Computer Science

CSC 665-1: Advanced Topics in Probabilistic Graphical Models

Monte Carlo Methods

Instructor: Prof. Jason Pacheco

Inference (and related) Tasks

- Simulation: $x \sim p(x) = \frac{1}{Z} f(x)$

- Compute expectations: $\mathbb{E}[\phi(x)] = \int p(x)\phi(x)\, dx$

- Optimization: $x^* = \arg\max_x f(x)$

- Compute normalizer: $Z = \int f(x)\, dx$

Inference (and related) Tasks

- Simulation: $x \sim p(x) = \frac{1}{Z} f(x)$

- Compute expectations: $\mathbb{E}[\phi(x)] = \int p(x)\phi(x)\,dx$

- Optimization: $x^* = \arg\max_x f(x)$

- Compute normalizer: $Z = \int f(x)\,dx$

Monte Carlo Integration

Estimate expectation over samples:

$$\hat{\phi} = \frac{1}{R}\sum_{r=1}^{r} \phi(x^{(r)}) \approx \mathbb{E}_p[\phi(x)], \quad \text{where } \{x^{(r)}\} \sim p(x)$$

How good is an estimate with R samples?

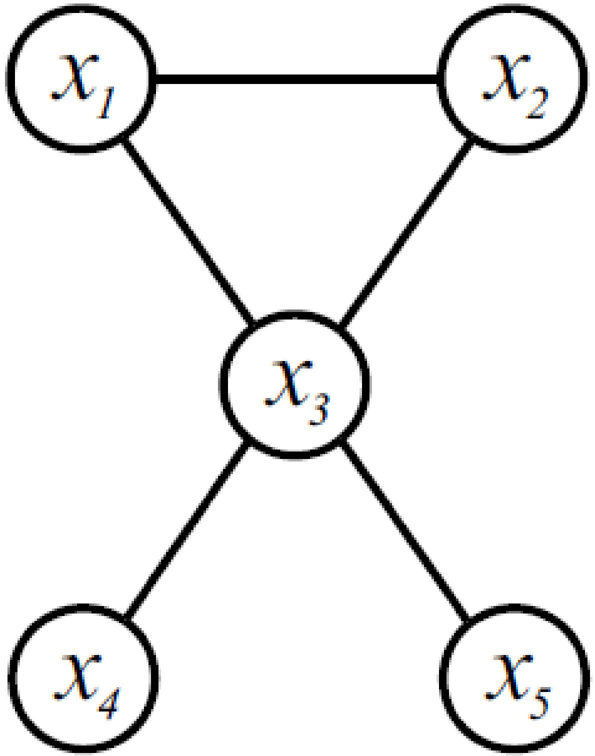
- Unbiased: $\mathbb{E}[\hat{\phi}] = \mathbb{E}[\phi]$
- Variance reduces at rate 1/R: $\text{var}(\hat{\phi}) = \frac{\text{var}(\phi)}{R}$

Variance independent of dimensionality of X

Markov Random Field

Consider the (pairwise) Markov Random Field :

$$p(x) = \frac{1}{Z} \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t)$$

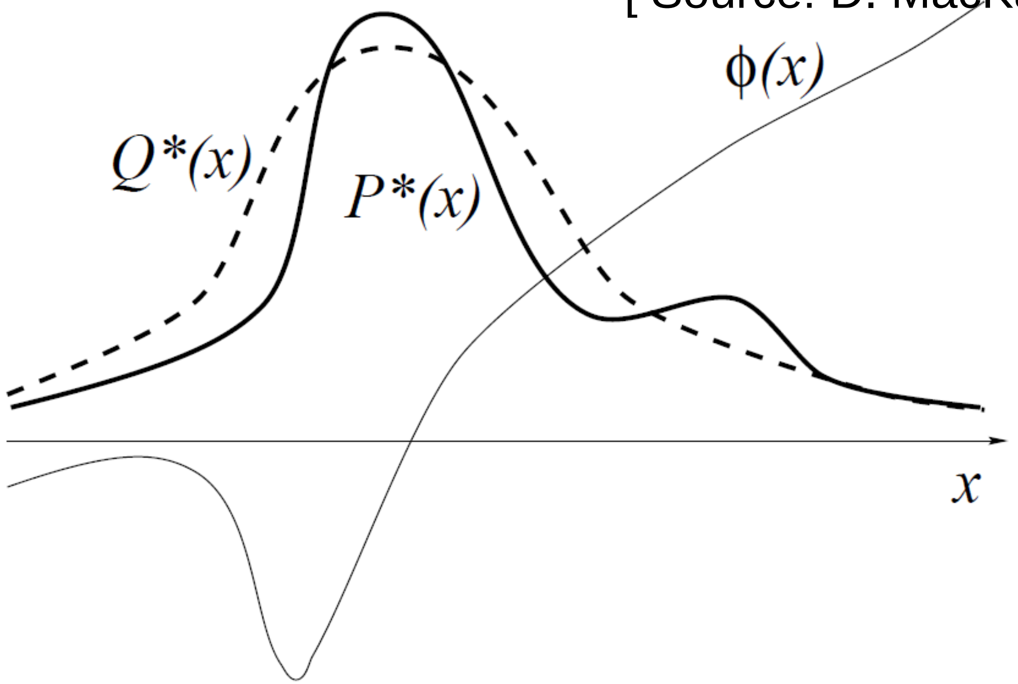


Specified up to unknown normalizer Z e.g.

$$p(x) = \frac{1}{Z} f(x)$$

Direct simulation is non-trivial in general…

Importance Sampling



Simulate from tractable distribution:

$$\{x^{(r)}\}_{r=1}^{R} \sim q(x)$$

Rewrite expectation:

$$\mathbb{E}_p[\phi(x)] = \int \frac{q(x)p(x)}{q(x)}\phi(x)\,dx$$

$$= \frac{1}{Z}\mathbb{E}_q\left[\frac{f(x)}{q(x)}\phi(x)\right]$$

$$\approx \sum_r \bar{w}_r \phi(x^{(r)})$$

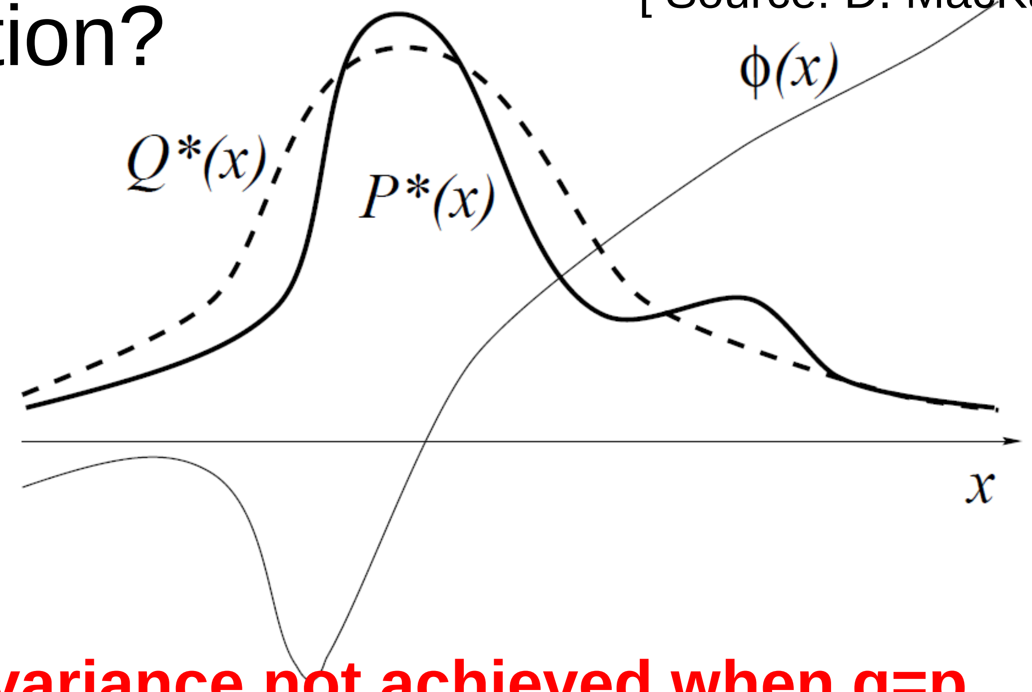
Normalized importance weights calculated without knowing Z:

$$w_r = \frac{f(x^{(r)})}{q(x^{(r)})} \qquad \bar{w}_r = \frac{w_r}{\sum_{r'} w_{r'}}$$

Unnormalized **Normalized**

Importance Sampling

[Source: D. MacKay]



Q: What is a good proposal distribution?

A: Minimize estimator variance

$$q^* = \arg\min_q \, \mathrm{var}_q(\hat{\phi})$$

minimum variance obtained when,

$$q^* \propto |\phi(x)| p(x)$$

Minimum variance not achieved when q=p

- Estimator variance scales catastrophically with dimension:

e.g. for N-dim. X and Gaussian q(x): $$\frac{w_r^{\max}}{w_r^{\mathrm{med}}} = \exp(\sqrt{2N})$$

Inference (and related) Tasks

- Simulation: $x \sim p(x) = \frac{1}{Z} f(x)$
- Compute expectations: $\mathbb{E}[\phi(x)] = \int p(x)\phi(x)\, dx$
- Optimization: $x^* = \arg\max_x f(x)$
- Compute normalizer: $Z = \int f(x)\, dx$

Markov Chain Monte Carlo (MCMC)

- Stochastic 1st order Markov process with transition kernel:

$$T(x^{(t)} \mid x^{(t-1)})$$

$$\cdots \longrightarrow x^{(t-1)} \longrightarrow x^{(t)} \longrightarrow x^{(t+1)} \longrightarrow \cdots$$

- Each $x^{(t)}$ full N-dimensional state vector
- MCMC samples $\ldots, x^{(t-1)}, x^{(t)}, x^{(t+1)}, \ldots$ **not independent**
- New superscript notation indicates dependence:

$\{x^{(r)}\}_{r=1}^{R}$

Independent

$\{x^{(t)}\}_{t=1}^{T}$

Dependent

Key Question: How many MCMC samples T are needed to draw R independent samples from p(x)?

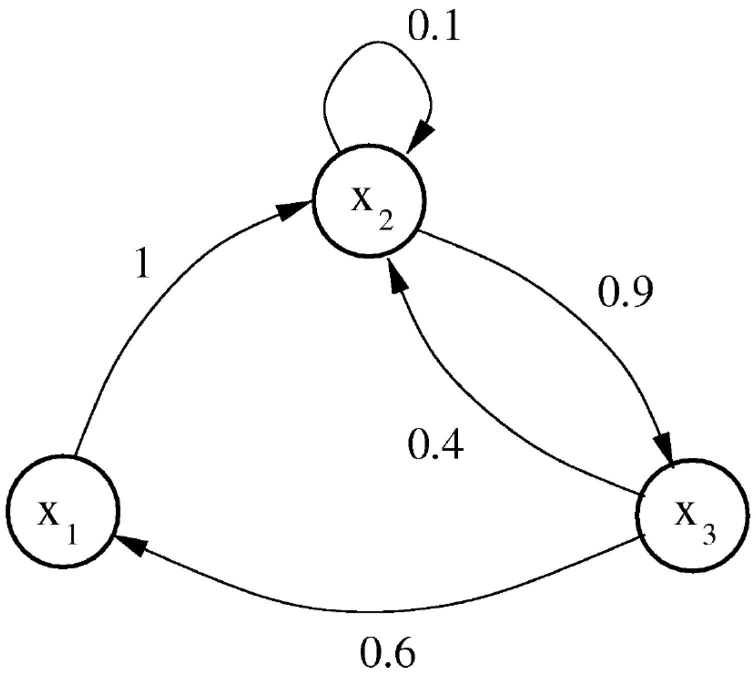
Markov Chain Monte Carlo (MCMC)

- Stochastic 1st order Markov process with transition kernel:

$$T(x^{(t)} \mid x^{(t-1)})$$

$$\cdots \longrightarrow x^{(t-1)} \longrightarrow x^{(t)} \longrightarrow x^{(t+1)} \longrightarrow \cdots$$

E.g. Let, $T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix}$



[Source: Andrieu et al.]

- Initial state dist'n: $\mu(x^{(1)}) = (0.5, 0.2, 0.3)$
- Repeated transitions converge to target

$$\mu(x^{(1)})T \cdot T \cdot \ldots \cdot T = (0.2, 0.4, 0.4) = p(x)$$

True for <u>any</u> initial state distribution

MCMC Theory

For any starting point chain converges to target $p(x)$ if T obeys:

- ***Aperiodicity*: ** Chain should not get trapped in cycles
- ***Irreducibility*: ** For any state $x \in \mathcal{X}$ there is positive probability of visiting any other state $x' \in \mathcal{X}$ in finite steps
- **Ergodicity:** Chain is *ergodic* if it is irreducible and aperiodic

Detailed Balance Sufficient (not necessary) condition:

$$p(x^{(t)})T(x^{(t-1)} \mid x^{(t)}) = p(x^{(t-1)})T(x^{(t)} \mid x^{(t-1)})$$

Summing over states yields target distribution:

$$p(x^{(t)}) = \sum_{x^{(t-1)}} p(x^{(t-1)})T(x^{(t)} \mid x^{(t-1)})$$

MCMC Theory

For any starting point chain converges to target $p(x)$ if T obeys:

- ***Aperiodicity*:** Chain should not get trapped in cycles
- ***Irreducibility*:** For any state $x \in \mathcal{X}$ there is positive probability of visiting any other state $x' \in \mathcal{X}$ in finite steps
- **Ergodicity:** Chain is *ergodic* if it is irreducible and aperiodic

Detailed Balance Sufficient (not necessary) condition:

$$p(x^{(t)})T(x^{(t-1)} \mid x^{(t)}) = p(x^{(t-1)})T(x^{(t)} \mid x^{(t-1)})$$

Summing over states yields target distribution:

$$p(x^{(t)}) = \sum_{x^{(t-1)}} p(x^{(t-1)})T(x^{(t)} \mid x^{(t-1)})$$

p(x) is eigenvector with largest eigenvalue 1

Metropolis-Hastings

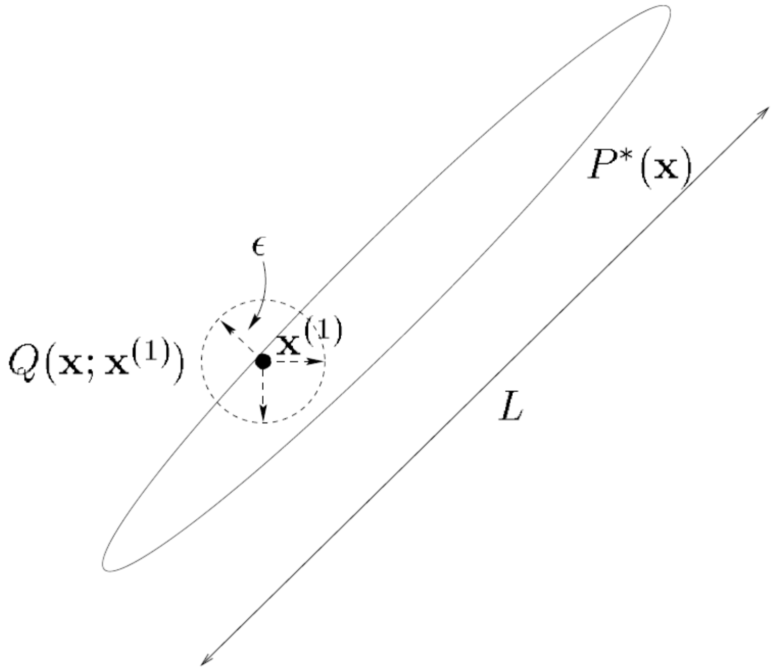
[Source: D. MacKay]

Transition kernel with target distribution:

$$p(x) = 1/Z f(x)$$

1. Sample proposal: $x' \mid x^{(t-1)} \sim q(\cdot)$
2. Accept with probability:

$$\min\{1, a\} \quad \text{where} \quad a = \frac{f(x')}{f(x^{(t-1)})} \frac{q(x^{(t-1)} \mid x')}{q(x' \mid x^{(t-1)})}$$



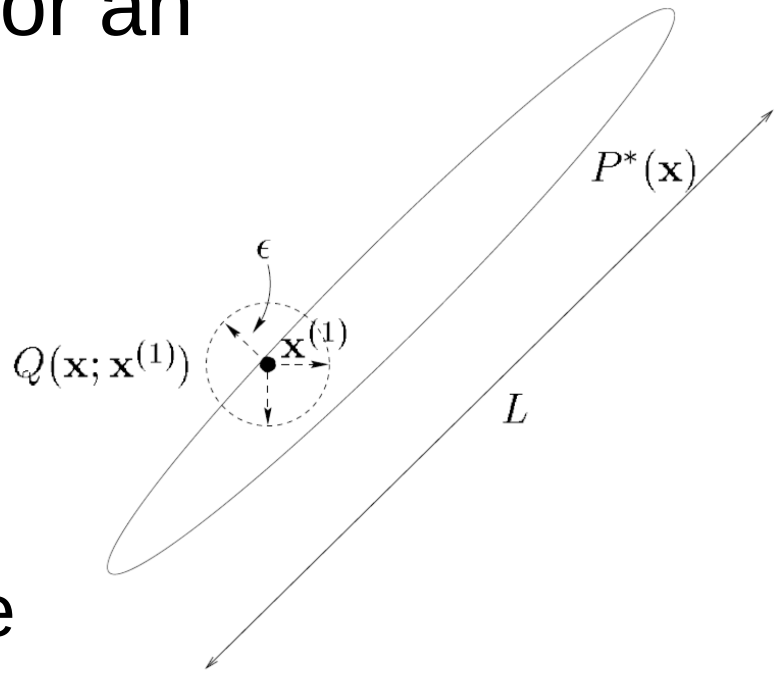
Example Gaussian proposal: $q(x^{(t)} \mid x^{(t-1)}) = \mathcal{N}(x^{(t-1)}, \epsilon^2)$

- ➢ Acceptance ratio simplifies to: $a = f(x')/f(x^{(t-1)})$
- ➢ True for any symmetric proposal: $q(x^{(t)} \mid x^{(t-1)}) = q(x^{(t-1)} \mid x^{(t)})$
- ➢ Known as Metropolis algorithm in this case

Independent Samples

[Source: D. MacKay]

Q How many M-H samples are required for an independent sample?



A Consider Gaussian proposal:

$$q(x^{(t)} \mid x^{(t-1)}) = \mathcal{N}(x^{(t-1)}, \epsilon^2)$$

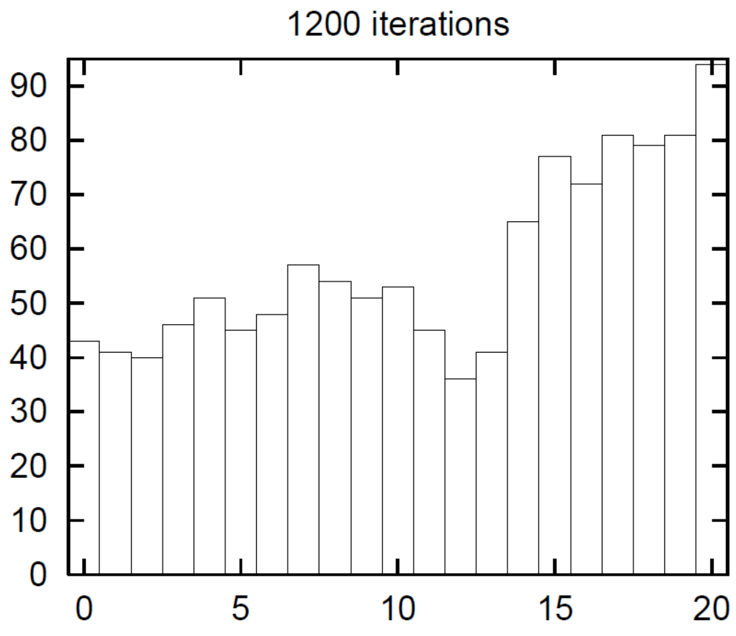
- Typically $\epsilon \ll L$ for adequate acceptance rate
- Leads to random walk dynamics, which can be slow to converge
- <u>Rule of Thumb:</u> If average acceptance is $f \in (0,1)$ need to run for roughly $T \approx (L/\epsilon)^2/f$ iterations for an independent sample

This is only a lower bound (and potentially very loose)

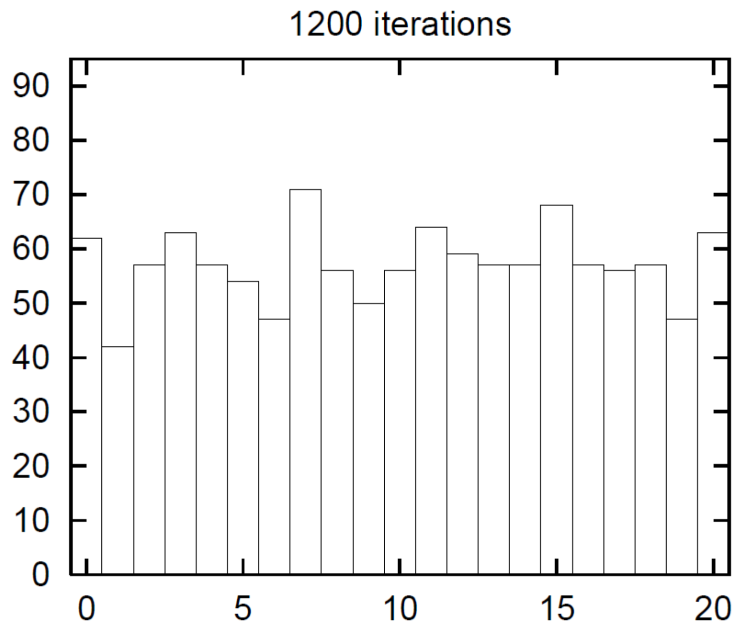
Example: Independent Samples

[Source: D. MacKay]

Metropolis



Independent



Proposal: $p(x) = \begin{cases} \frac{1}{21} & x \in \{0, \ldots, 20\} \\ 0 & \text{otherwise} \end{cases}$

Target: $q(x' \mid x) = \begin{cases} \frac{1}{2} & x' = x \pm 1 \\ 0 & \text{otherwise} \end{cases}$

From $x_0 = 10$ need ~400 steps to reach both end states (0 and 20). So, ~400 steps to generate 1 independent sample!

Very important to avoid random walk dynamics

Gibbs Sampling

[Source: Winn & Bishop]

Suppose target distribution is:

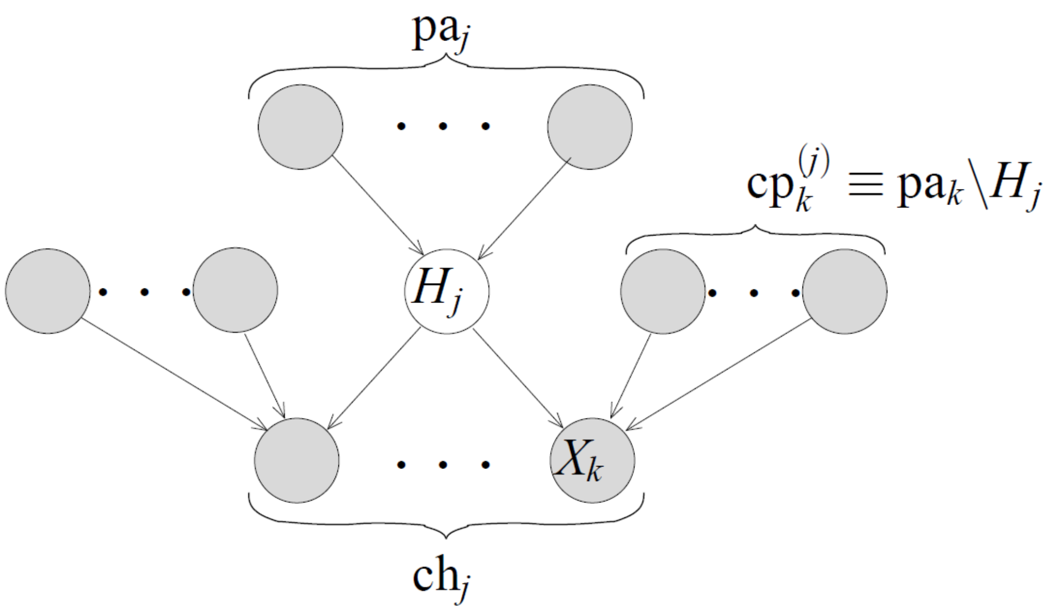
$$p(x) = \prod_{s \in \mathcal{V}} p(x_s \mid \text{Pa}(s))$$

where Pa(s) are parents of node *s*.

Metropolis-Hastings Proposal:

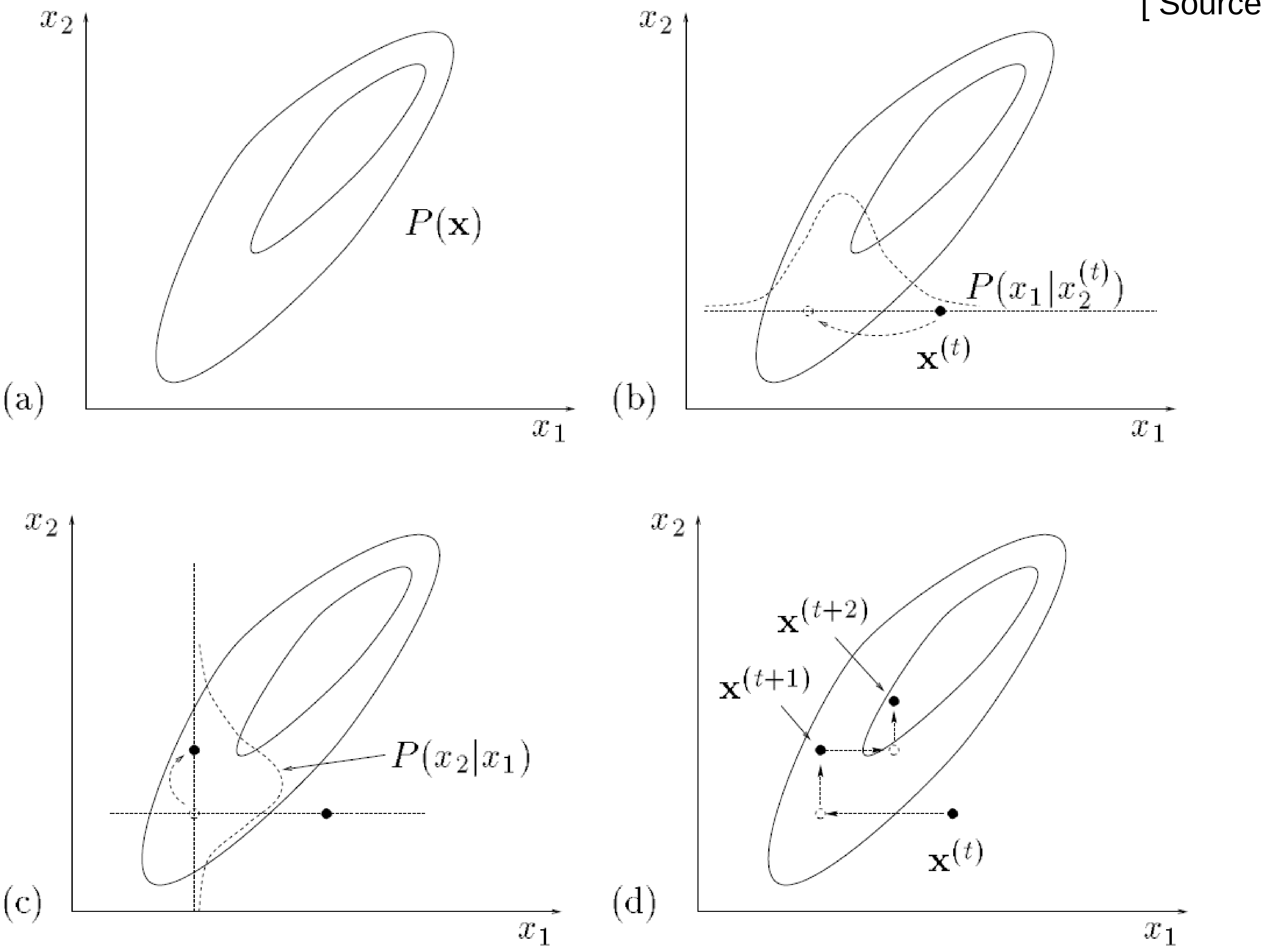
For system with K variables,

$$\begin{aligned} x_1^{(t+1)} &\sim P(x_1 | x_2^{(t)}, x_3^{(t)}, \ldots x_K^{(t)}) \\ x_2^{(t+1)} &\sim P(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots x_K^{(t)}) \\ x_3^{(t+1)} &\sim P(x_3 | x_1^{(t+1)}, x_2^{(t+1)}, \ldots x_K^{(t)}), \text{etc.} \end{aligned}$$



By conditional independence, Gibbs samples drawn from Markov blanket

[Source: D. MacKay]



Gibbs Sampling Properties

- Since Gibbs is an M-H sampler inherits all properties:
  - Aperiodicity, irreducibility, ergodicity
  - Stationary distribution is p(x)

- Proposal for x_s given by: $q(x \mid x^{(t)}) = \begin{cases} p(x_s \mid x_{\neg s}^{(t)}) & \text{If } x_{\neg s} = x_{\neg s}^{(t)} \\ 0 & \text{Otherwise} \end{cases}$

- Samples **always accepted**:

$$\Pr(\text{accept } x) = \min\left\{1, \frac{p(x)q(x^{(t)} \mid x)}{p(x^{(t)})q(x \mid x^{(t)})}\right\} = \min\left\{1, \frac{p(x)p(x_s^{(t)} \mid x_{\neg s}^{(t)})}{p(x^{(t)})p(x_s \mid x_{\neg s}^{(t)})}\right\}$$

$$= \min\left\{1, \frac{p(x_s \mid x_{\neg s}^{(t)})p(x_{\neg s}^{(t)})p(x_s^{(t)} \mid x_{\neg s}^{(t)})}{p(x_s^{(t)} \mid x_{\neg s}^{(t)})p(x_{\neg s}^{(t)})p(x_s \mid x_{\neg s}^{(t)})}\right\} = 1$$

Gibbs Sampling Extensions

Standard Gibbs suffers same random walk behavior as M-H
(but no adjustable parameters, so that's a plus…)

Block Gibbs Jointly sample subset $S \subset \mathcal{V}$ from $p(x_S \mid x_{\neg S})$

- Reduces random walk caused by highly correlated variables
- Requires that conditional $p(x_S \mid x_{\neg S})$ can be sampled efficiently

Collapsed Gibbs Marginalize some variables out of joint:

$$p(x_{\mathcal{V} \setminus S}) = \int p(x) dx_S$$

- Reduces dimensionality of space to be sampled
- Requires that marginals are computable in closed-form

Mixing MCMC Kernels

Consider a set of MCMC kernels $T_1, T_2, \ldots, T_K$ all having target distribution p(x) then the mixture:

$$T = \sum_{k=1}^{K} \pi_k T_k$$

Mixing weights

Is a valid MCMC kernel with target distribution p(x)

Mixture MCMC Transition kernel given by:

1. Sample $k \sim \pi$
2. Sample $x^{(t+1)} \sim T_k(x \mid x^{(t)})$

Inference (and related) Tasks

- Simulation: $x \sim p(x) = \frac{1}{Z} f(x)$
- Compute expectations: $\mathbb{E}[\phi(x)] = \int p(x)\phi(x)\, dx$
- Optimization: $x^* = \arg\max_x f(x)$
- Compute normalizer: $Z = \int f(x)\, dx$

Simulated Annealing

Let *annealing distribution* at temp τ be given by:

$$p_\tau(x) \propto (f(x))^{1/\tau}$$

As $\tau \to 0$ we have:

$$\lim_{\tau \to 0} p_\tau(x) = \delta(x^*) \qquad \text{where} \qquad x^* = \arg\max_x f(x)$$

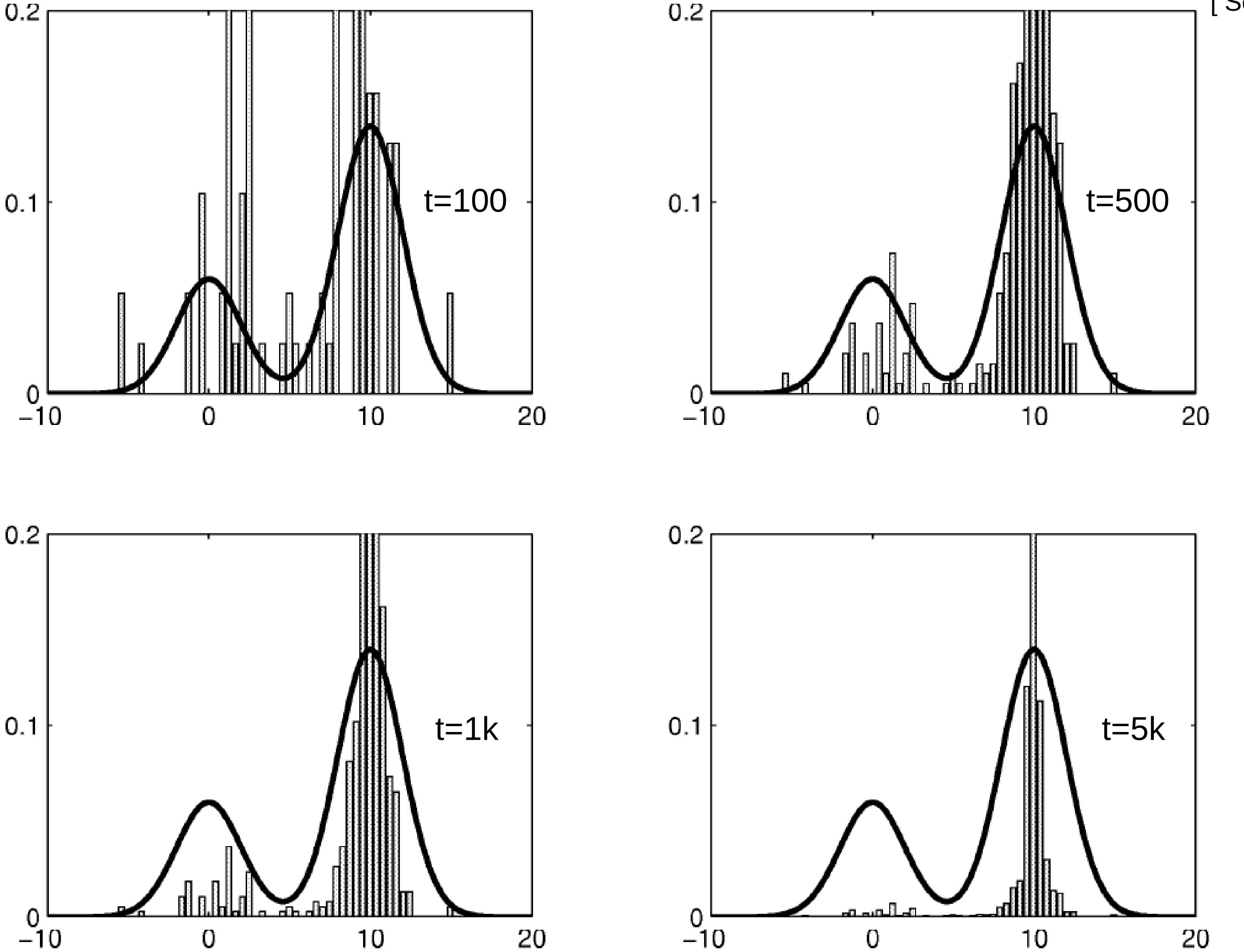
SA for Global Optimization:

Annealing schedule $\tau_0 \geq \ldots \geq \tau_t \geq \ldots \geq 0$

1. Sample $x^{(t)}$ from MCMC kernel T_t with target $p_{\tau_t}(x)$
2. Set τ_{t+1} according to annealing schedule

SA for Convergence: $\tau_0 \geq \ldots \geq 1$ Final temperature = 1

[Source: Andrieu et al.]



Inference (and related) Tasks

- Simulation: $x \sim p(x) = \frac{1}{Z} f(x)$
- Compute expectations: $\mathbb{E}[\phi(x)] = \int p(x)\phi(x)\, dx$
- Optimization: $x^* = \arg\max_x f(x)$
- Compute normalizer: $Z = \int f(x)\, dx$

Reverse IS, Chibb estimator, … Still active research area.

Comparison to Variational

- Asymptotically exact posterior samples (in theory)
- Easy to implement basic samplers (no derivatives)
- M-H broadly applicable, with few model constraints (Gibbs requires complete conditionals can be sampled)
- Diagnosing convergence is tricky (easy for variational)
- Unlike MCMC, variational inference provides:
  - Analytic posterior approximation
  - Bound of log-normalizer