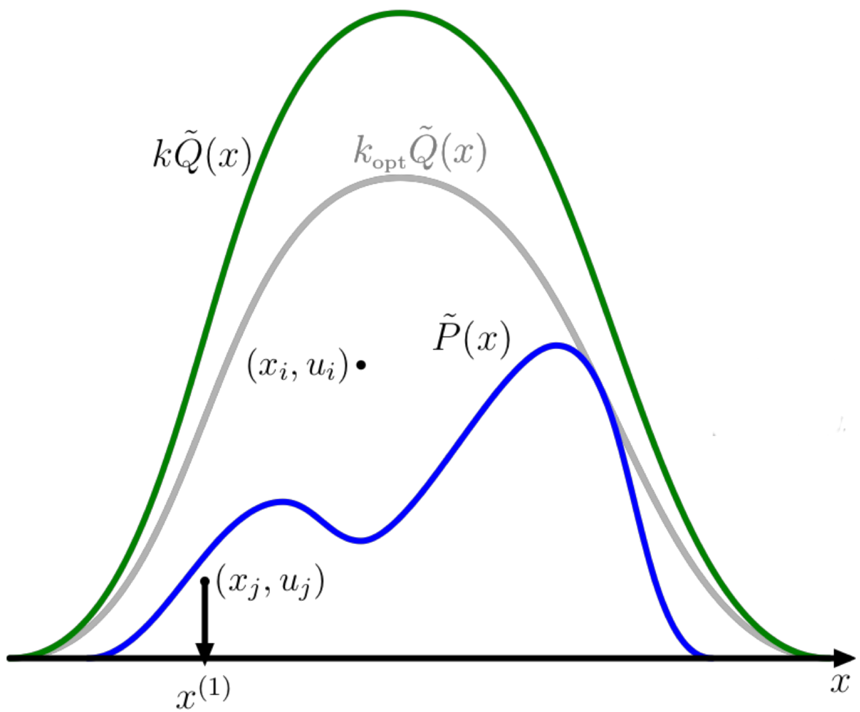
MCMC and Gibbs Sampling

Kayhan Batmanghelich

Approaches to inference

- Exact inference algorithms
  - The elimination algorithm
  - Message-passing algorithm (sum-product, belief propagation)
  - The junction tree algorithms
- Approximate inference techniques
  - Variational algorithms
    - Loopy belief propagation
    - Mean field approximation
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods

Recap: Rejection Sampling



Steps:

- Find Q(x) that is easy to sample from.
- Find k such that k such that:

$$\frac{\tilde{P}(x)}{kQ(x)} < 1$$

- Sample auxiliary variable y

$$\mathbb{P}(y = 1|x) = \frac{\tilde{P}(x)}{kQ(x)}$$

accept the sample with probability P(y=1|x)

Recap: Importance Sampling

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

$$\mathbb{E}_{x\sim p}\left[f(x)\right] = \int_x f(x)p(x) = \frac{\int_x f(x)\frac{\tilde{p}(x)}{\tilde{q}(x)}q(x)}{\int_x \frac{\tilde{p}(x)}{\tilde{q}(x)}q(x)}$$

Let $x^1, \cdots, x^L$ be samples from q(x).

$$\int_x f(x)p(x) \approx \frac{\sum_l f(x^l)\frac{\tilde{p}(x^l)}{\tilde{q}(x^l)}}{\sum_l \frac{\tilde{p}(x^l)}{\tilde{q}(x^l)}} = \sum_{l=1}^{L} f(x^l)w_l$$

Recap: Particle Filters

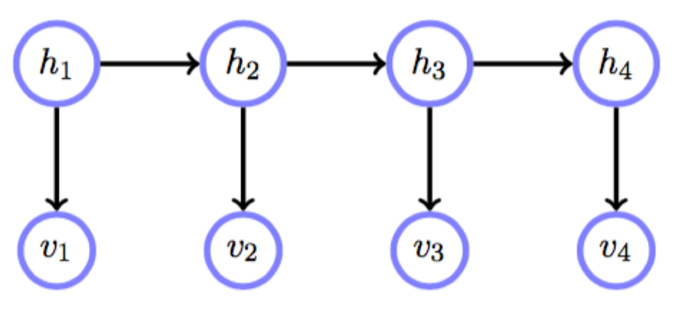
Apply the importance sampling to a temporal distribution $p(x_{1:t})$

General idea, change the proposal in each step: $q(x_t|x_{1:t})$

The recursion rule:

$$q(h_t|h_{1:t-1}) = p(h_t|h_{t-1})$$

$$\tilde{w}_t^l = \tilde{w}_{t-1}^l p(v_t|h_t^l)$$

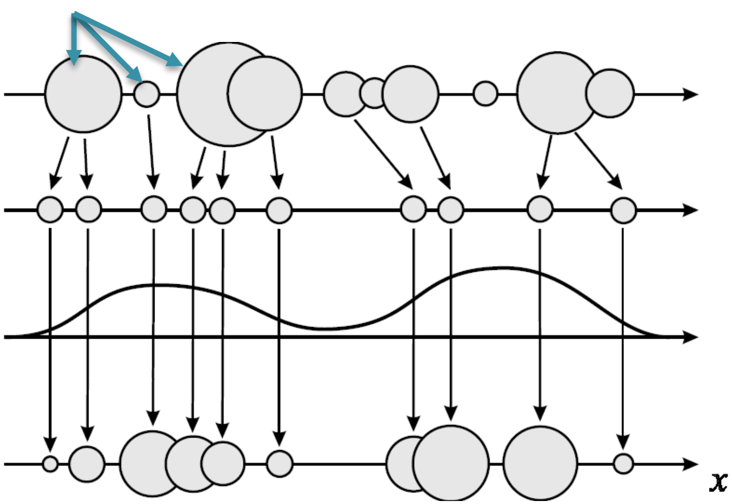


Forward message:

$$\rho(h_t) \propto p(h_t|v_{1:t})$$

$$\rho(h_t) \propto p(v_t|h_t) \int_{h_{t-1}} p(h_t|h_{t-1})\rho(h_{t-1})$$

$$\rho(h_t) \approx \frac{1}{Z} p(v_t|h_t) \sum_{l=1}^{L} p(h_t|h_{t-1}^l) w_{t-1}^l$$



Summary so far

<u>General ideas for the sampling approaches</u>

- Proposal distribution (q(x)): Use another distribution to sample from.
  - Change the proposal distribution with the iterations.
- Introduce an auxiliary variable to decide keeping a sample or not.
  - Why should we discard samples?
- Sampling from high-dimension is difficult.
  - Let's incorporate the graphical model into our sampling strategy.
- Can we use the gradient of the p(x)?

Summary so far

General ideas for the sampling approaches

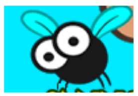
- Proposal distribution (q(x)): Use another distribution to sample from.
  - Change the proposal distribution with the iterations.

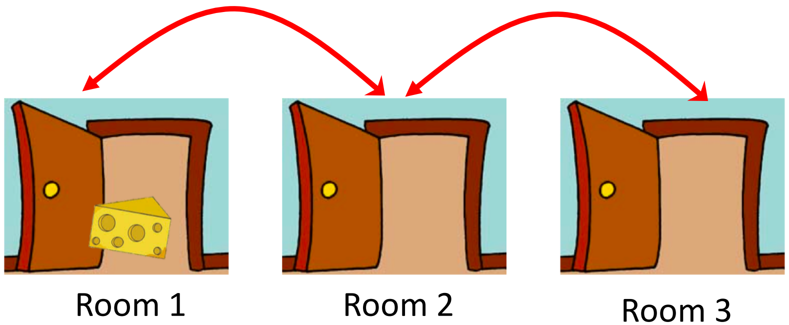
- Introduce an auxiliary variable to decide keeping a sample or not.
  - Why should we discard samples?

- Sampling from high-dimension is difficult.
  - Let's incorporate the graphical model into our sampling strategy.

- Can we use the gradient of the p(x)?

Random Walks of the Annoying Fly





$$p(x_{t+1} = i | x_t = j) = M_{ij}$$

$$\begin{bmatrix} 0.7 & 0.5 & 0 \\ 0.3 & 0.3 & 0.5 \\ 0 & 0.2 & 0.5 \end{bmatrix}$$

Stationary distribution:

$$M v_\infty = \boxed{v_\infty}$$

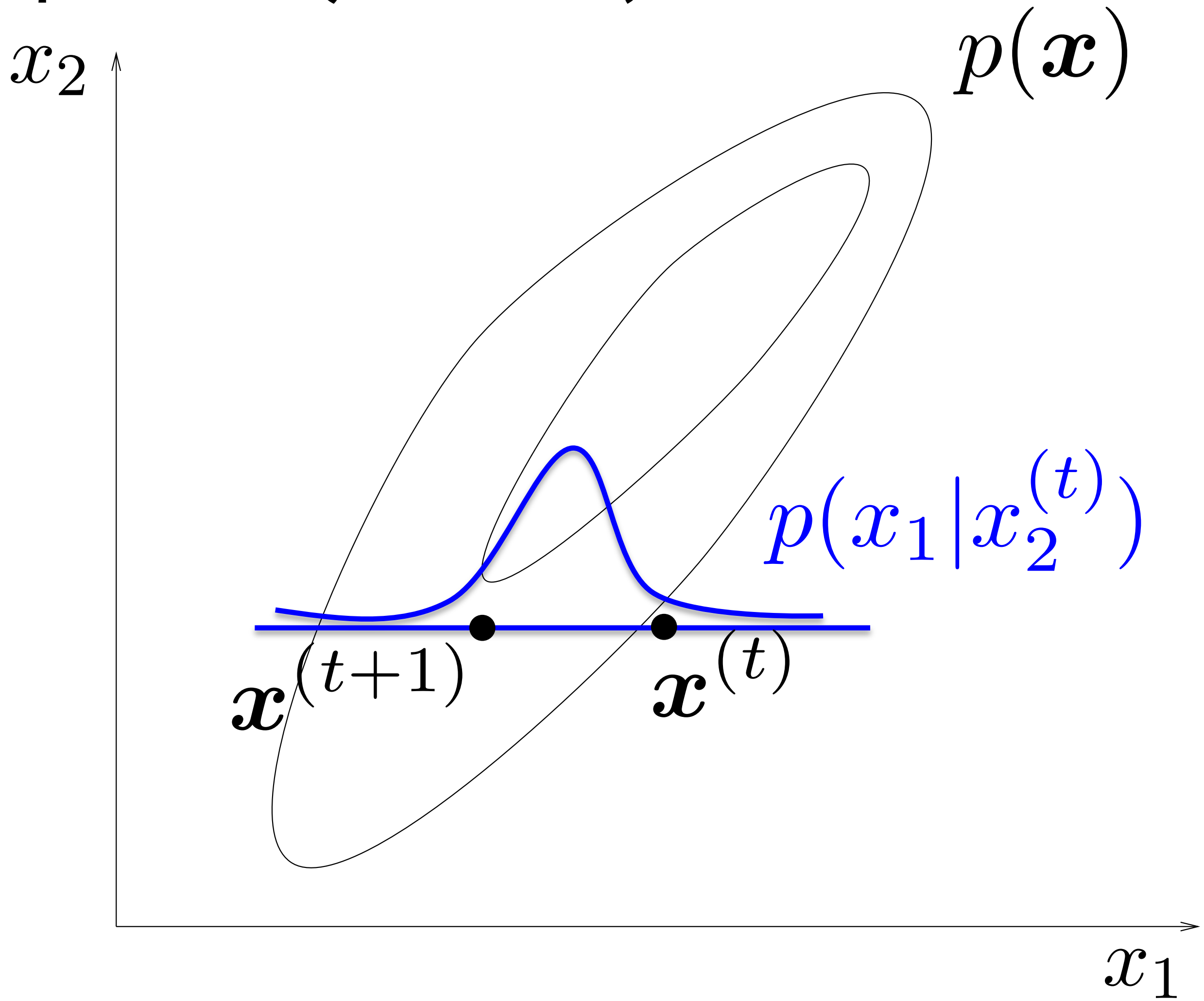
Eigen vector of the Markov matrix

Exploiting the structure

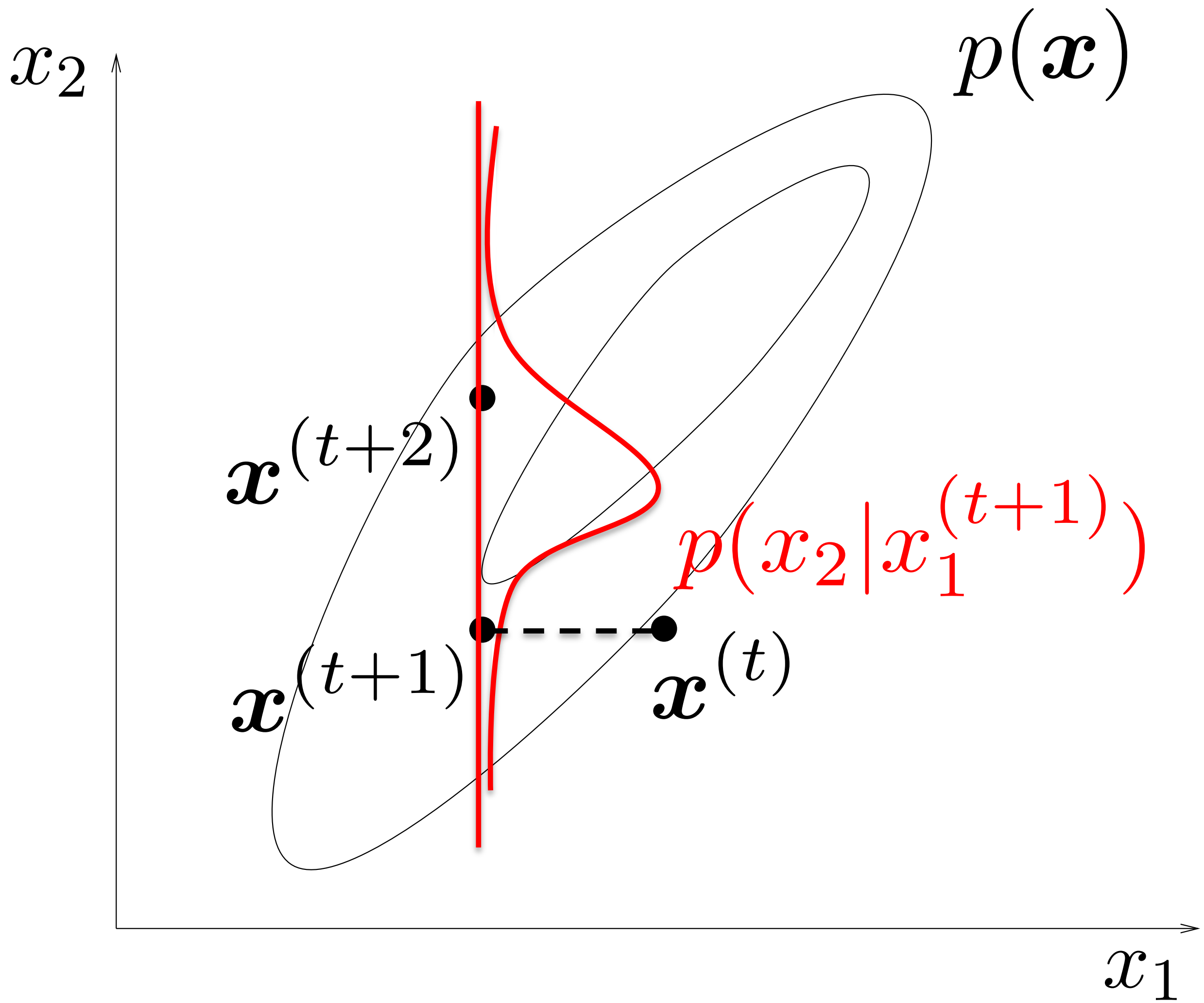
GIBBS SAMPLING

Gibbs Sampling

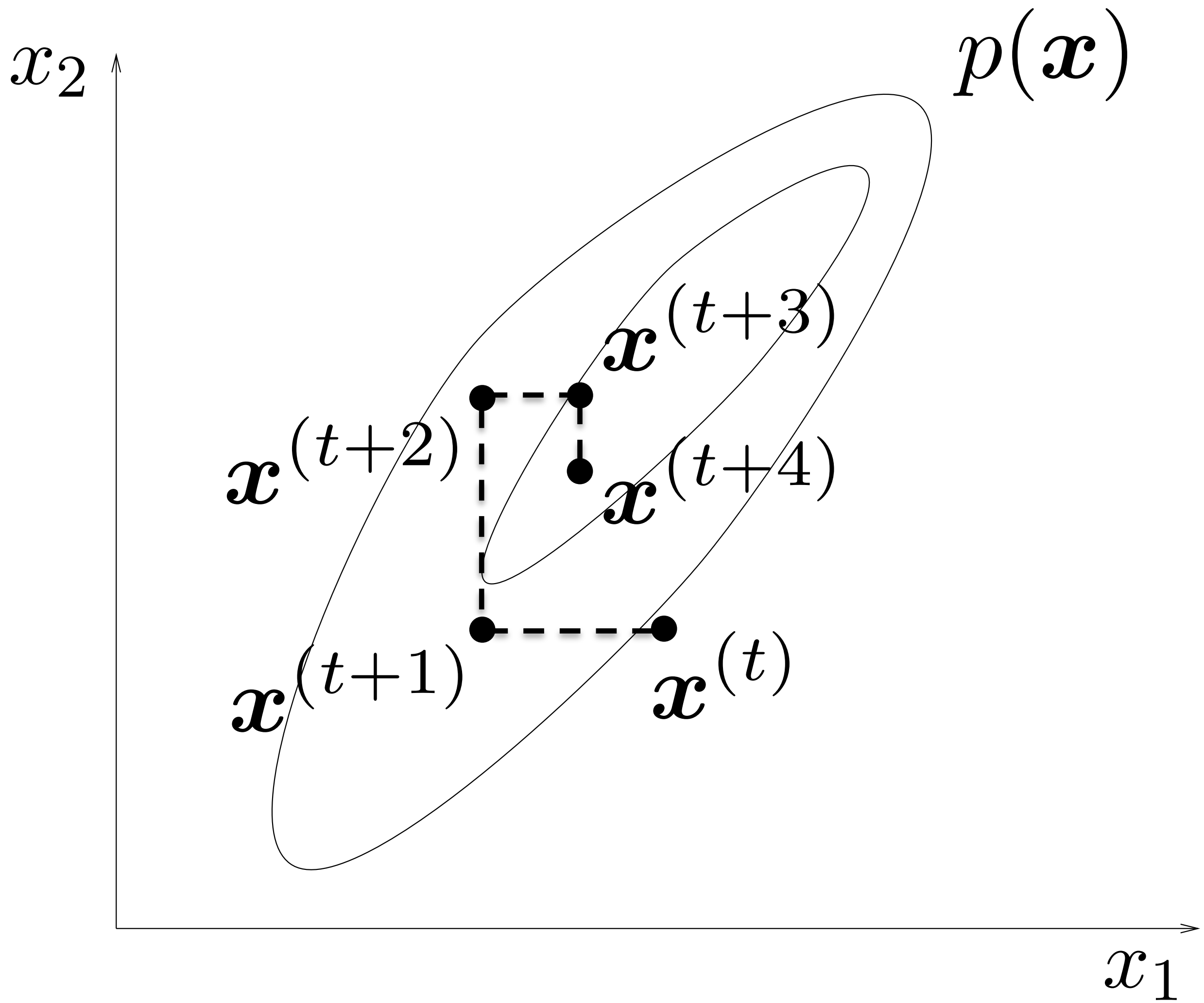
- Sample one (block of) variable at the time



Gibbs Sampling



Gibbs Sampling



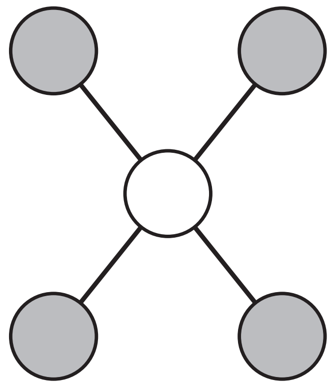
Gibbs Sampling

Link:
https://www.youtube.com/watch?v=AEwY6QXWoUg
https://www.youtube.com/watch?v=ZaKwpVgmKTY

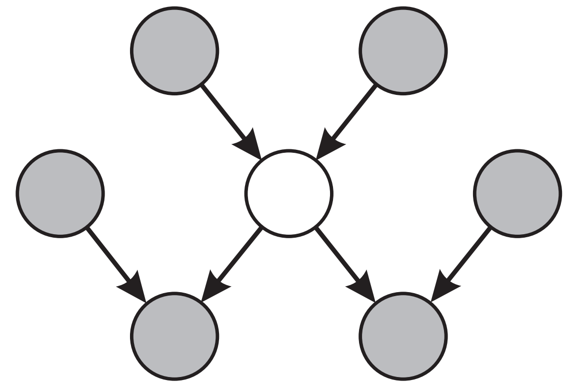
Ingredients for Gibb Recipe

Full conditionals only need to condition on the Markov Blanket

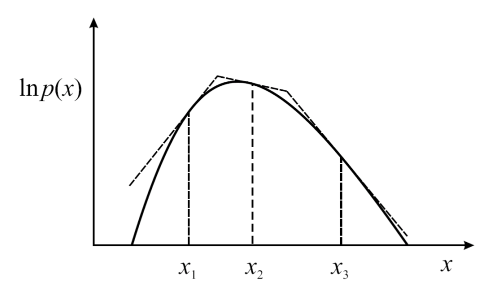
MRF



Bayes Net



- Must be "easy" to sample from conditionals



Gibbs Sampling

- Sample one (block of) variable at the time

$$p(x) = \boxed{p(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)} p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

$$p(x_i | x_{\setminus i}) = \frac{1}{Z} p(x_i | \mathrm{pa}(x_i)) \prod_{j \in \mathrm{ch}(i)} p(x_j | \mathrm{pa}(x_j))$$

Markov Blanket

Easy to compute

- The proposal distribution:

$$q(x^{l+1} | x^l, i) = p(x_i^{l+1} | x_{\setminus i}^l) \boxed{\prod_{j \neq i} \delta\left(x_j^{l+1}, x_j^l\right)}$$

Make sure other variables do not change

$$q(x^{l+1} | x^l) = \sum_i q(x^{l+1} | x^l, i) \boxed{q(i),}$$

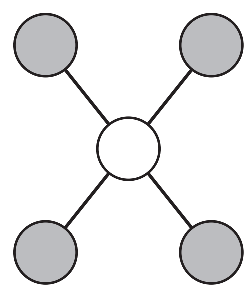
Choose one of the variables randomly with probability q(i)

Again….

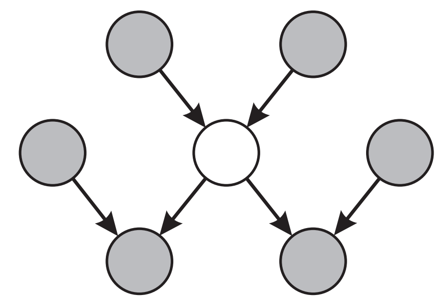
$$p(x_i|x_{\setminus i}) = \frac{1}{Z} p(x_i|\mathrm{pa}(x_i)) \prod_{j \in \mathrm{ch}(i)} p(x_j|\mathrm{pa}(x_j))$$

Full conditionals only need to condition on the Markov Blanket

MRF

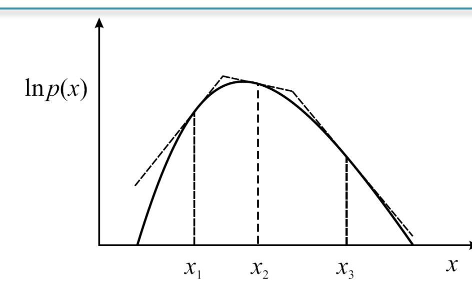


Bayes Net



$$p(x_i|x_{\setminus i}) = \frac{1}{Z} p(x_i|\mathrm{pa}(x_i)) \prod_{j \in \mathrm{ch}(i)} p(x_j|\mathrm{pa}(x_j))$$

- Must be "easy" to sample from conditionals
- Many conditionals are log-concave and are amenable to adaptive rejection sampling



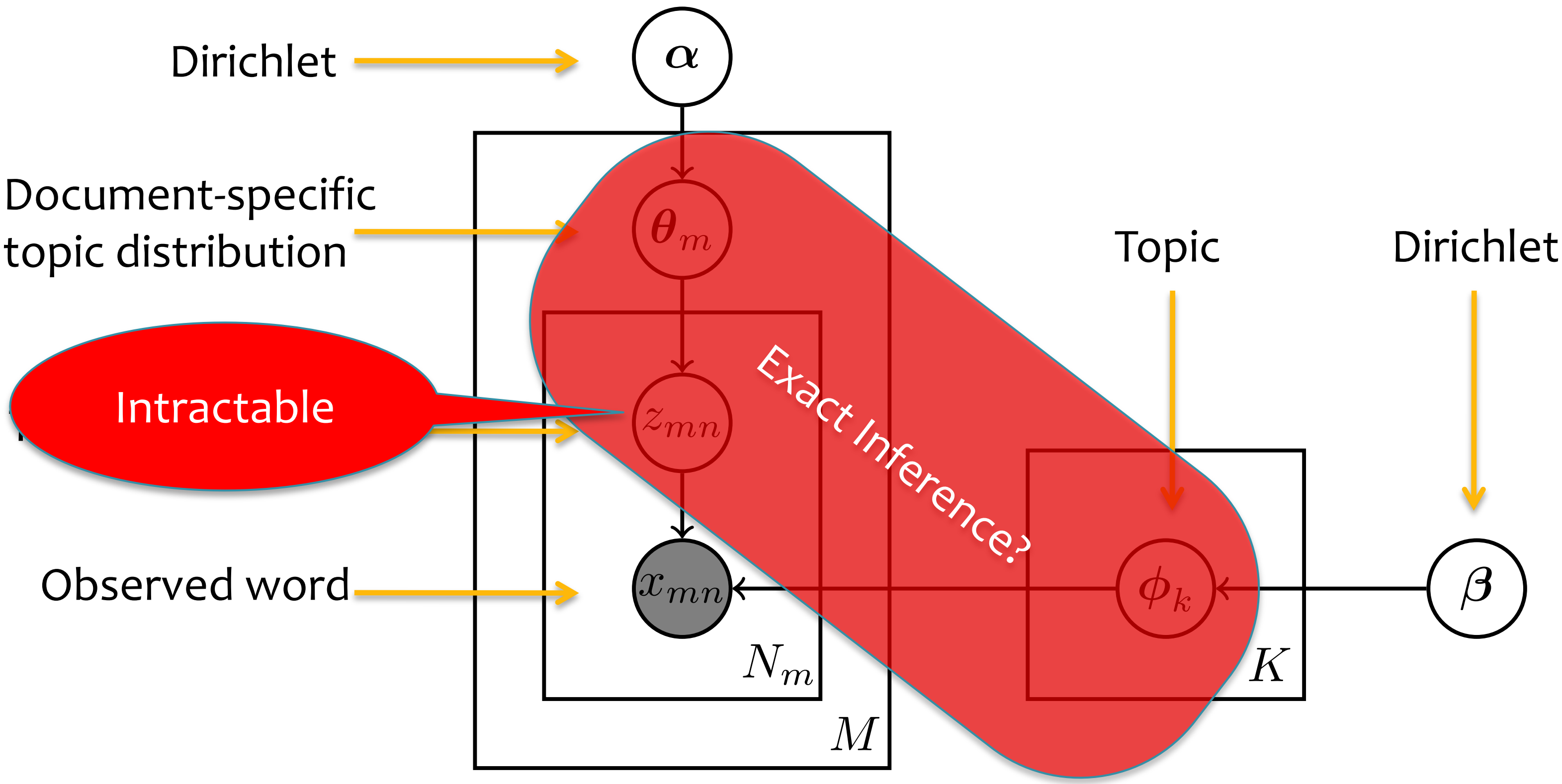
Whiteboard

- Gibbs Sampling as M-H

Case Study: LDA

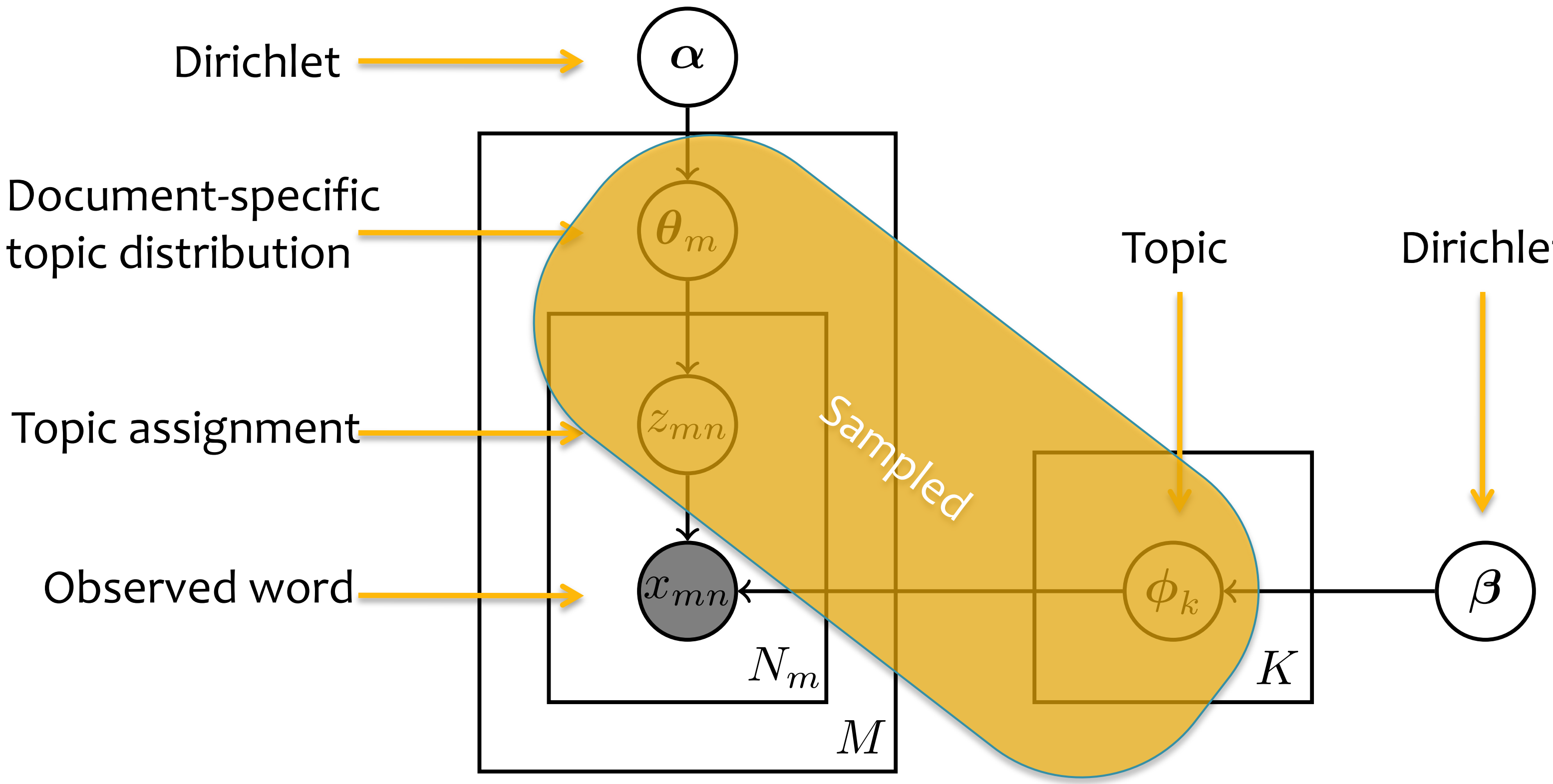
LDA Inference

- Bayesian Approach



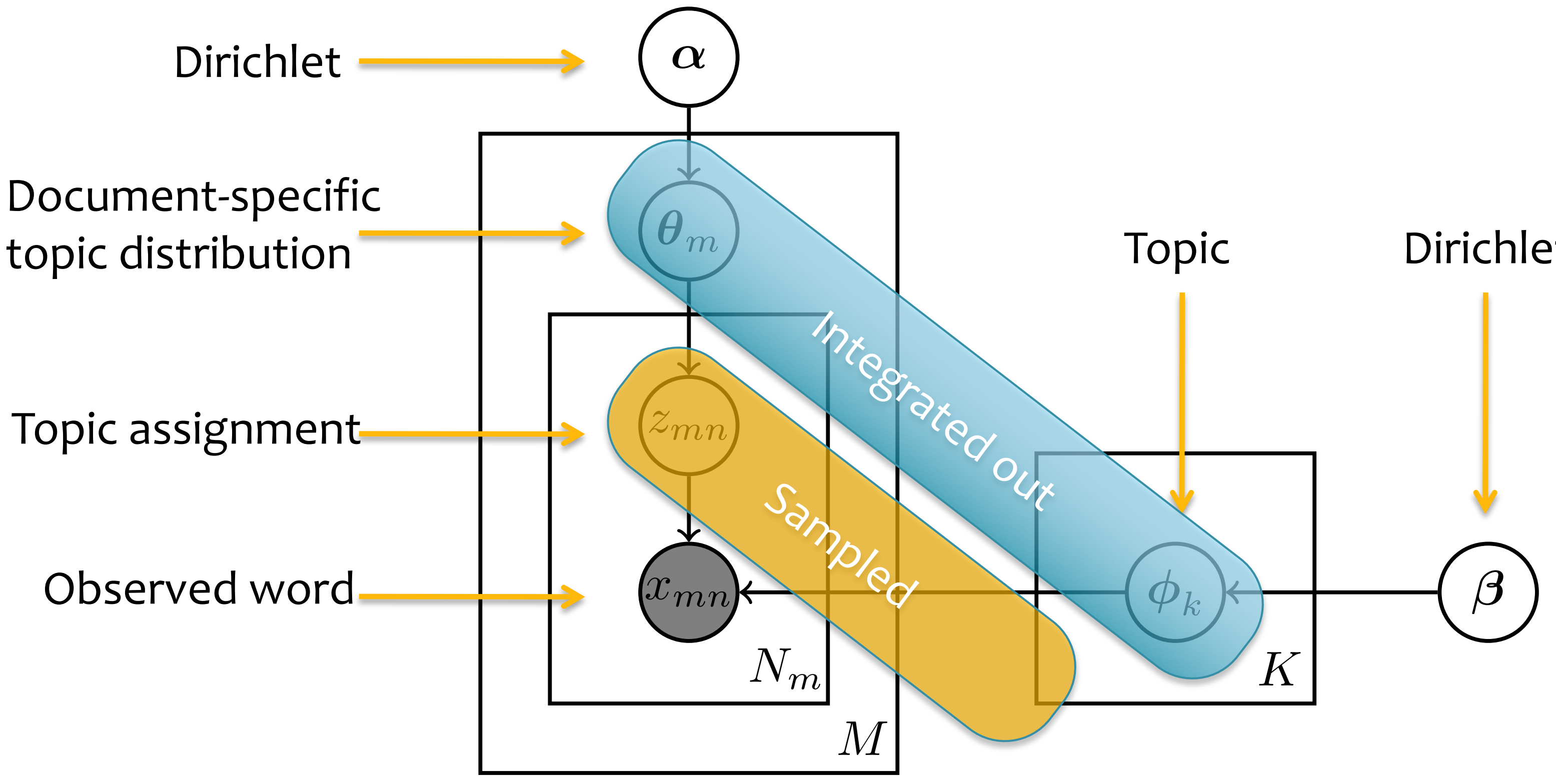
LDA Inference

- Explicit Gibbs Sampler



LDA Inference

- Collapsed Gibbs Sampler



Sampling

Goal:

- Draw samples from the posterior $p(Z|X, \alpha, \beta)$
- Integrate out topics ϕ and document-specific distribution over topics θ

Algorithm:

- While not done…
  - For each document, m:
    - For each word, n:
      - » Resample a single topic assignment using the full conditionals for z_{mn}

Sampling

- What queries can we answer with samples of z_{mn}?
  - Mean of z_{mn}
  - Mode of z_{mn}
  - Estimate posterior over z_{mn}
  - Estimate of topics ϕ and document-specific distribution over topics θ

Gibbs Sampling for LDA

- Full conditionals

$$p(z_i = j | z_{-i}, X, \alpha, \beta) \propto \frac{n_{-i,j}^{(x_i)} + \beta}{n_{-i,j}^{(\cdot)} + T\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}$$

$n_{-i,j}^{(x_i)}$ the number of instances of word x_i assigned to topic j, not including current word.

$n_{-i,j}^{(\cdot)}$ total number of words assigned to topic j, not including the current one.

$n_{-i,j}^{(d_i)}$ the number of words for document d_i assigned to topic j.

$n_{-i,\cdot}^{(d_i)}$ total number of words in the document d_i not including the current one.

Link: https://people.cs.umass.edu/~wallach/courses/s11/cmpsci791ss/readings/griffiths02gibbs.pdf

Gibbs Sampling for LDA

- Sketch of the derivation of the full conditionals

$$
\begin{aligned}
p(z_i = k | Z^{-i}, X, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{p(X, Z|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(X, Z^{-i}|\boldsymbol{\alpha}, \boldsymbol{\beta})} \\
&\propto p(X, Z|\boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= p(X|Z, \boldsymbol{\beta}) p(Z|\boldsymbol{\alpha}) \\
&= \int_{\Phi} p(X|Z, \Phi) p(\Phi|\boldsymbol{\beta}) \, d\Phi \int_{\Theta} p(Z|\Theta) p(\Theta|\boldsymbol{\alpha}) \, d\Theta \\
&= \left(\prod_{k=1}^{K} \frac{B(\vec{n}_k + \boldsymbol{\beta})}{B(\boldsymbol{\beta})} \right) \left(\prod_{m=1}^{M} \frac{B(\vec{n}_m + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \right)
\end{aligned}
$$

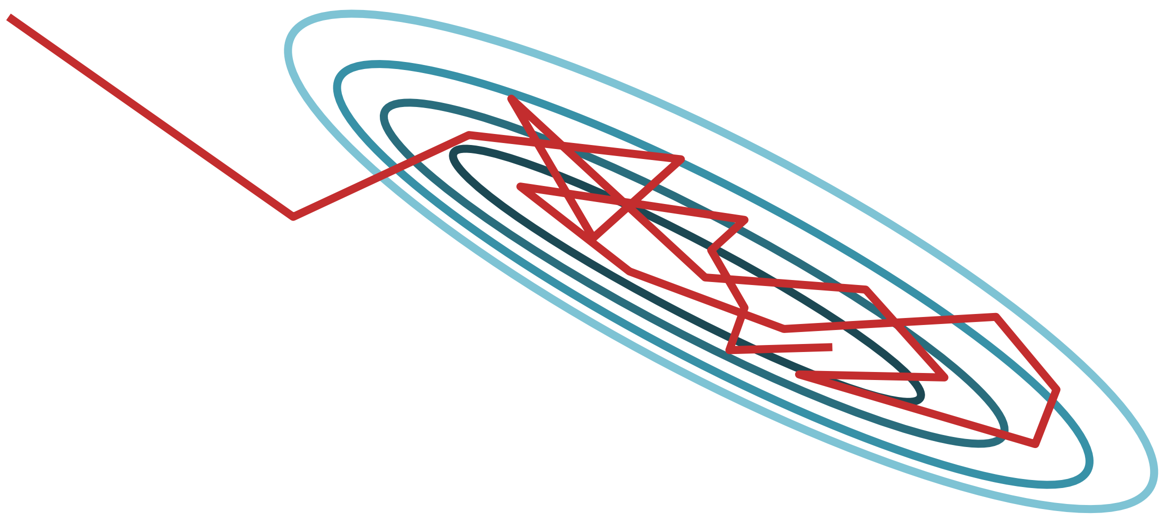
$$
p(z_i = j | z_{-i}, X, \alpha, \beta) \propto \frac{n_{-i,j}^{(x_i)} + \beta}{n_{-i,j}^{(\cdot)} + T\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}
$$

Definitions and Theoretical Justification for MCMC

MARKOV CHAINS

MCMC

- **Goal:** Draw approximate, correlated samples from a target distribution p(x)
- **MCMC:** Performs a biased random walk to explore the distribution



Simulations of MCMC

Visualization of Metroplis-Hastings, Gibbs Sampling, and Hamiltonian MCMC:

https://www.youtube.com/watch?v=Vv3f0QNWvWQ

Metropolis-Hastings Sampling

- Consider this mixture for the proposal

$$q(x'|x) = \tilde{q}(x'|x) f(x', x) + \delta(x', x)\left(1 - \int_{x''} \tilde{q}(x''|x) f(x'', x)\right)$$

Stay where you are Or ....

Sample from a distribution depends on current location (x)

Of course! It should be a proper density!

Choose between those options with a probability that depends on a proposed point and current point ($0 \leq f(x', x) \leq 1$)

Metropolis-Hastings Sampling

- Consider this mixture for the proposal

$$q(x'|x) = \tilde{q}(x'|x)f(x',x) + \delta(x',x)\left(1 - \int_{x''} \tilde{q}(x''|x)f(x'',x)\right)$$

- Is it a proper density?

$$\int_{x'} q(x'|x) = \int_{x'} \tilde{q}(x'|x)f(x',x) + 1 - \int_{x''} \tilde{q}(x''|x)f(x'',x) = 1$$

Metropolis-Hastings Sampling

- Consider this mixture for the proposal

$$q(x'|x) = \tilde{q}(x'|x)f(x',x) + \delta(x',x)\left(1 - \int_{x''} \tilde{q}(x''|x)f(x'',x)\right)$$

- How to choose $f(x',x)$ and $\tilde{q}(x'|x)$?

p(x) must be the Stationary distribution

$$p(x') = \int_x q(x'|x)p(x)$$

$$\int_x \tilde{q}(x'|x)f(x',x)p(x) = \int_x \tilde{q}(x|x')f(x,x')p(x')$$

Designing $f(x', x)$

$$\int_x \tilde{q}(x'|x) f(x', x) p(x) = \int_x \tilde{q}(x|x') f(x, x') p(x')$$

- MH acceptance function:

$$f(x', x) = \min\left(1, \frac{\tilde{q}(x|x')p(x')}{\tilde{q}(x'|x)p(x)}\right)$$

Designing $f(x', x)$

- MH acceptance function:

$$f(x', x) = \min\left(1, \frac{\tilde{q}(x|x')p(x')}{\tilde{q}(x'|x)p(x)}\right)$$

- Detailed Balance:

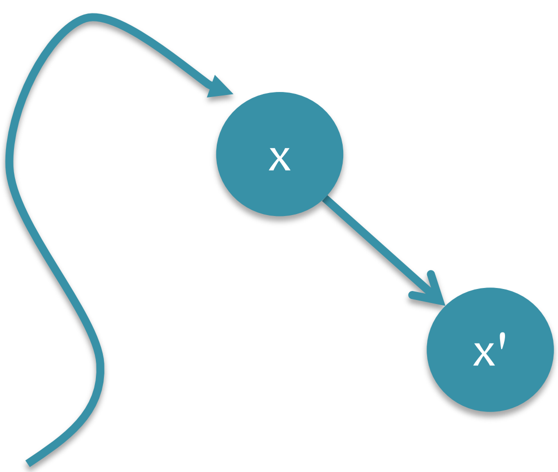
$$f(x', x)\tilde{q}(x'|x)p(x) = f(x, x')\tilde{q}(x|x')p(x')$$

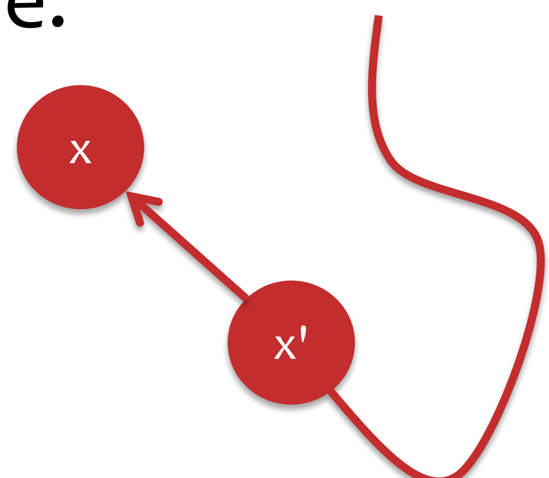
Detailed Balance

$$f(x', x)\tilde{q}(x'|x)p(x) = f(x, x')\tilde{q}(x|x')p(x')$$

Detailed balance means that, for each pair of states x and x',

arriving at x then x' and arriving at x' then x are equiprobable.





A choice for $q(x'|x)$

- For Metropolis-Hastings, a generic proposal distribution is: $q(x|x^{(t)}) = \mathcal{N}(0, \epsilon^2)$
- If ϵ is large, many rejections
- If ϵ is small, slow mixing

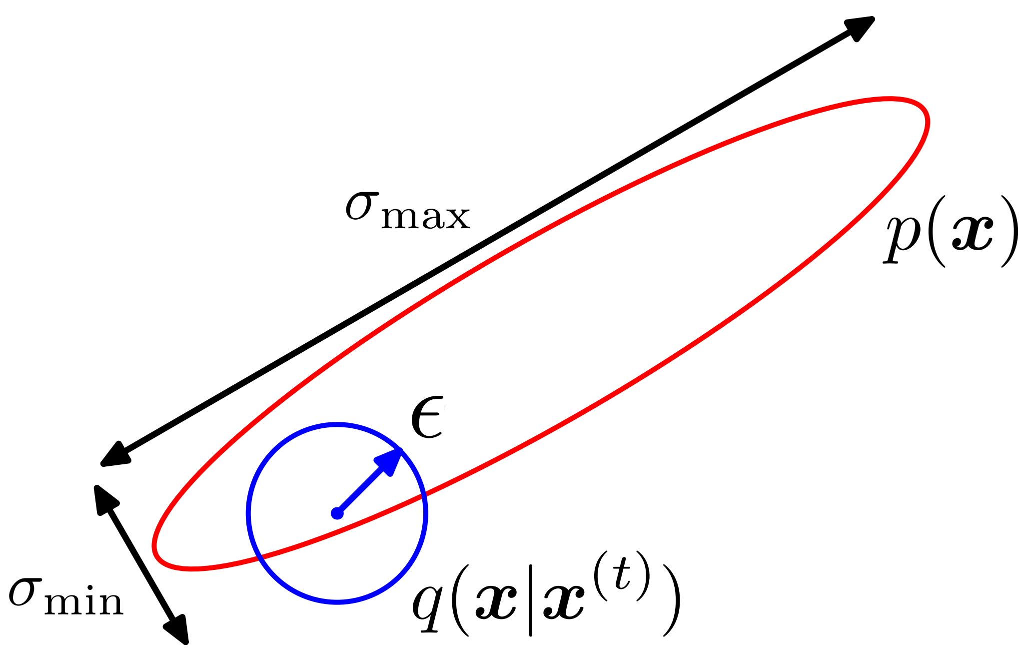
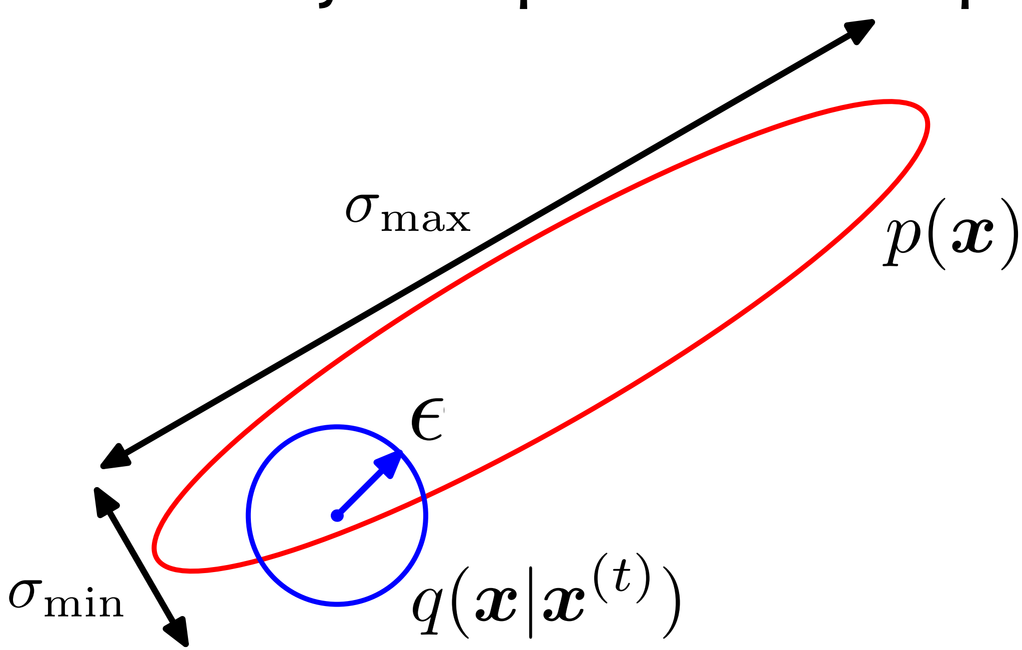


Figure from Bishop (2006)

A choice for $q(x'|x)$

- For Rejection Sampling, the accepted samples are are **independent**
- But for Metropolis-Hastings, the samples are **correlated**
- **Question:** How long must we wait to get effectively independent samples?



A: independent states in the M-H random walk are separated by roughly $(\sigma_{\max}/\sigma_{\min})^2$ steps

Metropolis-Hastings Algorithm

Choose a starting point x^1.
for $i = 2$ to L **do**
    Draw a candidate sample x^{cand} from the proposal $\tilde{q}(x'|x^{l-1})$.
    Let $a = \frac{\tilde{q}(x^{l-1}|x^{cand})p(x^{cand})}{\tilde{q}(x^{cand}|x^{l-1})p(x^{l-1})}$
    if $a \geq 1$ **then** $x^l = x^{cand}$
    else
        draw a random value u uniformly from the unit interval $[0, 1]$.
        if $u < a$ **then** $x^l = x^{cand}$
        else
            $x^l = x^{l-1}$
        end if
    end if
end for

Practical Issues

- **Question:** Is it better to move along one dimension or many?

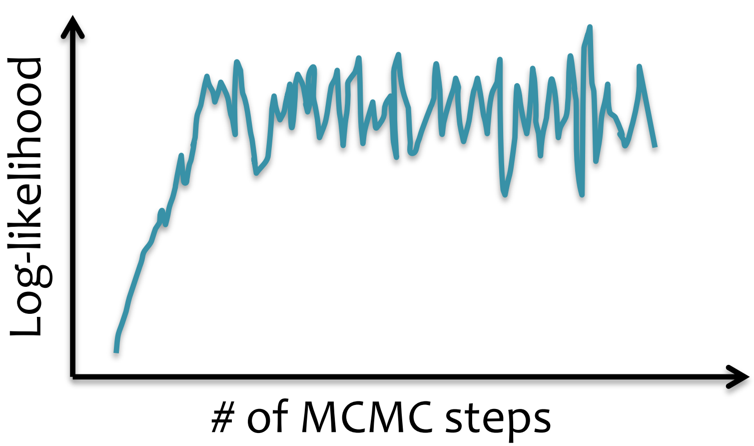
- **Answer:** For Metropolis-Hasings, it is sometimes better to sample one dimension at a time
  - Q: Given a sequence of 1D proposals, compare rate of movement for **one-at-a-time** vs. **concatenation.**

- **Answer:** For Gibbs Sampling, sometimes better to sample a block of variables at a time
  - Q: When is it tractable to sample a block of variables?

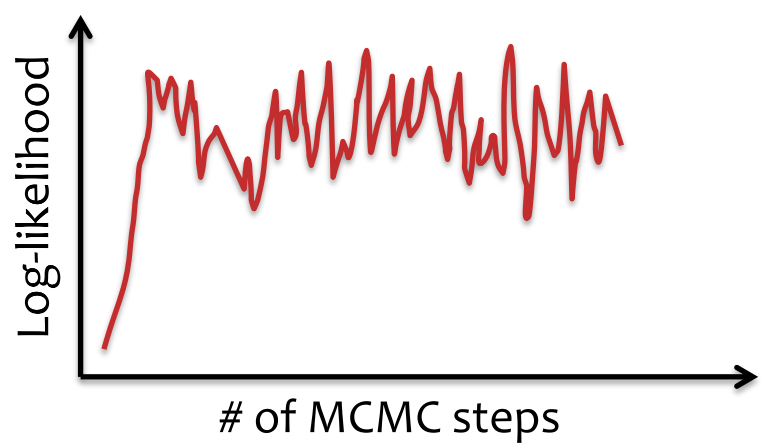
Practical Issues

- **Question:** How do we assess convergence of the Markov chain?
- **Answer:** It's not easy!
  - Compare statistics of multiple independent chains
  - Ex: Compare log-likelihoods

Chain 1

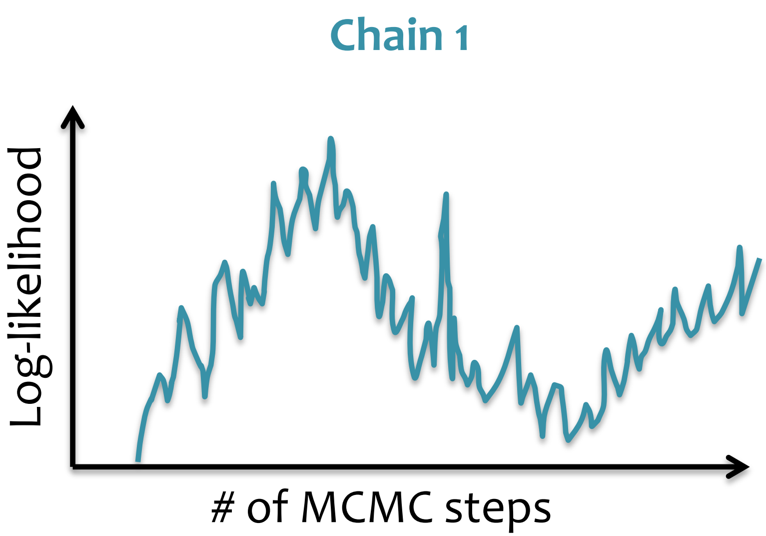


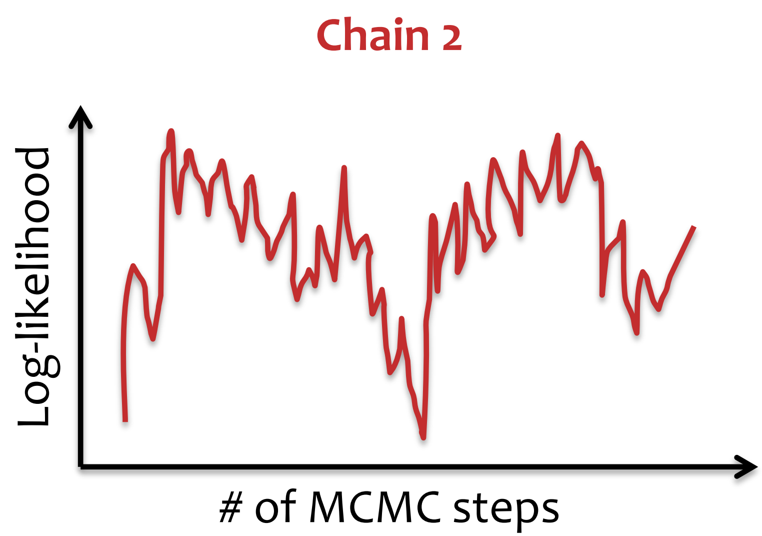
Chain 2



Practical Issues

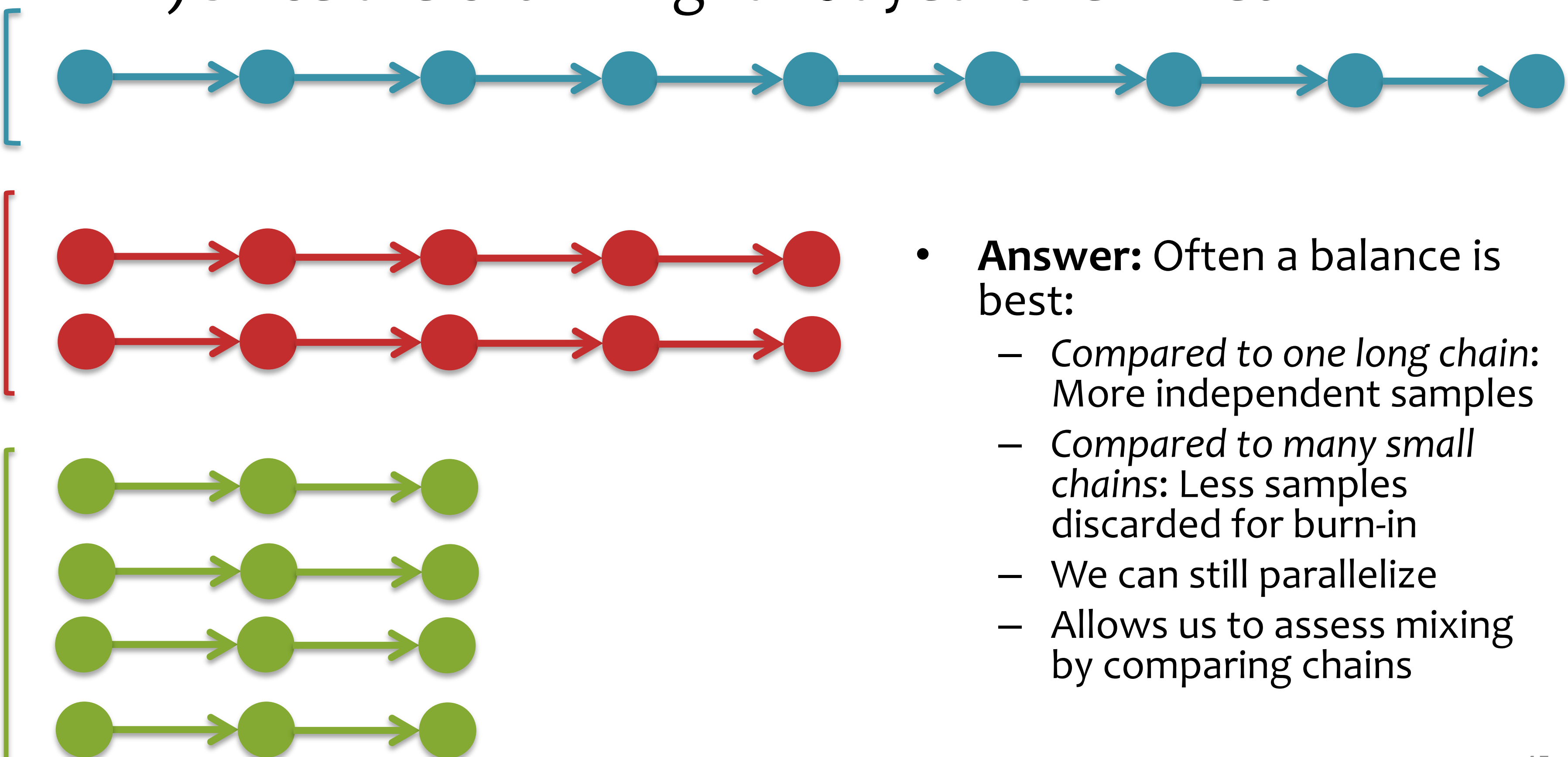
- **Question:** How do we assess convergence of the Markov chain?
- **Answer:** It's not easy!
  - Compare statistics of multiple independent chains
  - Ex: Compare log-likelihoods





Practical Issues

- **Question:** Is one long Markov chain better than many short ones?
- **Note:** typical to discard initial samples (aka. "burn-in") since the chain might not yet have mixed



- **Answer:** Often a balance is best:
  - *Compared to one long chain*: More independent samples
  - *Compared to many small chains*: Less samples discarded for burn-in
  - We can still parallelize
  - Allows us to assess mixing by comparing chains