MCMC methods for Bayesian Inference

A. Taylan Cemgil

Signal Processing and Communications Lab.

UNIVERSITY OF CAMBRIDGE
Department of Engineering

5R1 Stochastic Processes
March 06, 2008

Outline

Goal: Provide motivating examples to the theory of Markov chains (that Sumeet Singh has covered)

- Bayesian Inference, Probability models and Graphical model notation
- The Gibbs sampler
- Metropolis-Hastings, MCMC Transition Kernels,
- Sketch of convergence results
- Simulated annealing and iterative improvement

Bayes' Theorem

Thomas Bayes (1702-1761)

"What you know about a parameter $\boldsymbol{\lambda}$ after the data $\mathcal{D}$ arrive is what you knew before about $\boldsymbol{\lambda}$ and what the data $\mathcal{D}$ told you[1]."

$$p(\lambda|\mathcal{D}) = \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

[1](Janes 2003 (ed. by Bretthorst); MacKay 2003)

An application of Bayes' Theorem: "Source Separation"

Given two fair dice with outcomes λ and y,

$$\mathcal{D} = \lambda + y$$

What is λ when $\mathcal{D} = 9$?

“Burocratical” derivation

Formally we write

$$\begin{aligned} p(\lambda) &= \mathcal{C}(\lambda; [\ 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6\ ]) \\ p(y) &= \mathcal{C}(y; [\ 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6\ ]) \\ p(\mathcal{D}|\lambda, y) &= \delta(\mathcal{D} - (\lambda + y)) \end{aligned}$$

Kronecker delta function denoting a degenerate (deterministic) distribution $\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases}$

$$\begin{aligned} p(\lambda, y|\mathcal{D}) &= \frac{1}{p(\mathcal{D})} \times p(\mathcal{D}|\lambda, y) \times p(y)p(\lambda) \\ \text{Posterior} &= \frac{1}{\text{Evidence}} \times \text{Likelihood} \times \text{Prior} \\ p(\lambda|\mathcal{D}) &= \sum_y p(\lambda, y|\mathcal{D}) \quad \text{Posterior Marginal} \end{aligned}$$

An application of Bayes' Theorem: "Source Separation"

$$\mathcal{D} = \lambda + y = 9$$

$\mathcal{D} = \lambda + y$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$
$\lambda = 1$	2	3	4	5	6	7
$\lambda = 2$	3	4	5	6	7	8
$\lambda = \mathbf{3}$	4	5	6	7	8	**9**
$\lambda = \mathbf{4}$	5	6	7	8	**9**	10
$\lambda = \mathbf{5}$	6	7	8	**9**	10	11
$\lambda = \mathbf{6}$	7	8	**9**	10	11	12

Bayes theorem "upgrades" $p(\lambda)$ into $p(\lambda|\mathcal{D})$.

But you have to provide an observation model: $p(\mathcal{D}|\lambda)$

Another application of Bayes' Theorem: "Model Selection"

Given an unknown number of fair dice with outcomes $\lambda_1, \lambda_2, \dots, \lambda_n$,

$$\mathcal{D} = \sum_{i=1}^{n} \lambda_i$$

How many dice are there when $\mathcal{D} = 9$?

Assume that any number n is equally likely

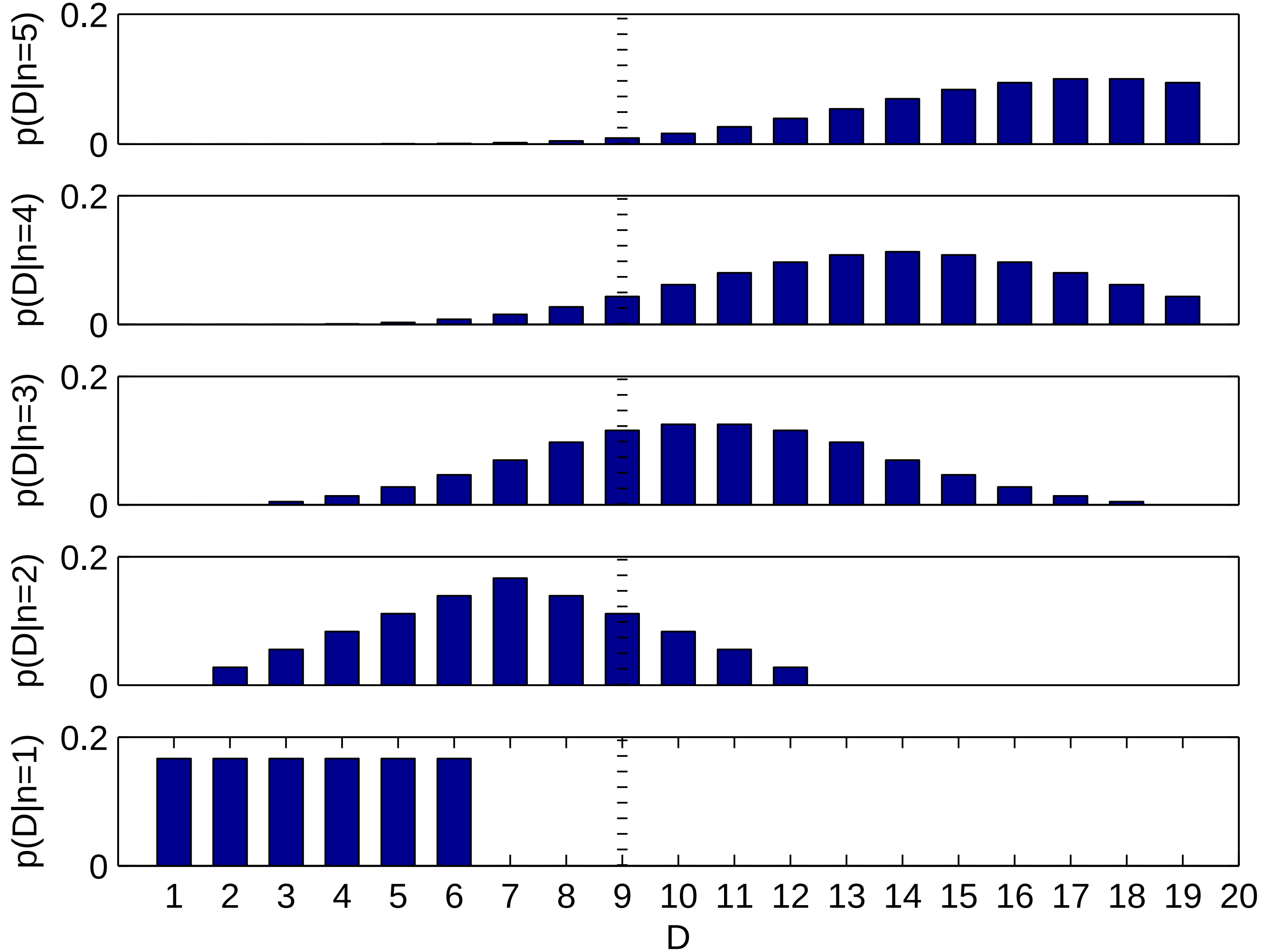
Another application of Bayes' Theorem: "Model Selection"

Given all n are equally likely (i.e., $p(n)$ is flat), we calculate (formally)

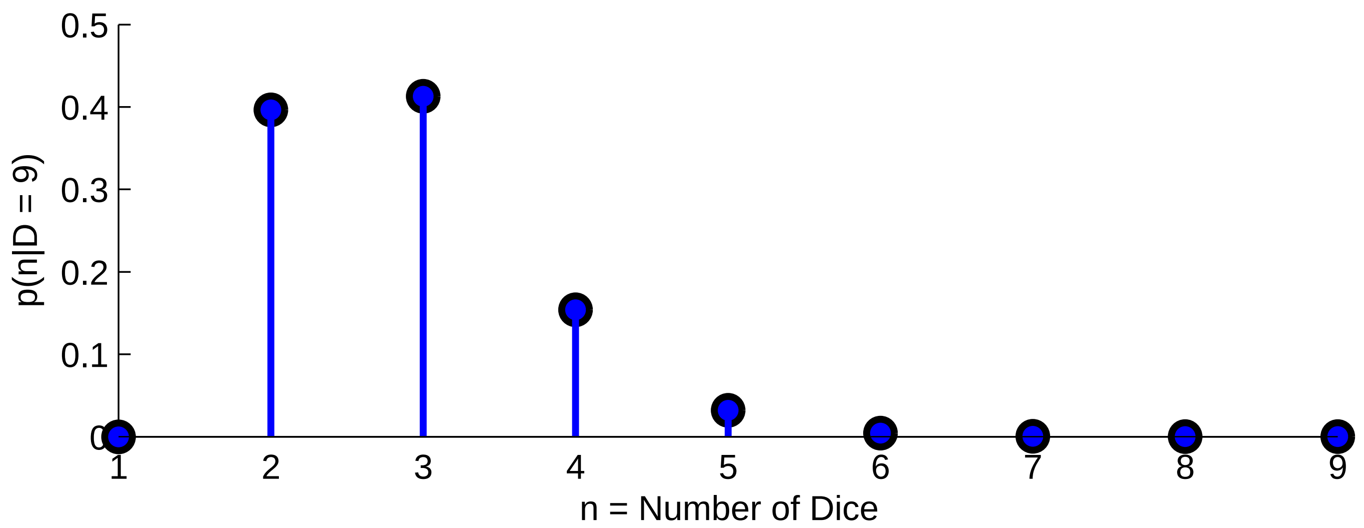
$$p(n|\mathcal{D}=9) \;=\; \frac{p(\mathcal{D}=9|n)p(n)}{p(\mathcal{D})} \propto p(\mathcal{D}=9|n)$$

$$\begin{aligned}
p(\mathcal{D}|n=1) &= \sum_{\lambda_1} p(\mathcal{D}|\lambda_1)p(\lambda_1) \\
p(\mathcal{D}|n=2) &= \sum_{\lambda_1}\sum_{\lambda_2} p(\mathcal{D}|\lambda_1,\lambda_2)p(\lambda_1)p(\lambda_2) \\
\dots \\
p(\mathcal{D}|n=n') &= \sum_{\lambda_1,\dots,\lambda_{n'}} p(\mathcal{D}|\lambda_1,\dots,\lambda_{n'})\prod_{i=1}^{n'} p(\lambda_i)
\end{aligned}$$

$$p(\mathcal{D}|n) = \sum_{\boldsymbol{\lambda}} p(\mathcal{D}|\boldsymbol{\lambda}, n)p(\boldsymbol{\lambda}|n)$$



Another application of Bayes' Theorem: "Model Selection"



- Complex models are more flexible but they spread their probability mass

- Bayesian inference inherently prefers "simpler models" – Occam's razor

- Computational burden: We need to sum over all parameters λ

Probabilistic Inference

A huge spectrum of applications – all boil down to computation of

- **expectations** of functions under probability distributions: **Integration**

$$\langle f(x) \rangle = \int_{\mathcal{X}} dx p(x) f(x) \qquad \langle f(x) \rangle = \sum_{x \in \mathcal{X}} p(x) f(x)$$

- **modes** of functions under probability distributions: **Optimization**

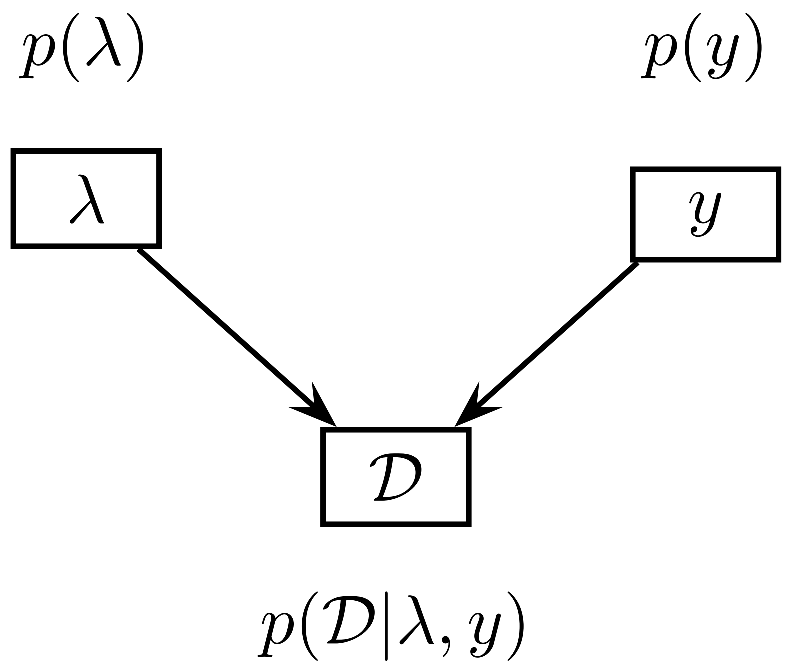
$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} p(x) f(x)$$

- any “mix” of the above: e.g.,

$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} p(x) = \operatorname*{argmax}_{x \in \mathcal{X}} \int_{\mathcal{Z}} dz p(z) p(x|z)$$

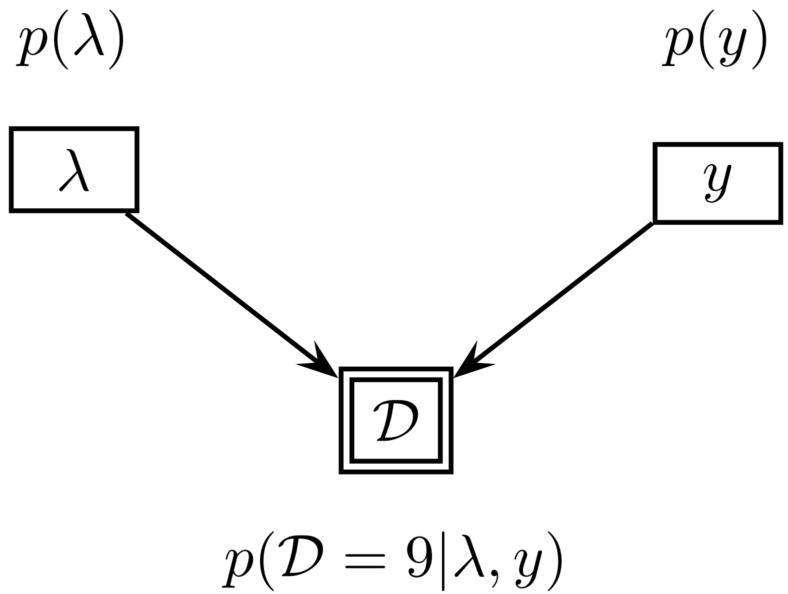
Directed Acyclic Graphical (DAG) Models
and
Factor Graphs

DAG Example: Two dice



$$p(\mathcal{D}, \lambda, y) = p(\mathcal{D}|\lambda, y)p(\lambda)p(y)$$

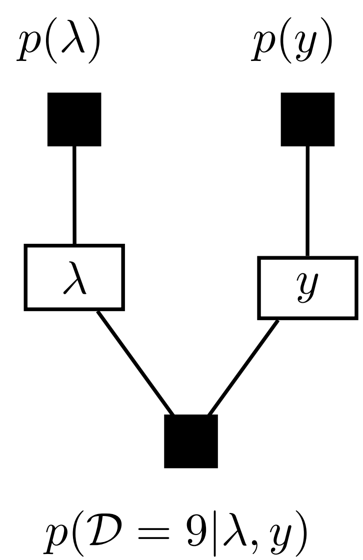
DAG with observations



$$\phi_{\mathcal{D}}(\lambda, y) \quad = \quad p(\mathcal{D} = 9|\lambda, y)p(\lambda)p(y)$$

Factor graphs (Kschischang et. al.)

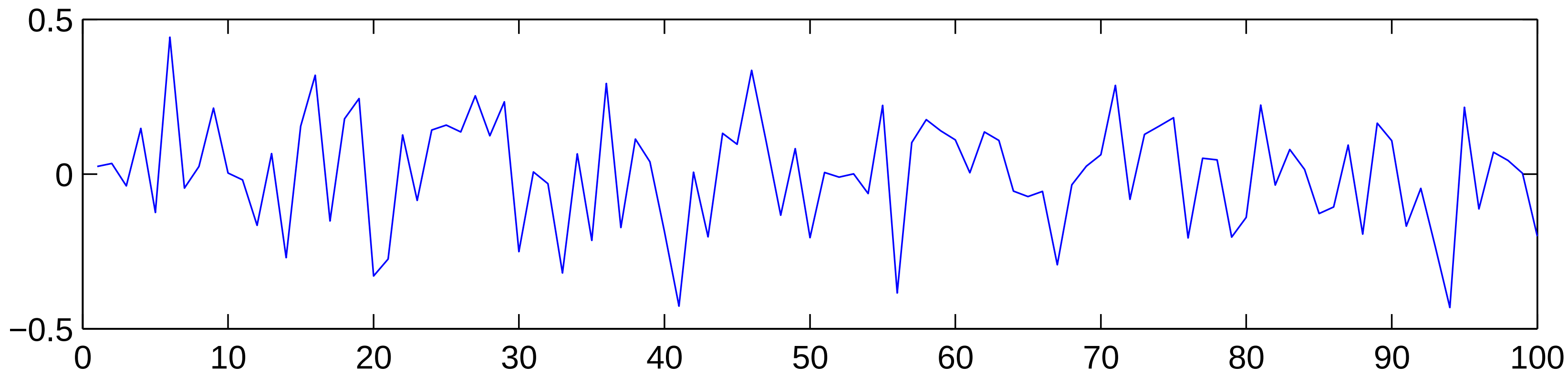
- A bipartite graph. A powerful graphical representation of the inference problem
  - **Factor nodes**: Black squares. Factor potentials (local functions) defining the posterior.
  - **Variable nodes**: White Nodes. Define collections of random variables
  - **Edges**: denote membership. A variable node is connected to a factor node if a member variable is an argument of the local function.



$$\phi_{\mathcal{D}}(\lambda, y) \quad = \quad p(\mathcal{D} = 9|\lambda, y)p(\lambda)p(y) = \phi_1(\lambda, y)\phi_2(\lambda)\phi_3(y)$$

Probability Models

Example: AR(1) model



$$x_k = Ax_{k-1} + \epsilon_k \qquad k = 1 \dots K$$

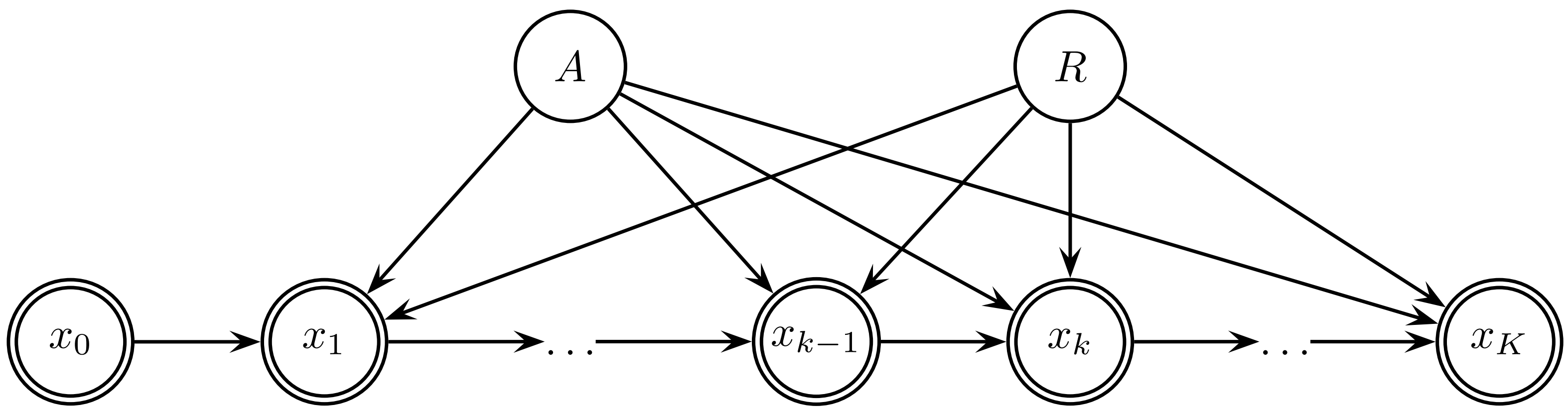
ϵ_k is i.i.d., zero mean and normal with variance R.

Estimation problem:

Given $x_0, \dots, x_K$, determine coefficient A and variance R (both scalars).

AR(1) model, Generative Model notation

$$
\begin{aligned}
A &\sim \mathcal{N}(A; 0, P) \\
R &\sim \mathcal{IG}(R; \nu, \beta/\nu) \\
x_k | x_{k-1}, A, R &\sim \mathcal{N}(x_k; A x_{k-1}, R) \qquad\qquad x_0 = \hat{x}_0
\end{aligned}
$$



Observed variables are shown with double circles

Example, Univariate Gaussian

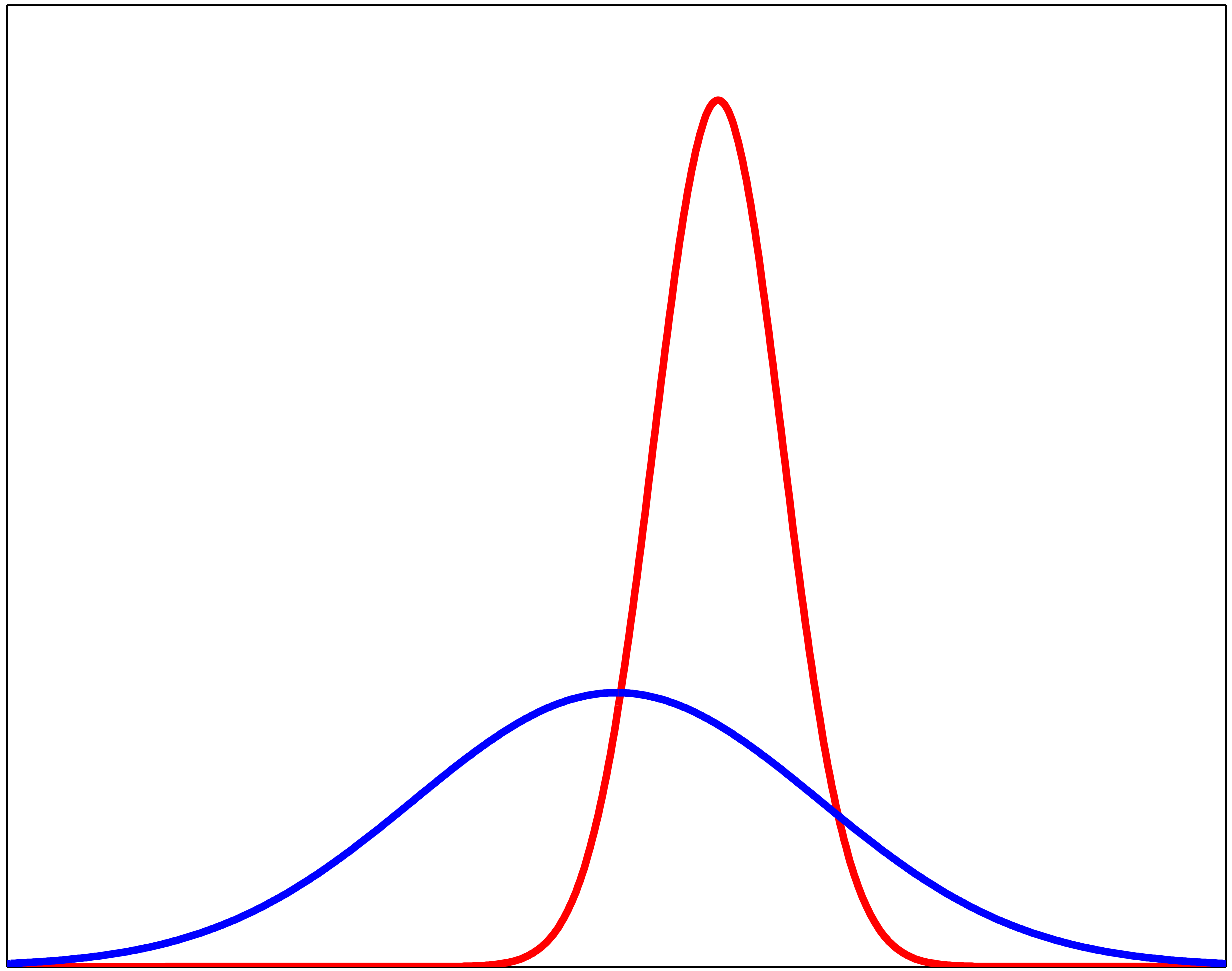
The Gaussian distribution with mean m and covariance S has the form

$$
\begin{aligned}
\mathcal{N}(x; m, S) &= (2\pi S)^{-1/2} \exp\{-\frac{1}{2}(x-m)^2/S\} \\
&= \exp\{-\frac{1}{2}(x^2 + m^2 - 2xm)/S - \frac{1}{2}\log(2\pi S)\} \\
&= \exp\left\{\frac{m}{S}x - \frac{1}{2S}x^2 - \left(\frac{1}{2}\log(2\pi S) + \frac{1}{2S}m^2\right)\right\} \\
&= \exp\{\underbrace{\begin{pmatrix} m/S \\ -\frac{1}{2}/S \end{pmatrix}^\top}_{\theta} \underbrace{\begin{pmatrix} x \\ x^2 \end{pmatrix}}_{\psi(x)} - c(\theta)\}
\end{aligned}
$$

Hence by matching coefficients we have

$$
\exp\left\{-\tfrac{1}{2}Kx^2 + hx + g\right\} \Leftrightarrow S = K^{-1} \quad m = K^{-1}h
$$

Example, Gaussian



The Multivariate Gaussian Distribution

μ is the mean and P is the covariance:

$$
\begin{aligned}
\mathcal{N}(s;\mu,P) &= |2\pi P|^{-1/2}\exp\left(-\frac{1}{2}(s-\mu)^T P^{-1}(s-\mu)\right)\\
&= \exp\left(-\frac{1}{2}s^T P^{-1}s + \mu^T P^{-1}s - \frac{1}{2}\mu^T P^{-1}\mu - \frac{1}{2}|2\pi P|\right)\\
\log \mathcal{N}(s;\mu,P) &= -\frac{1}{2}s^T P^{-1}s + \mu^T P^{-1}s + \text{ const}\\
&= -\frac{1}{2}\,\mathbf{Tr}\,P^{-1}ss^T + \mu^T P^{-1}s + \text{ const}\\
&=^+ -\frac{1}{2}\,\mathbf{Tr}\,P^{-1}ss^T + \mu^T P^{-1}s
\end{aligned}
$$

Notation: $\log f(x) =^+ g(x) \iff f(x) \propto \exp(g(x)) \iff \exists c \in \mathbb{R} : f(x) = c\exp(g(x))$

$$
\log p(s) \quad =^+ \quad -\frac{1}{2}\,\mathbf{Tr}\,Kss^T + h^\top s \qquad \Rightarrow \qquad p(s) = \mathcal{N}(s; K^{-1}h, K^{-1})
$$

Example, Inverse Gamma

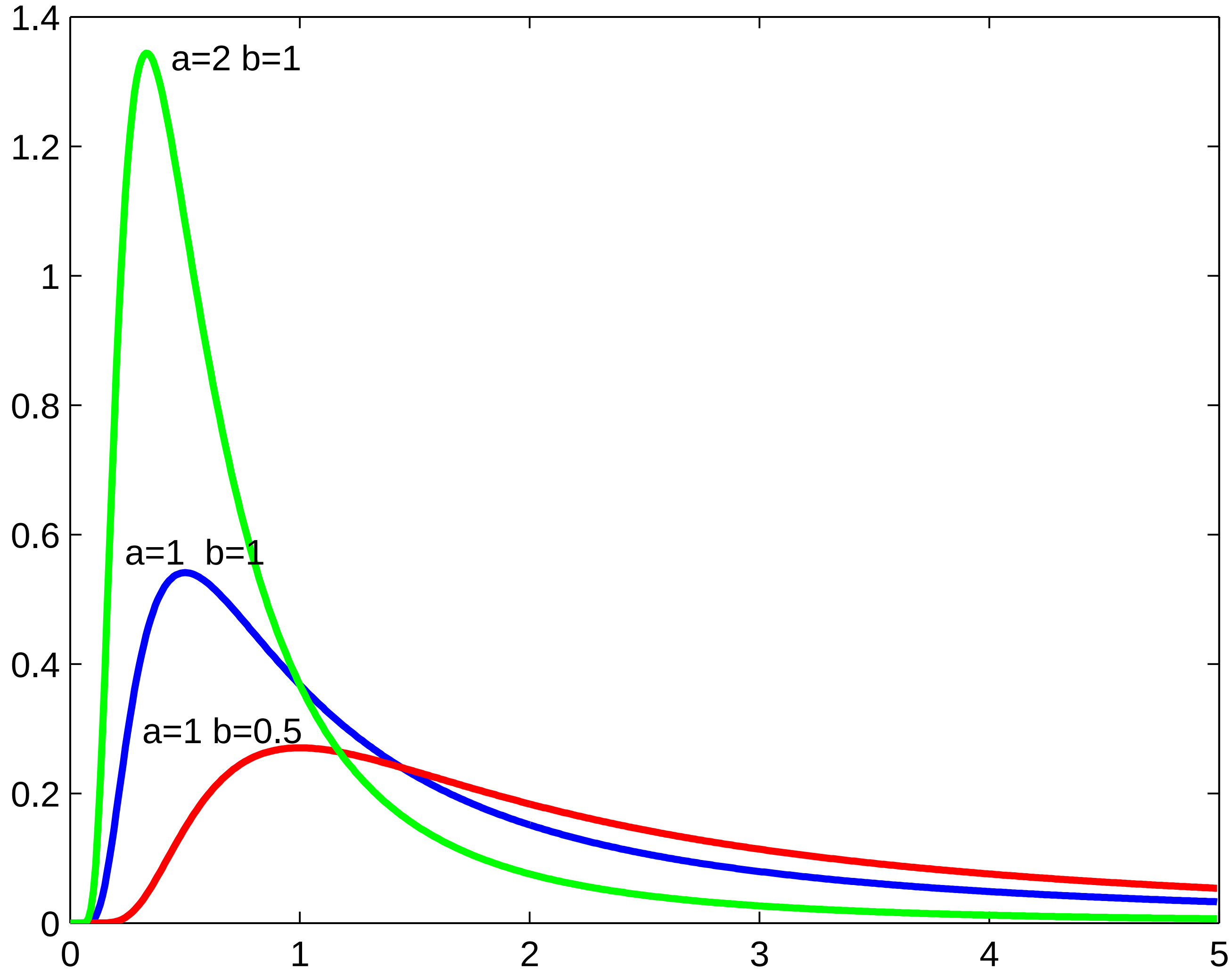
The inverse Gamma distribution with shape a and scale b

$$
\begin{aligned}
\mathcal{IG}(r; a, b) &= \frac{1}{\Gamma(a)} \frac{r^{-(a+1)}}{b^a} \exp(-\frac{1}{br}) \\
&= \exp\left(-(a+1)\log r - \frac{1}{br} - \log\Gamma(a) - a\log b \right) \\
&= \exp\left(\left(\begin{array}{c} -(a+1) \\ -1/b \end{array} \right)^\top \left(\begin{array}{c} \log r \\ 1/r \end{array} \right) - \log\Gamma(a) - a\log b \right)
\end{aligned}
$$

Hence by matching coefficients, we have

$$
\exp\left\{ \alpha \log r + \beta \frac{1}{r} + c \right\} \Leftrightarrow a = -\alpha - 1 \qquad b = -1/\beta
$$

Example, Inverse Gamma



Basic Distributions : Exponential Family

- Following distributions are used often as elementary building blocks:
  - Gaussian
  - Gamma, Inverse Gamma, (Exponential, Chi-square, Wishart)
  - Dirichlet
  - Discrete (Categorical), Bernoulli, multinomial

- All of those distributions can be written as

$$p(x|\theta) = \exp\{\theta^\top \psi(x) - c(\theta)\}$$

$$c(\theta) = \log \int_{\mathcal{X}^n} dx \; \exp(\theta^\top \psi(x)) \quad \text{log-partition function}$$

$$\theta \quad \text{canonical parameters}$$

$$\psi(x) \quad \text{sufficient statistics}$$

Conjugate priors: Posterior is in the same family as the prior.

Example: posterior inference for the variance R of a zero mean Gaussian.

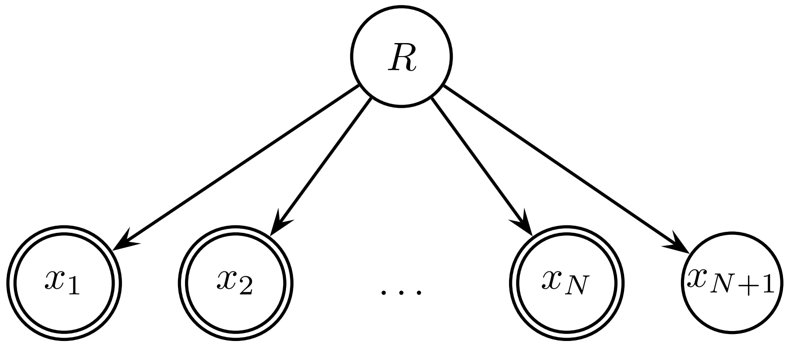
$$
\begin{aligned}
p(x|R) &= \mathcal{N}(x; 0, R) \\
p(R) &= \mathcal{IG}(R; a, b)
\end{aligned}
$$

$$
\begin{aligned}
p(R|x) &\propto p(R)p(x|R) \\
&\propto \exp\left(-(a+1)\log R - (1/b)\frac{1}{R}\right)\exp\left(-(x^2/2)\frac{1}{R} - \frac{1}{2}\log R\right) \\
&= \exp\left(\left(\begin{array}{c} -(a+1+\frac{1}{2}) \\ -(1/b + x^2/2) \end{array}\right)^\top \left(\begin{array}{c} \log R \\ 1/R \end{array}\right)\right) \\
&\propto \mathcal{IG}(R; a+\frac{1}{2}, \frac{2}{x^2+2/b})
\end{aligned}
$$

Like the prior, this is an inverse-Gamma distribution.

Conjugate priors: Posterior is in the same family as the prior.

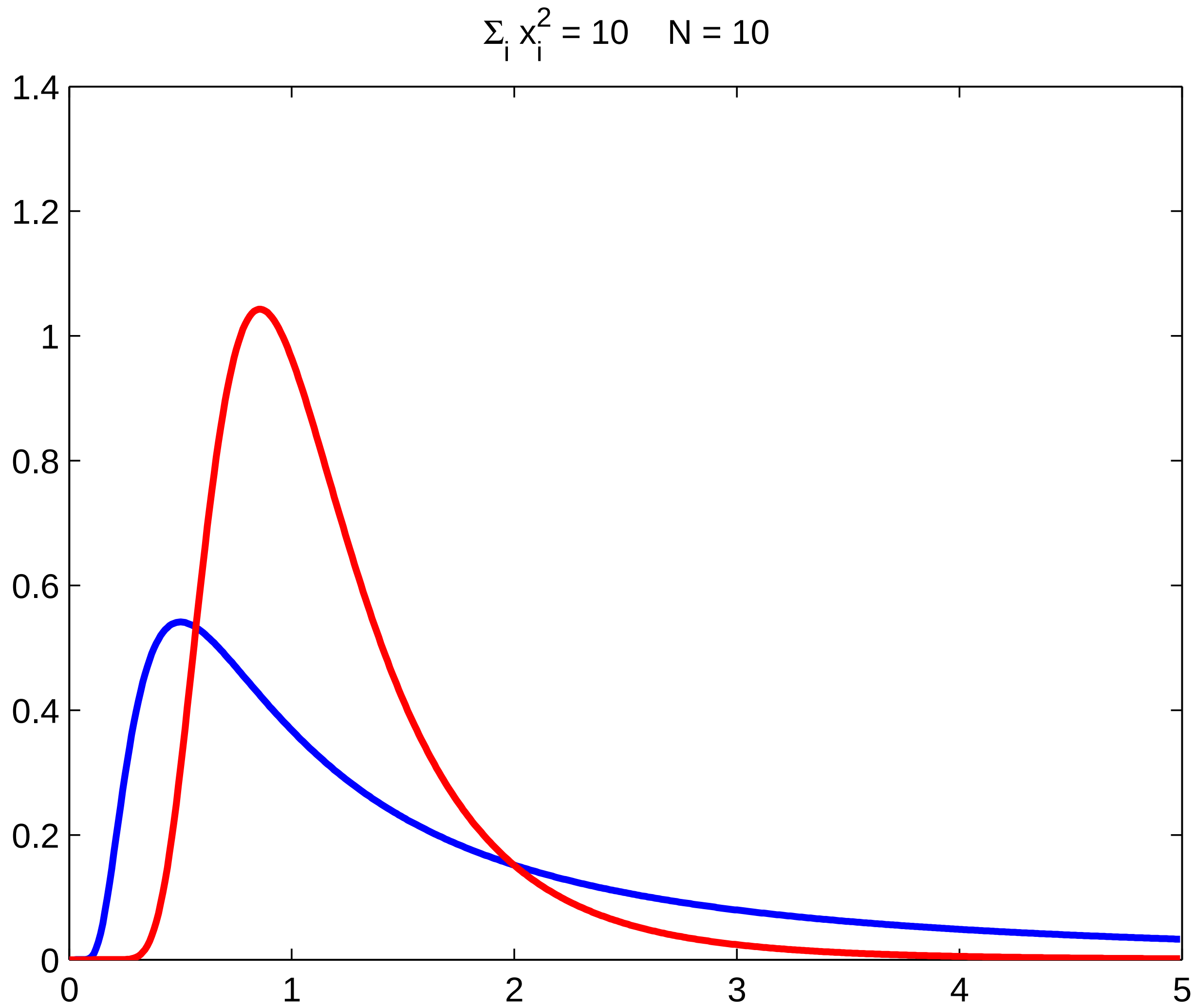
Example: posterior inference of variance R from $x_1, \dots, x_N$.



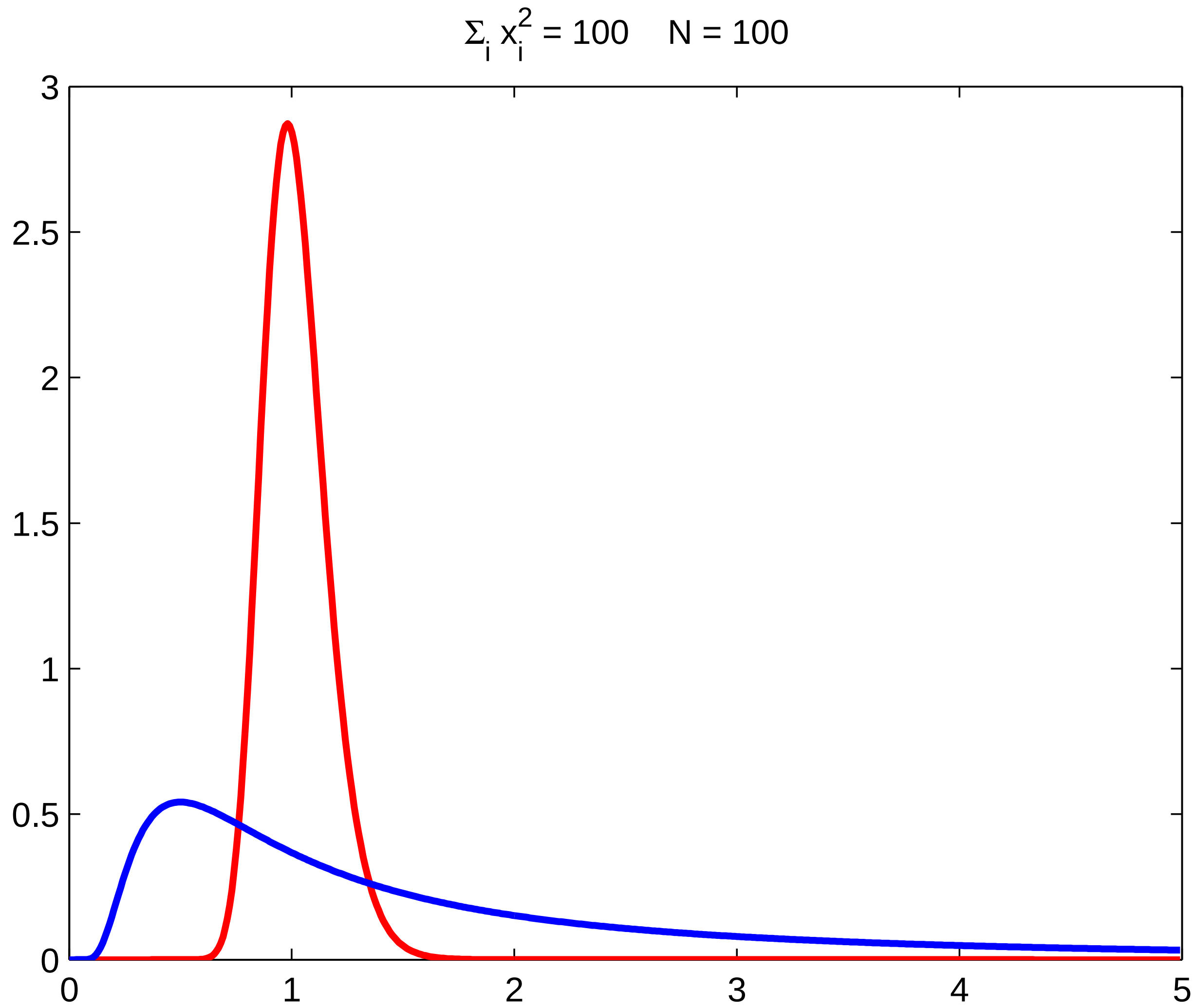
$$
\begin{aligned}
p(R|x) &\propto p(R) \prod_{i=1}^{N} p(x_i|R) \\
&\propto \exp\left(-(a+1)\log R - (1/b)\frac{1}{R}\right) \exp\left(-\left(\frac{1}{2}\sum_i x_i^2\right)\frac{1}{R} - \frac{N}{2}\log R\right) \\
&= \exp\left(\begin{pmatrix} -(a+1+\frac{N}{2}) \\ -(1/b + \frac{1}{2}\sum_i x_i^2) \end{pmatrix}^{\top} \begin{pmatrix} \log R \\ 1/R \end{pmatrix}\right) \propto \mathcal{IG}(R; a + \frac{N}{2}, \frac{2}{\sum_i x_i^2 + 2/b})
\end{aligned}
$$

Sufficient statistics are **additive**

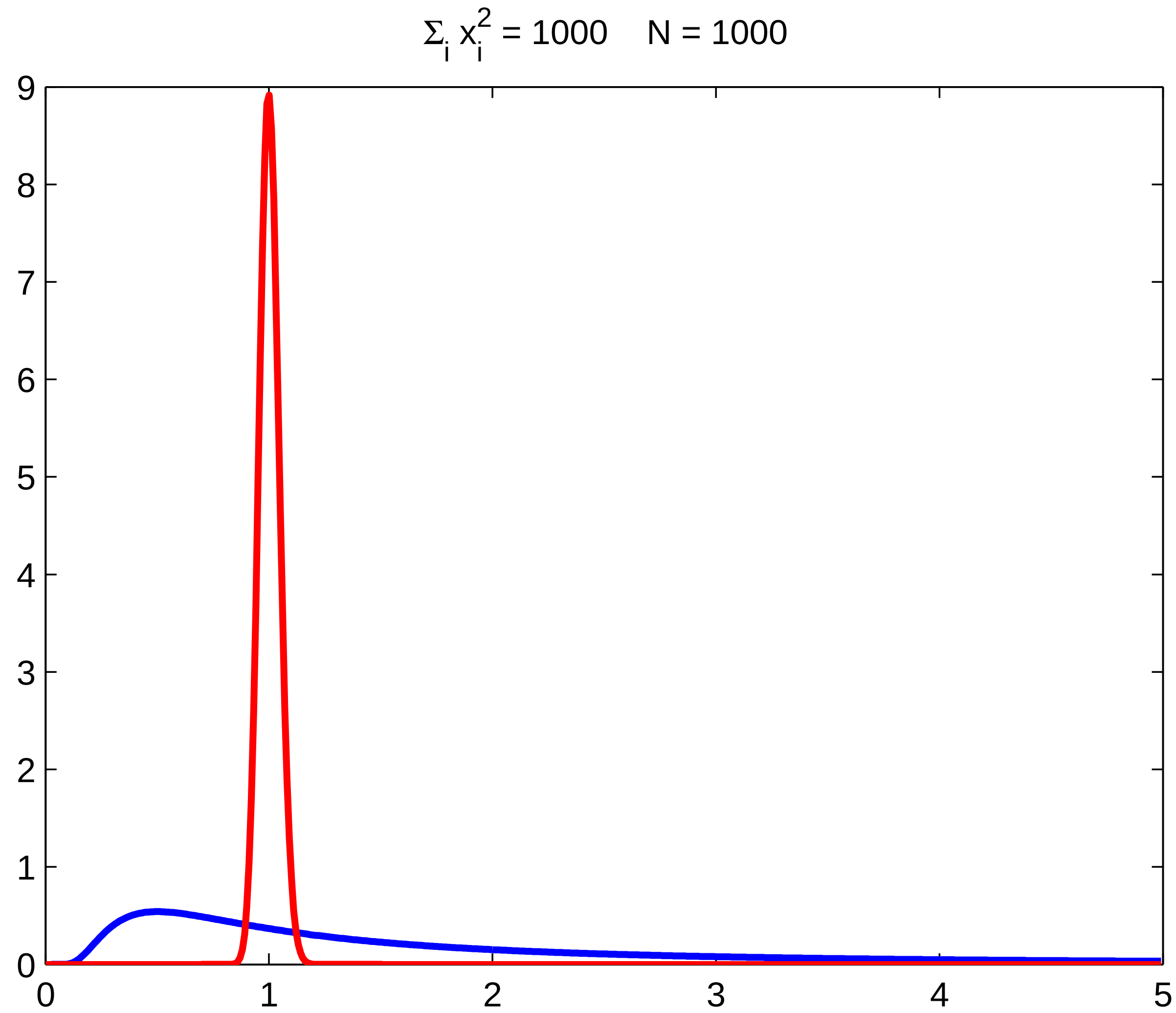
Inverse Gamma, $\sum_i x_i^2 = 10 \quad N = 10$



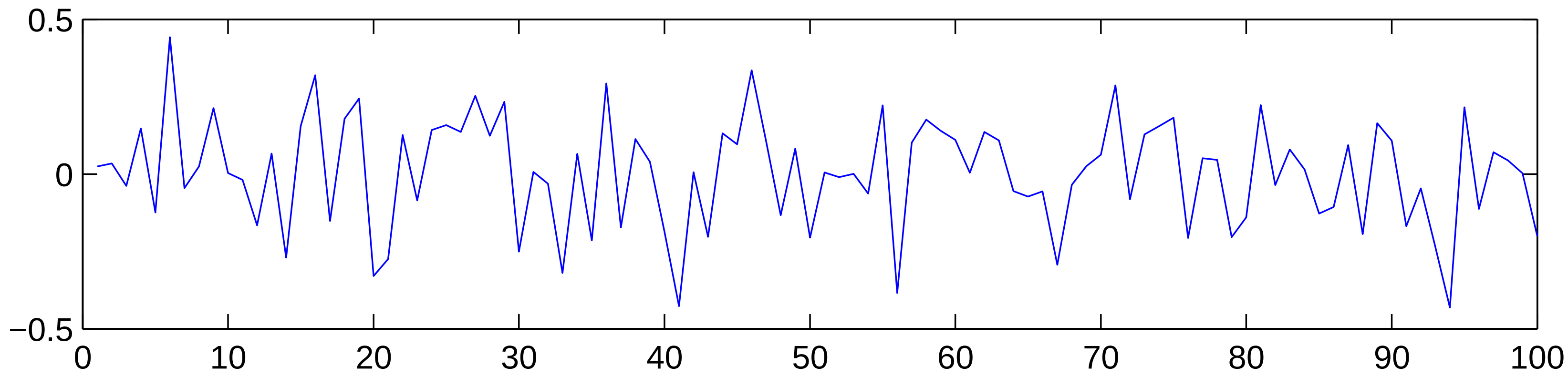
Inverse Gamma, $\sum_i x_i^2 = 100 \quad N = 100$



Inverse Gamma, $\sum_i x_i^2 = 1000 \quad N = 1000$



Example: AR(1) model



$$x_k = Ax_{k-1} + \epsilon_k \qquad\qquad k = 1 \dots K$$

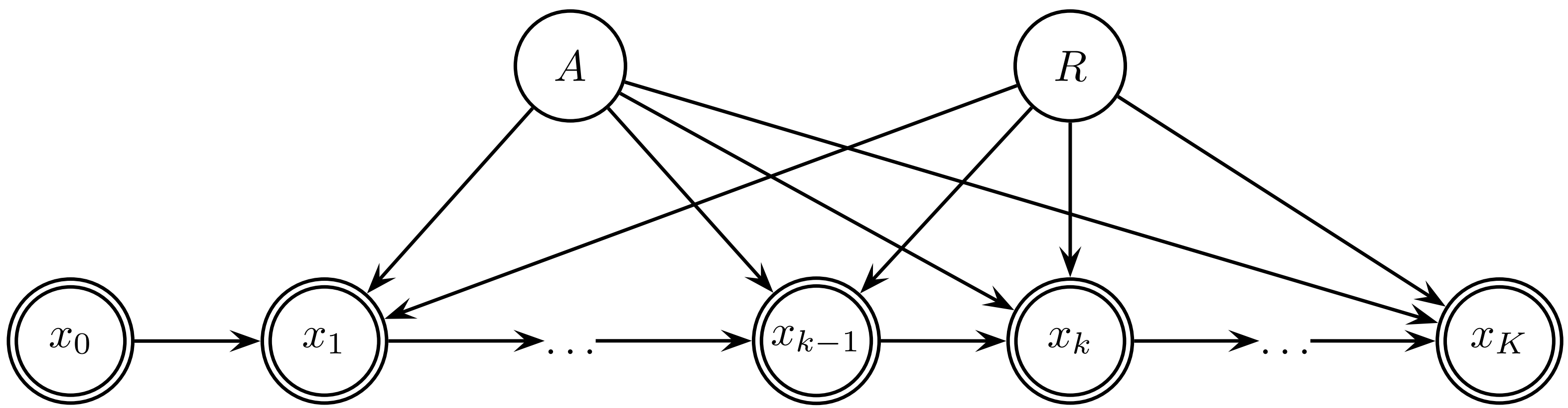
ϵ_k is i.i.d., zero mean and normal with variance R.

Estimation problem:

Given $x_0, \dots, x_K$, determine coefficient A and variance R (both scalars).

AR(1) model, Generative Model notation

$$
\begin{aligned}
A &\sim \mathcal{N}(A; 0, P) \\
R &\sim \mathcal{IG}(R; \nu, \beta/\nu) \\
x_k | x_{k-1}, A, R &\sim \mathcal{N}(x_k; A x_{k-1}, R) \qquad x_0 = \hat{x}_0
\end{aligned}
$$



Gaussian : $\mathcal{N}(x; \mu, V) \equiv |2\pi V|^{-\frac{1}{2}} \exp(-\frac{1}{2}(x-\mu)^2/V)$

Inverse-Gamma distribution: $\mathcal{IG}(x; a, b) \equiv \Gamma(a)^{-1} b^{-a} x^{-(a+1)} \exp(-1/(bx)) \quad x \geq 0$

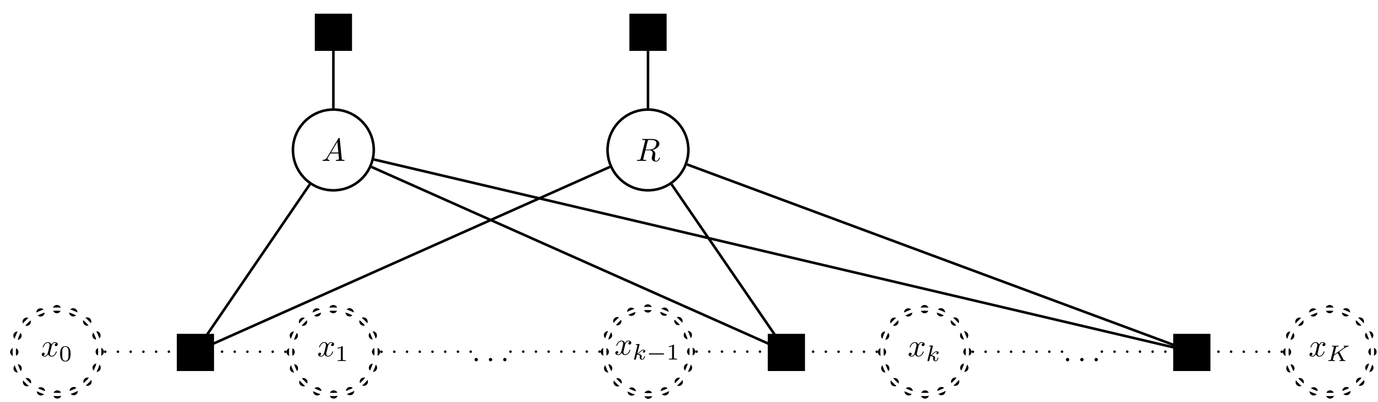
Observed variables are shown with double circles

AR(1) Model. Bayesian Posterior Inference

$$p(A, R|x_0, x_1, \dots, x_K) \quad \propto \quad p(x_1, \dots, x_K|x_0, A, R)p(A, R)$$
$$\text{Posterior} \quad \propto \quad \text{Likelihood} \times \text{Prior}$$

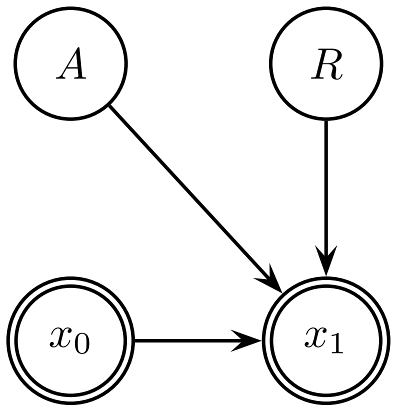
Using the Markovian (conditional independence) structure we have

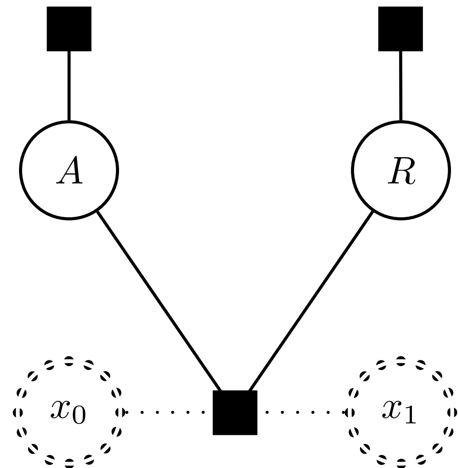
$$p(A, R|x_0, x_1, \dots, x_K) \quad \propto \quad \left(\prod_{k=1}^{K} p(x_k|x_{k-1}, A, R) \right) p(A)p(R)$$



Numerical Example

Suppose $K = 1$,





By Bayes' Theorem and the structure of AR(1) model

$$
\begin{aligned}
p(A, R|x_0, x_1) &\propto p(x_1|x_0, A, R)p(A)p(R) \\
&= \mathcal{N}(x_1; Ax_0, R)\mathcal{N}(A; 0, P)\mathcal{IG}(R; \nu, \beta/\nu)
\end{aligned}
$$

Numerical Example

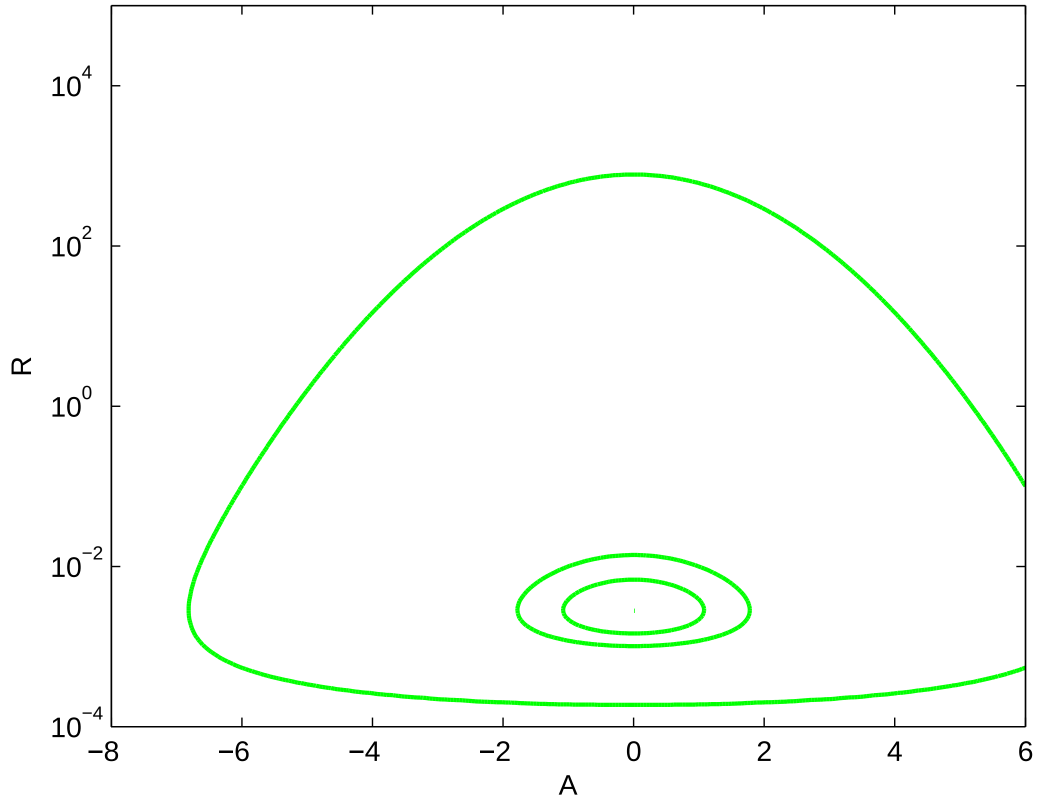
$$
\begin{aligned}
p(A, R|x_0, x_1) &\propto p(x_1|x_0, A, R)p(A)p(R) \\
&= \mathcal{N}(x_1; Ax_0, R)\mathcal{N}(A; 0, P)\mathcal{IG}(R; \nu, \beta/\nu) \\
&\propto \exp\left(-\frac{1}{2}\frac{x_1^2}{R} + x_0 x_1 \frac{A}{R} - \frac{1}{2}\frac{x_0^2 A^2}{R} - \frac{1}{2}\log 2\pi R\right) \\
&\quad \exp\left(-\frac{1}{2}\frac{A^2}{P}\right)\exp\left(-(\nu+1)\log R - \frac{\nu}{\beta}\frac{1}{R}\right)
\end{aligned}
$$

This posterior has a nonstandard form

$$
\exp\left(\alpha_1 \frac{1}{R} + \alpha_2 \frac{A}{R} + \alpha_3 \frac{A^2}{R} + \alpha_4 \log R + \alpha_5 A^2\right)
$$

Numerical Example, the prior $p(A, R)$

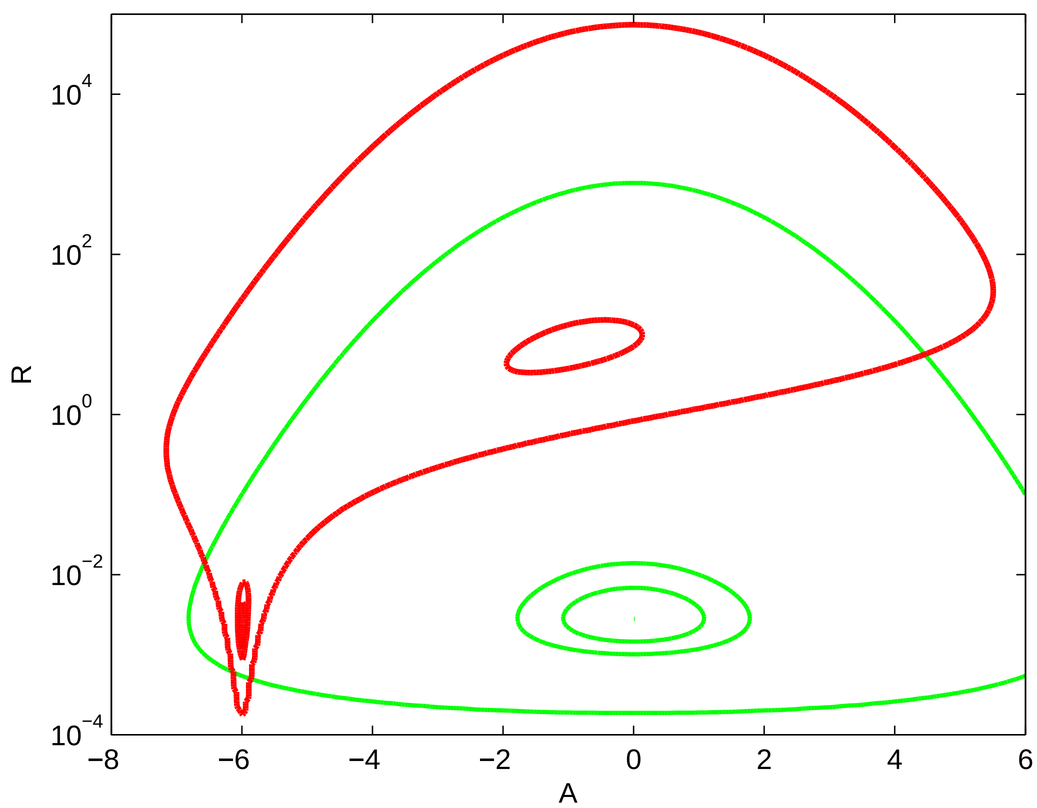
Equiprobability contour of $p(A)p(R)$



$$A \sim \mathcal{N}(A; 0, 1.2) \qquad R \sim \mathcal{IG}(R; 0.4, 250)$$

Suppose: $x_0 = 1 \qquad x_1 = -6 \qquad x_1 \sim \mathcal{N}(x_1; Ax_0, R)$

Numerical Example, the posterior $p(A, R|x)$



Note the bimodal posterior with $x_0 = 1, x_1 = -6$

- $A \approx -6 \Leftrightarrow$ low noise variance R.
- $A \approx 0 \Leftrightarrow$ high noise variance R.

Remarks

- Even very simple models can lead easily to complicated posterior distributions

- Ambiguous data usually leads to a multimodal posterior, each mode corresponding to one possible explanation

- *A-priori* independent variables often become dependent *a-posteriori* (“Explaining away”)

- (Unfortunately), exact posterior inference is only possible for few special cases

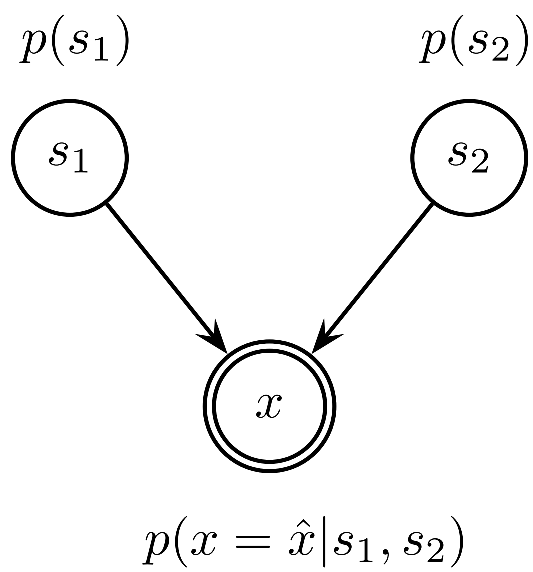
$\Rightarrow$ We need numerical approximate inference methods

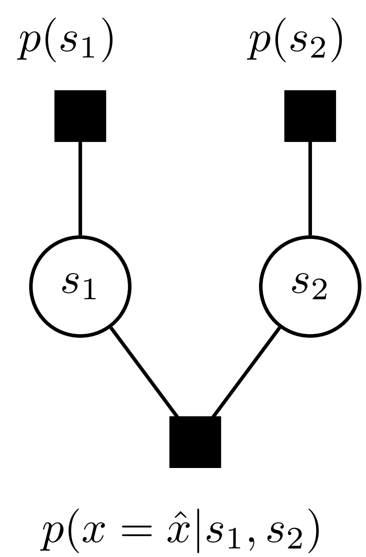
Approximate Inference

- Markov Chain Monte Carlo, Gibbs sampler

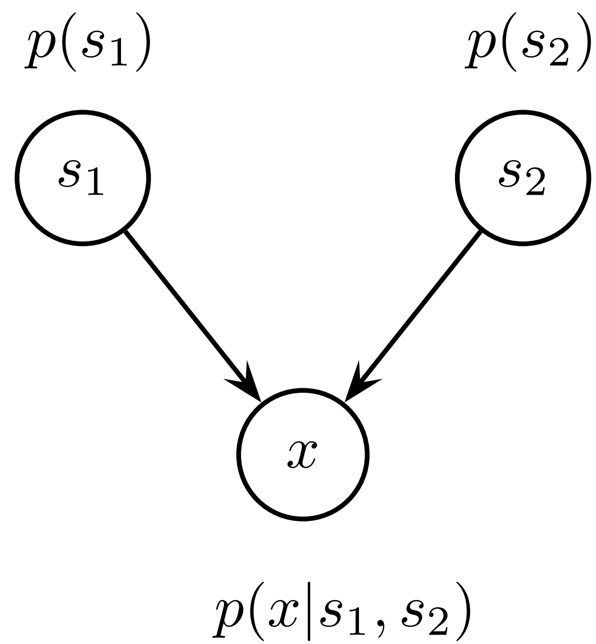
It turns out that the Gibbs sampler can be viewed as a message passing algorithm on a factor graph

- Lets focus on a simpler graph to illustrate these algorithms





Toy Model : “One sample source separation”

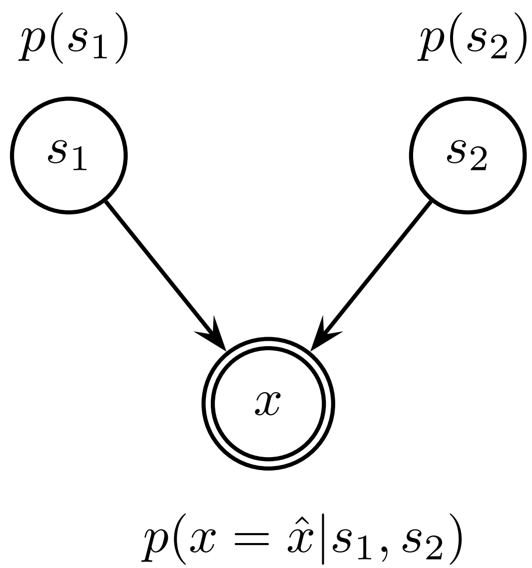


This graph encodes the joint: $p(x, s_1, s_2) = p(x|s_1, s_2)p(s_1)p(s_2)$

$$
\begin{aligned}
s_1 &\sim p(s_1) = \mathcal{N}(s_1; \mu_1, P_1) \\
s_2 &\sim p(s_2) = \mathcal{N}(s_2; \mu_2, P_2) \\
x|s_1, s_2 &\sim p(x|s_1, s_2) = \mathcal{N}(x; s_1 + s_2, R)
\end{aligned}
$$

Toy example

Suppose, we observe $x = \hat{x}$.



- By Bayes' theorem, the posterior is given by:

$$\mathcal{P} \equiv p(s_1, s_2|x = \hat{x}) = \frac{1}{Z_{\hat{x}}} p(x = \hat{x}|s_1, s_2) p(s_1) p(s_2) \equiv \frac{1}{Z_{\hat{x}}} \phi(s_1, s_2)$$

- The function $\phi(s_1, s_2)$ is proportional to the exact posterior. ($Z_{\hat{x}} \equiv p(x = \hat{x})$)

Toy example, cont.

$$
\begin{aligned}
\log p(s_1) &= \mu_1^T P_1^{-1} s_1 - \frac{1}{2} s_1^T P_1^{-1} s_1 + \text{const} \\
\log p(s_2) &= \mu_2^T P_2^{-1} s_2 - \frac{1}{2} s_2^T P_2^{-1} s_2 + \text{const} \\
\log p(x|s_1, s_2) &= \hat{x}^T R^{-1}(s_1 + s_2) - \frac{1}{2}(s_1 + s_2)^T R^{-1}(s_1 + s_2) + \text{const}
\end{aligned}
$$

$$
\begin{aligned}
\log \phi(s_1, s_2) &= \log p(x = \hat{x}|s_1, s_2) + \log p(s_1) + \log p(s_2) \\
&=^+ \left(\mu_1^T P_1^{-1} + \hat{x}^T R^{-1}\right) s_1 + \left(\mu_2^T P_2^{-1} + \hat{x}^T R^{-1}\right) s_2 \\
&\quad -\frac{1}{2}\mathbf{Tr}\left(P_1^{-1} + R^{-1}\right) s_1 s_1^T - \underbrace{s_1^T R^{-1} s_2}_{(*)} - \frac{1}{2}\mathbf{Tr}\left(P_2^{-1} + R^{-1}\right) s_2 s_2^T
\end{aligned}
$$

- The (*) term is the cross correlation term that makes s_1 and s_2 a-posteriori dependent.

Toy example, cont.

Completing the square

$$\log \phi(s_1, s_2) \quad =^+ \quad \begin{pmatrix} P_1^{-1}\mu_1 + R^{-1}\hat{x} \\ P_2^{-1}\mu_2 + R^{-1}\hat{x} \end{pmatrix}^\top \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$$

$$-\frac{1}{2}\begin{pmatrix} s_1 \\ s_2 \end{pmatrix}^\top \begin{pmatrix} P_1^{-1} + R^{-1} & R^{-1} \\ R^{-1} & P_2^{-1} + R^{-1} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$$

Remember: $\log \mathcal{N}(s; m, \Sigma) \quad =^+ \quad (\Sigma^{-1}m)^\top s - \frac{1}{2}s^\top \Sigma^{-1} s$

$$\Sigma \quad = \quad \begin{pmatrix} P_1^{-1} + R^{-1} & R^{-1} \\ R^{-1} & P_2^{-1} + R^{-1} \end{pmatrix}^{-1} \qquad m = \Sigma \begin{pmatrix} P_1^{-1}\mu_1 + R^{-1}\hat{x} \\ P_2^{-1}\mu_2 + R^{-1}\hat{x} \end{pmatrix}$$

Gibbs sampler

- We define the following iterative schema to generate a Markov Chain

$$\begin{aligned} s_1^{(t+1)} &\sim p(s_1|s_2^{(t)}, x = \hat{x}) && \propto \phi(s_1, s_2^{(t)}) \\ s_2^{(t+1)} &\sim p(s_2|s_1^{(t+1)}, x = \hat{x}) && \propto \phi(s_1^{(t+1)}, s_2) \end{aligned}$$

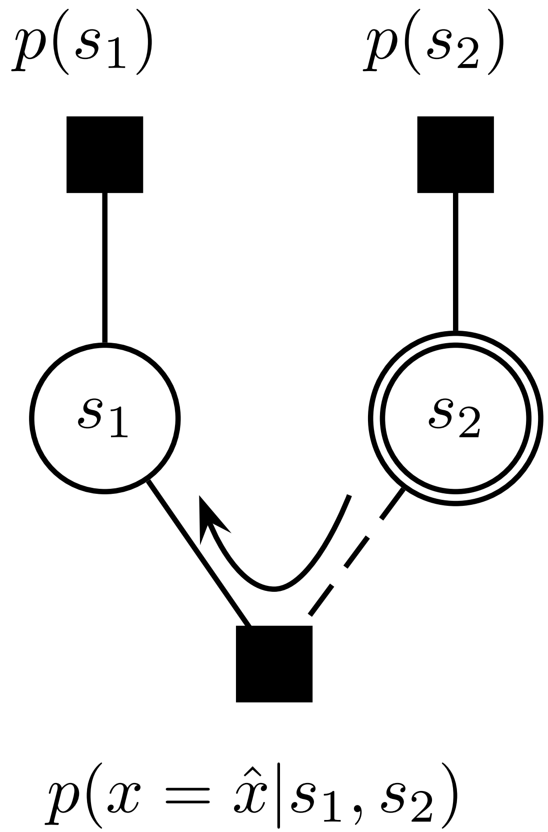
- The desired posterior $\mathcal{P}$ is the stationary distribution of T (why? – later...).

- A remarkable fact is that we can estimate any desired expectation by ergodic averages

$$\langle f(\mathbf{s}) \rangle_{\mathcal{P}} \approx \frac{1}{t - t_0} \sum_{n=t_0}^{t} f(\mathbf{s}^{(n)})$$

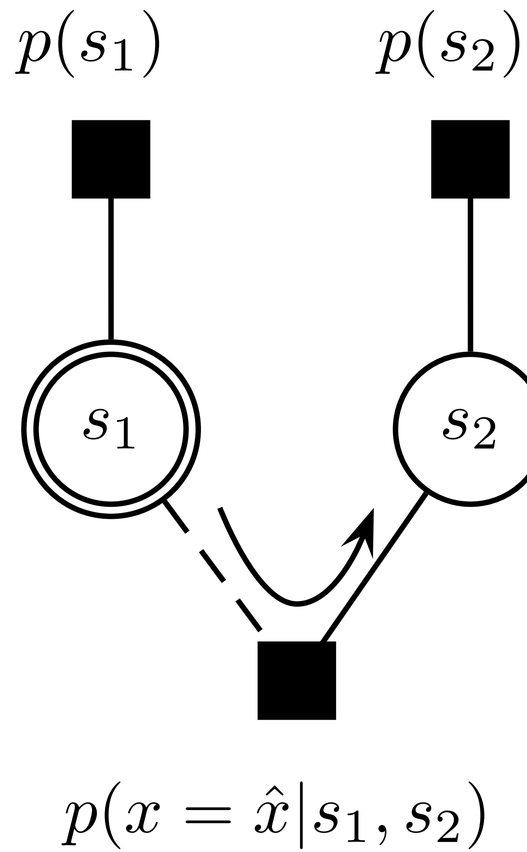
- Consecutive samples $\mathbf{s}^{(t)}$ are dependent but we can “pretend” as if they are independent!

Gibbs Sampling



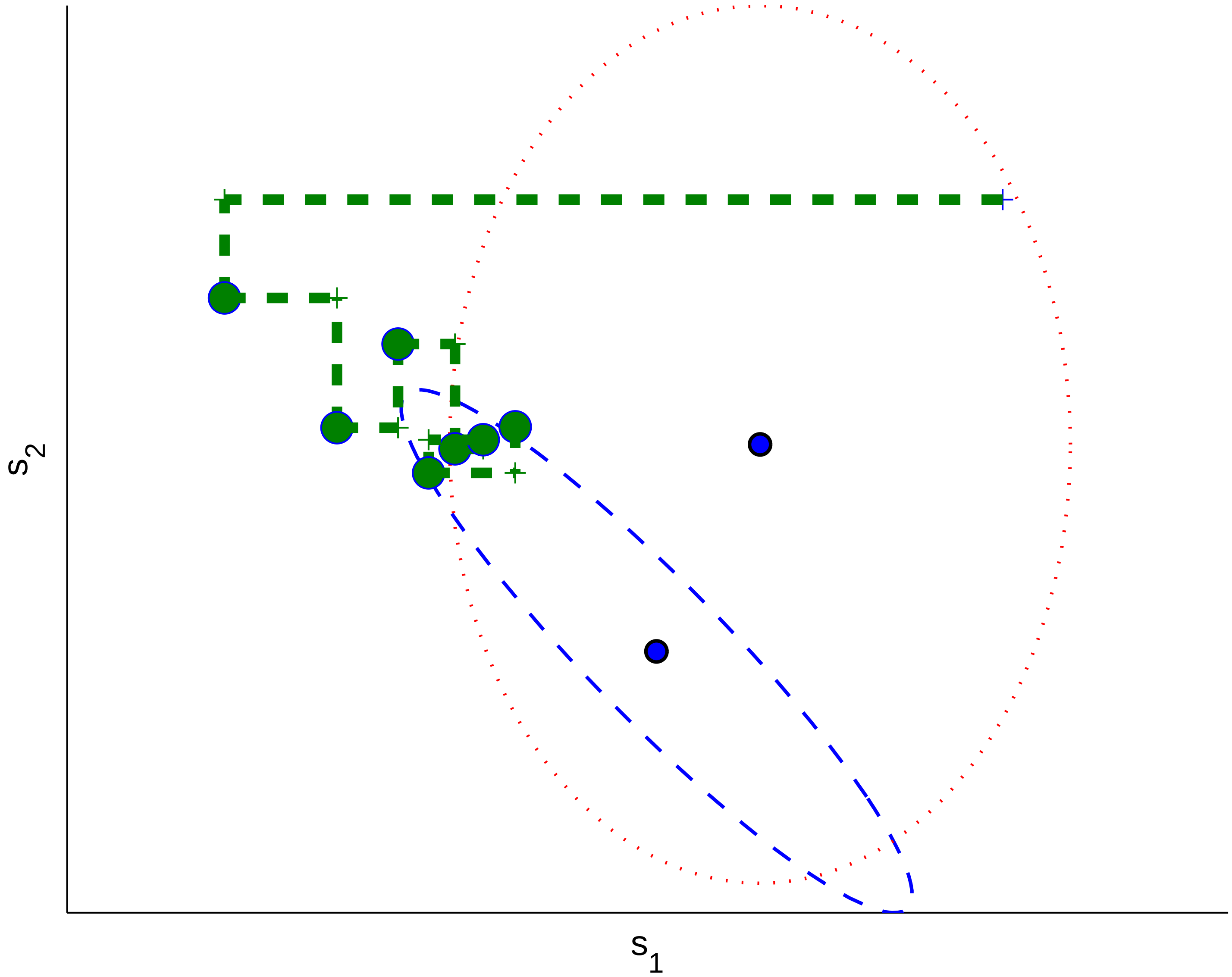
$$\color{red}{s_1}^{(t+1)} \sim \mathcal{N}(s_1; m_1(s_2^{(t)}), S_1)$$

Gibbs Sampling

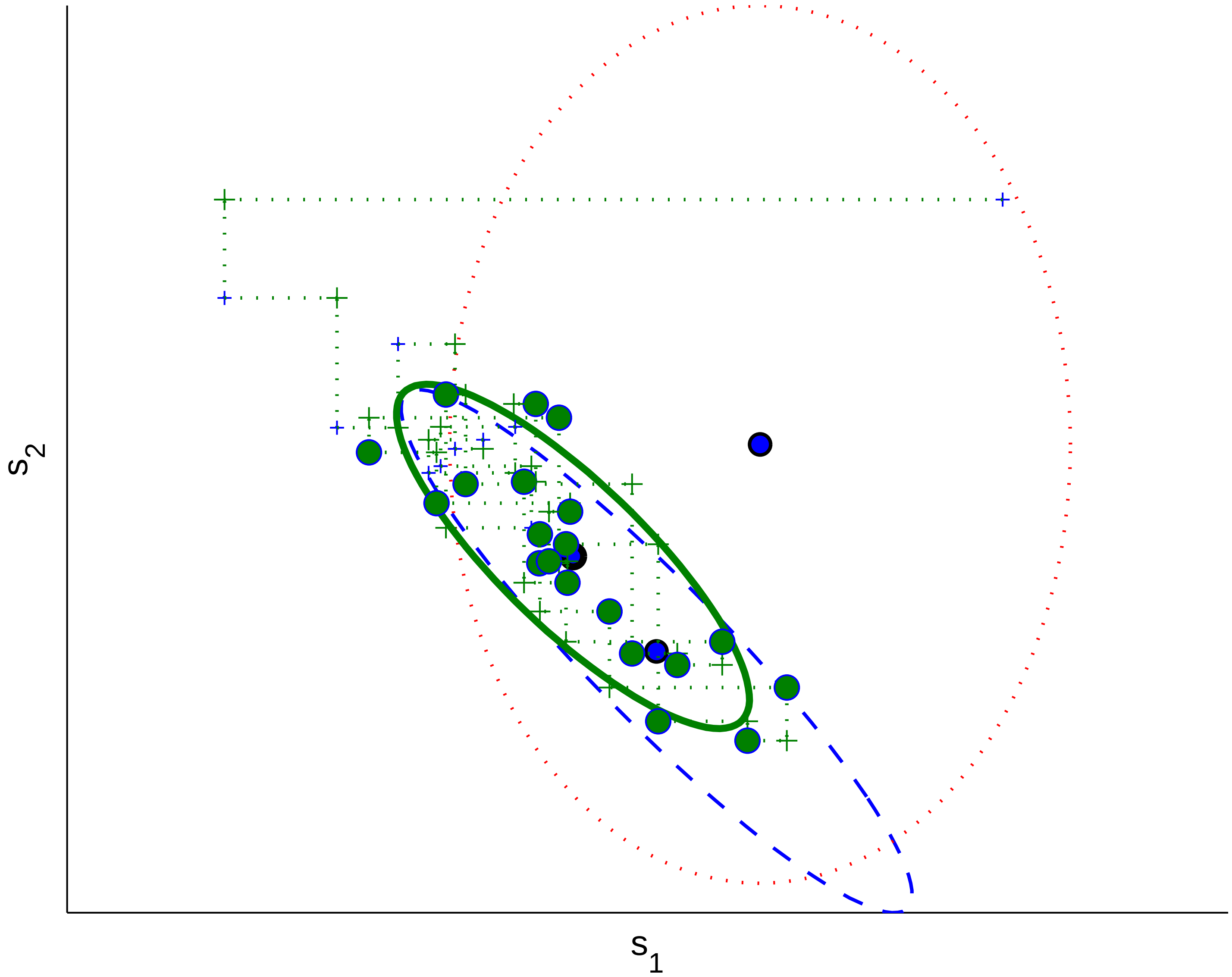


$$s_2^{(t+1)} \sim \mathcal{N}(s_2; m_2(s_1^{(t+1)}), S_2)$$

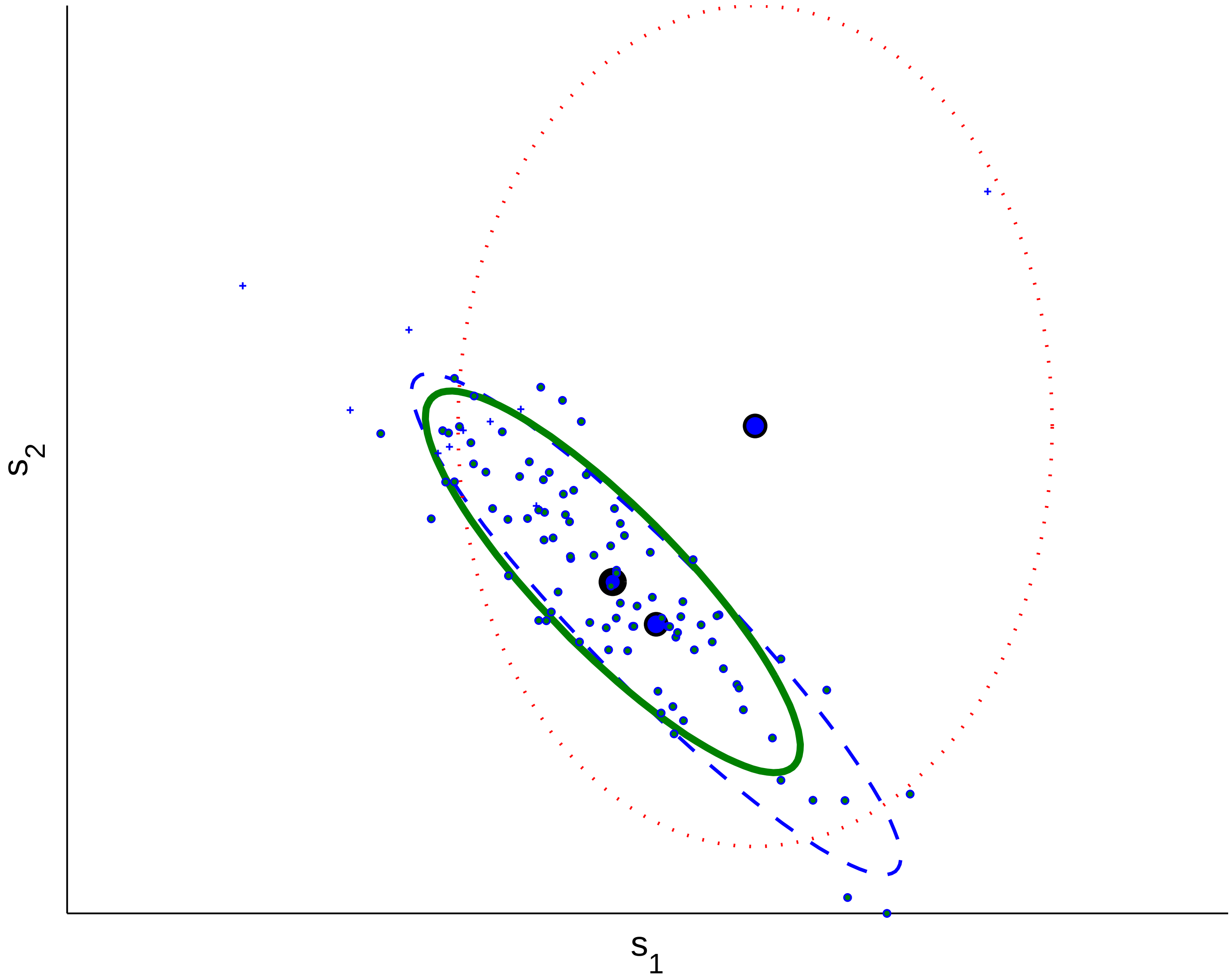
Gibbs Sampling



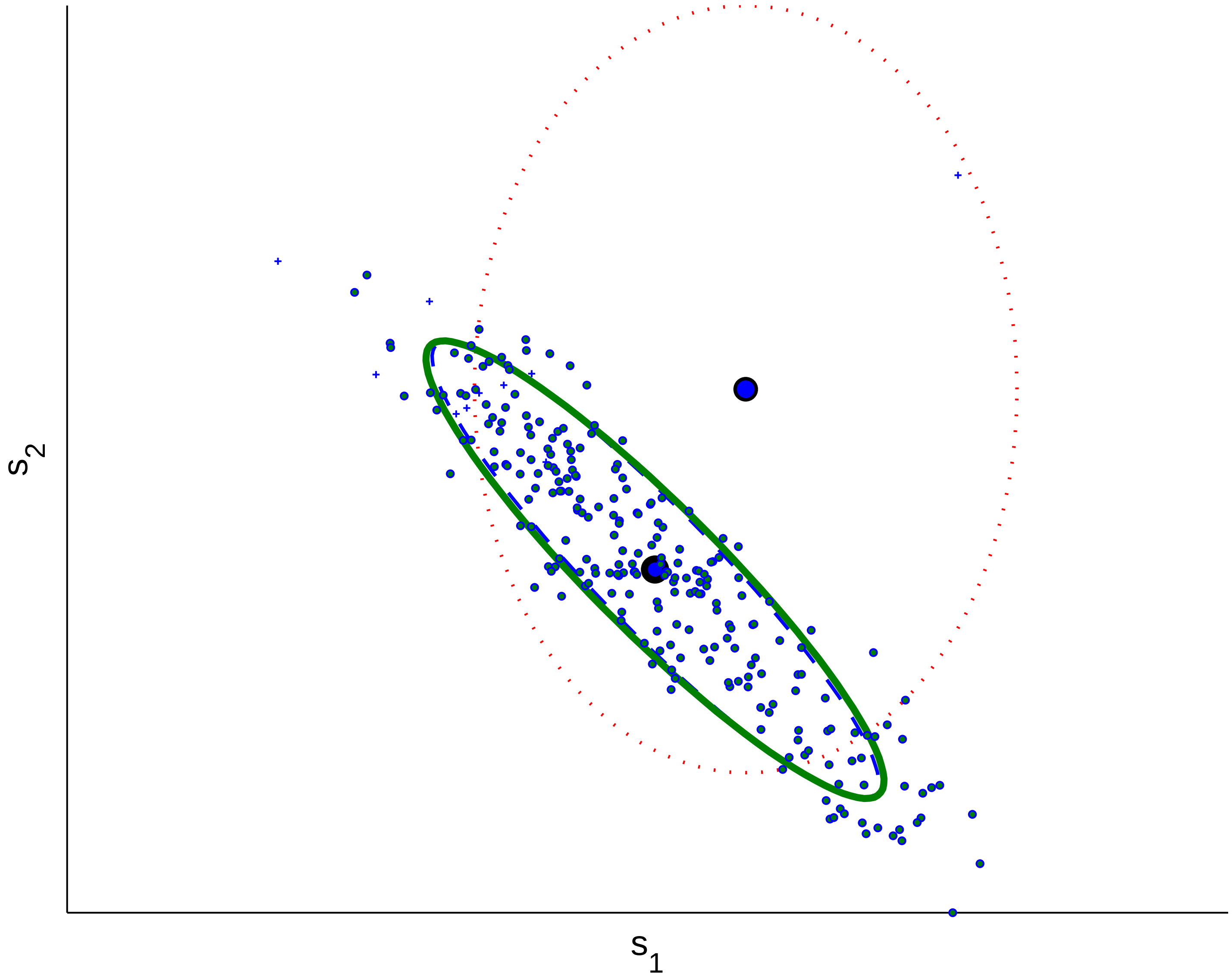
Gibbs Sampling, $t = 20$



Gibbs Sampling, $t = 100$



Gibbs Sampling, $t = 250$



Finding the full conditionals

$$s_1{}^{(t+1)} \quad \sim \quad p(s_1|s_2^{(t)}, x = \hat{x}) \qquad \propto \quad \phi(s_1, s_2^{(t)})$$

Eliminate terms that don't depend on s_1

$$\log \phi(s_1, s_2^{(t)}) \quad = \quad \log p(x = \hat{x}|s_1, s_2^{(t)}) \; + \log p(s_1) \; + \log p(s_2^{(t)})$$

$$=^+ \quad \underbrace{\mu_1^\top P_1^{-1} s_1 - \frac{1}{2} s_1{}^\top P_1^{-1} s_1}_{\log p(s_1)} + \underbrace{\hat{x}^\top R^{-1}(s_1 + s_2^{(t)}) - \frac{1}{2}(s_1 + s_2^{(t)})^\top R^{-1}(s_1 + s_2^{(t)})}_{p(x=\hat{x}|s_1, s_2^{(t)})}$$

$$=^+ \quad \left(\mu_1^\top P_1^{-1} + (\hat{x} - s_2^{(t)})^\top R^{-1}\right) s_1 - \frac{1}{2} \mathbf{Tr} \left(P_1^{-1} + R^{-1}\right) s_1 s_1{}^\top$$

$$p(s_1|s_2^{(t)}, x = \hat{x}) \quad = \quad \mathcal{N}(s_1; m_1, S_1)$$

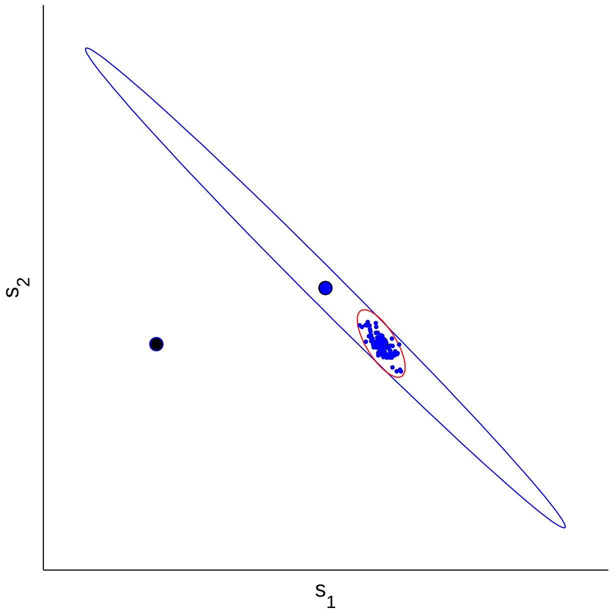
$$S_1 \quad = \quad \left(P_1^{-1} + R^{-1}\right)^{-1} \qquad\qquad m_1(s_2^{(t)}) = S_1 \left(P_1^{-1}\mu_1 + R^{-1}(\hat{x} - s_2^{(t)})\right)$$

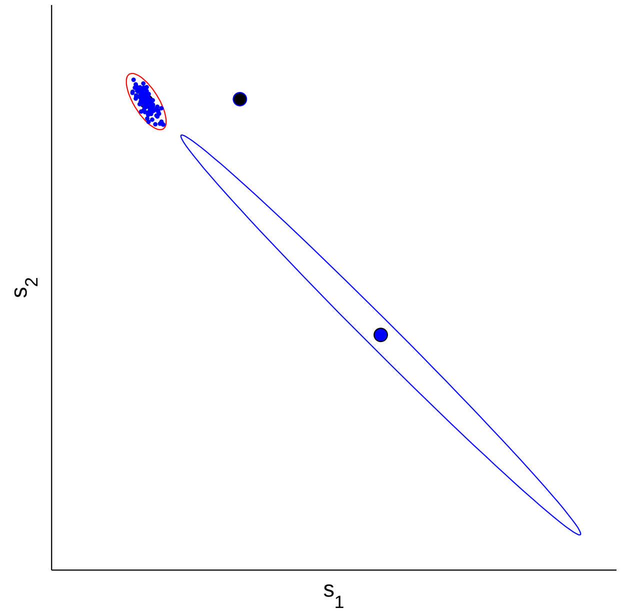
The transition kernel

$$
\begin{aligned}
T(s_1^{(t+1)}, s_2^{(t+1)} | s_1^{(t)}, s_2^{(t)}) &= T(s_2^{(t+1)} | s_1^{(t+1)}, s_1^{(t)}, s_2^{(t)}) T(s_1^{(t+1)} | s_1^{(t)}, s_2^{(t)}) \\
&= T(s_2^{(t+1)} | s_1^{(t+1)}) T(s_1^{(t+1)} | s_2^{(t)}) \\
&= \mathcal{N}(s_2^{(t+1)}; m_2(s_1^{(t+1)}), S_2) \mathcal{N}(s_1^{(t+1)}; m_1(s_2^{(t)}), S_1)
\end{aligned}
$$

Therefore, the transition kernel is also Gaussian.

The transition kernel





But why does the chain converge to the target distribution?

Markov Chain Monte Carlo (MCMC)

- Construct a transition kernel $T(\mathbf{s}'|\mathbf{s})$ with the stationary distribution $\mathcal{P} = \phi(\mathbf{s})/Z_x \equiv \pi(\mathbf{s})$ for any initial distribution $r(\mathbf{s})$.

$$\pi(\mathbf{s}) = T^{\infty} r(\mathbf{s}) \tag{1}$$

- Sample $\mathbf{s}^{(0)} \sim r(\mathbf{s})$
- For $t = 1 \dots \infty$, Sample $\mathbf{s}^{(t)} \sim T(\mathbf{s}|\mathbf{s}^{(t-1)})$
- Estimate any desired expectation by the average

$$\langle f(\mathbf{s}) \rangle_{\pi(\mathbf{s})} \approx \frac{1}{t - t_0} \sum_{n=t_0}^{t} f(\mathbf{s}^{(n)})$$

where t_0 is a preset burn-in period.

But how to construct T and verify that $\pi(\mathbf{s})$ is indeed its stationary distribution ?

Proof Technique

- Show that the target distribution is a stationary distribution of the Markov chain
  - Verify detailed balance

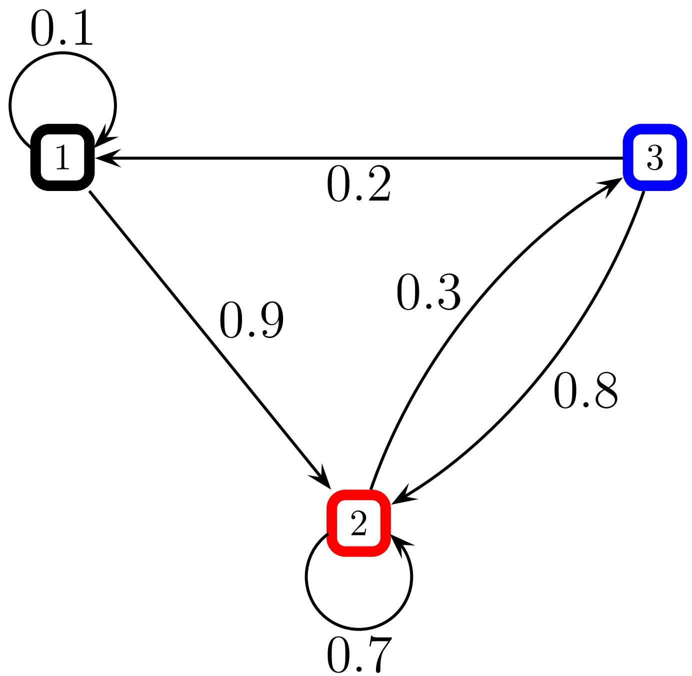
- Show that the transition kernel T has a unique stationary distribution
  - Verify irreducibility and aperiodicity $\Rightarrow$ unique stationary distribution
    - Irreducibility (probabilisic connectedness): Every state s' can be reached from every s

$$T(s'|s) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$ is **not** irreducible

    - Aperiodicity : Cycling around is not allowed

$$T(s'|s) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$ is **not** aperiodic

Reminder of Theory of Markov Chains



$$\begin{pmatrix} 0.1 & 0 & 0.2 \\ 0.9 & 0.7 & 0.8 \\ 0 & 0.3 & 0 \end{pmatrix}$$

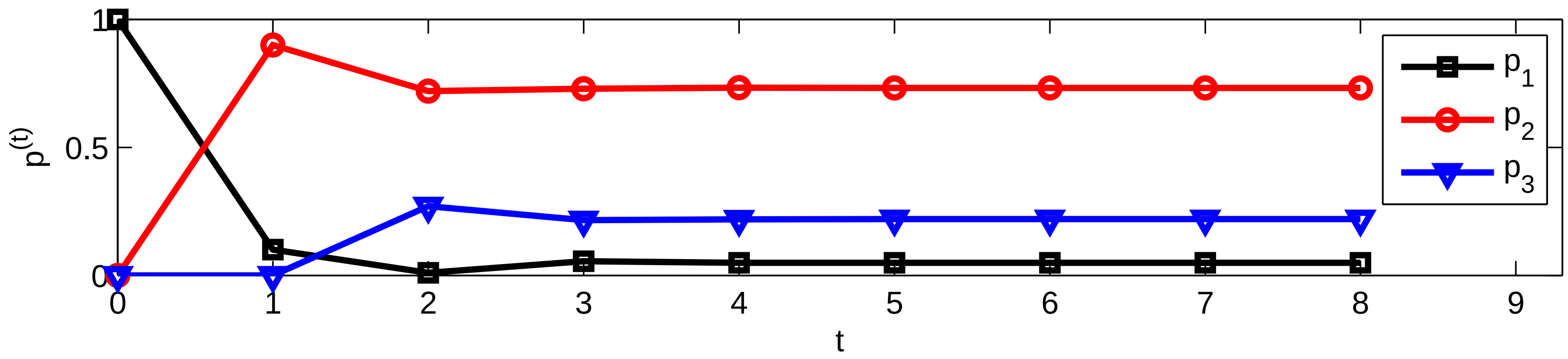
- Suppose the inital state is 1, we have

$$p^{(1)} = \mathbf{T}p^{(0)} = \begin{pmatrix} 0.1 & 0 & 0.2 \\ 0.9 & 0.7 & 0.8 \\ 0 & 0.3 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.9 \\ 0 \end{pmatrix}$$

Numeric Example

- Continue

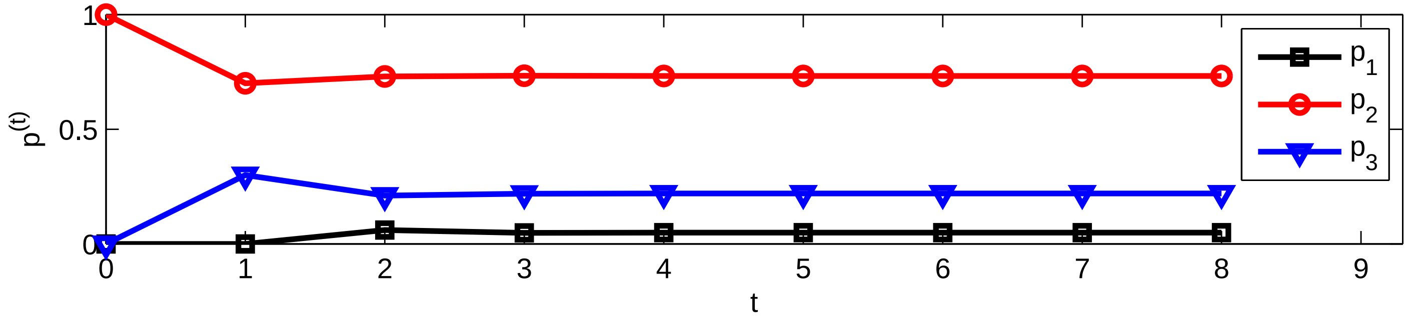
$$
\begin{aligned}
p^{(2)} &= \mathbf{T}\begin{pmatrix} 0.1 \\ 0.9 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.01 \\ 0.72 \\ 0.27 \end{pmatrix} \\
p^{(3)} &= \mathbf{T}\begin{pmatrix} 0.01 \\ 0.72 \\ 0.27 \end{pmatrix} = \begin{pmatrix} 0.05 \\ 0.73 \\ 0.22 \end{pmatrix}
\end{aligned}
$$



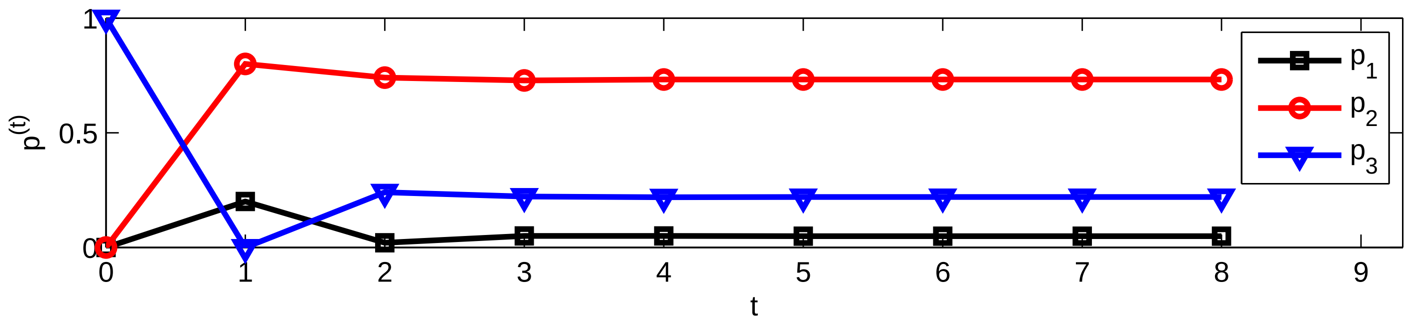
Convergence to a stationary distribution

Starting from other configurations does not alter the picture

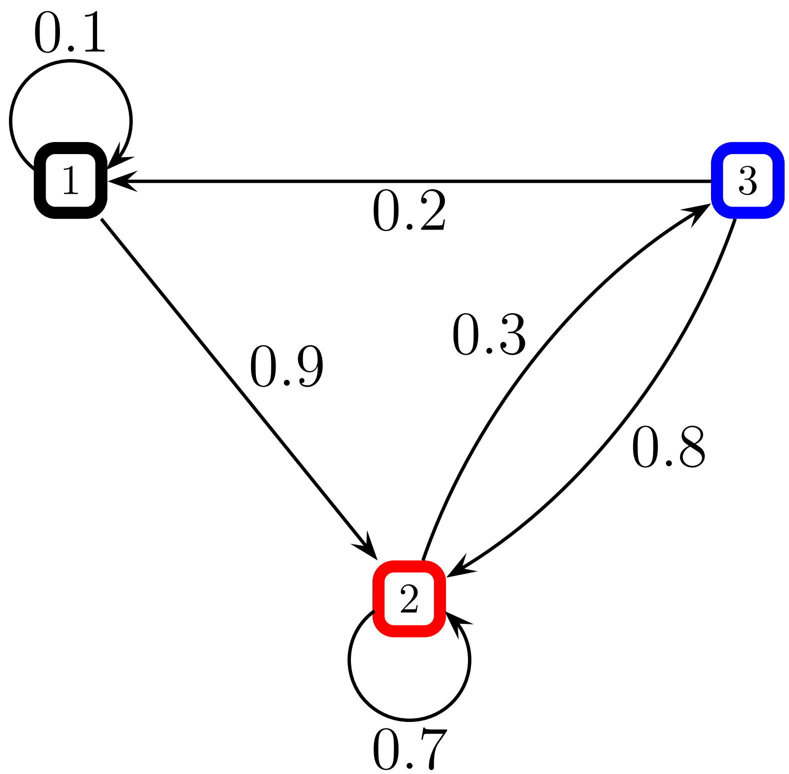
- $p^{(0)} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^\top$



- $p^{(0)} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^\top$



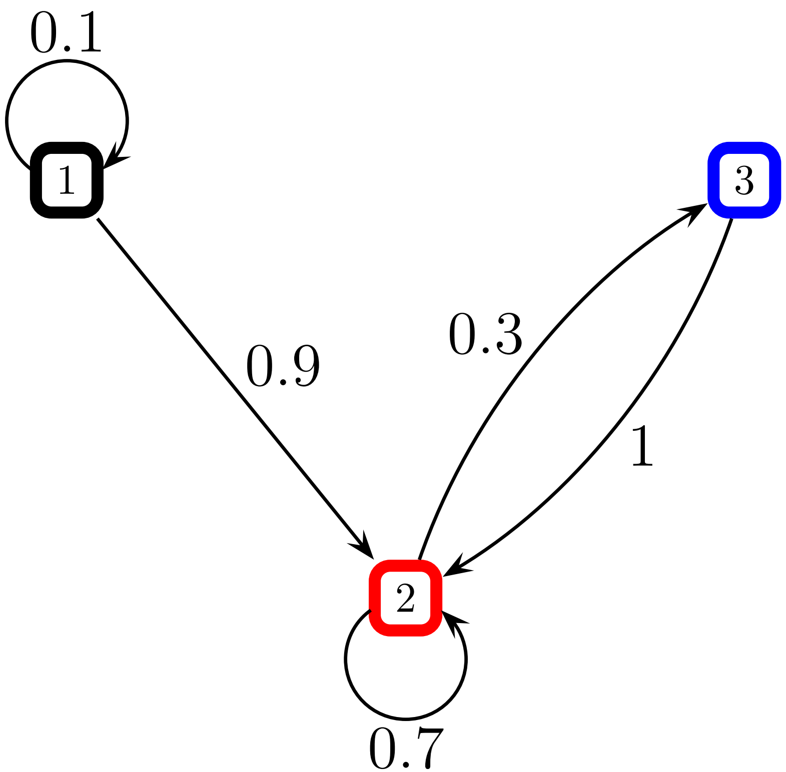
Examples: Irreducable chain



$$\begin{pmatrix} 0.1 & 0 & 0.2 \\ 0.9 & 0.7 & 0.8 \\ 0 & 0.3 & 0 \end{pmatrix}$$

- All states communicate $\Rightarrow$ Chain is said to be irreducable
- All states recurrent

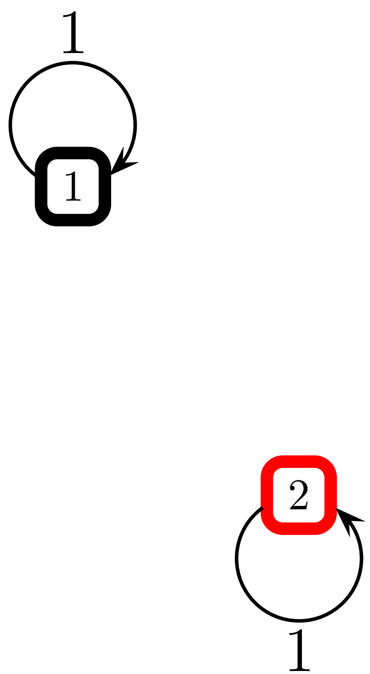
Examples: Transient states



$$\begin{pmatrix} 0.1 & 0 & 0 \\ 0.9 & 0.7 & 1 \\ 0 & 0.3 & 0 \end{pmatrix}$$

- When the chain leaves state 1, it never returns $\Rightarrow$ State 1 is transient

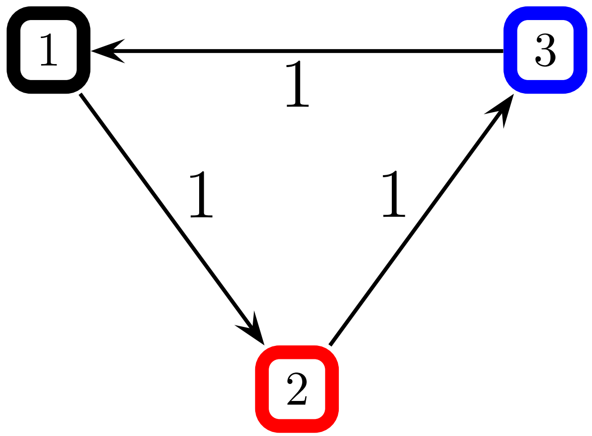
Examples: Reducable chains



$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- Disconnected subgraphs in state transition diagram $\Rightarrow$ Chain is reducable
- No unique stationary distribution

Example: Periodic



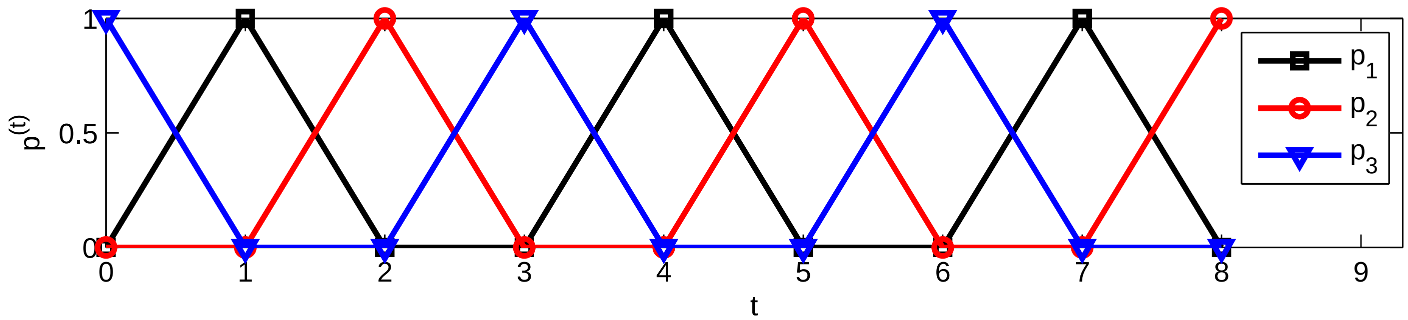
$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

- All states communicate, but ...
- Effect of Initial distribution $p(s^0)$ on $p(s^t)$ does not diminish when $t \to \infty$

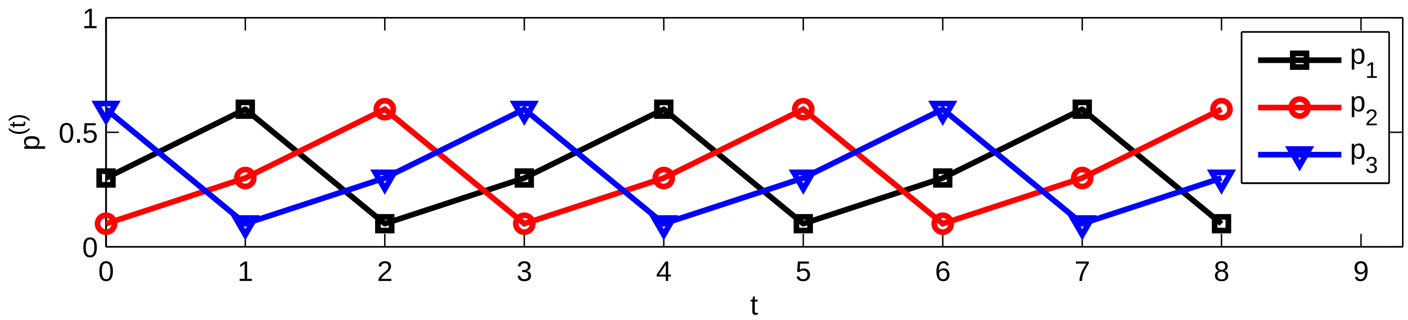
Example: Periodic

There is no stationary distribution

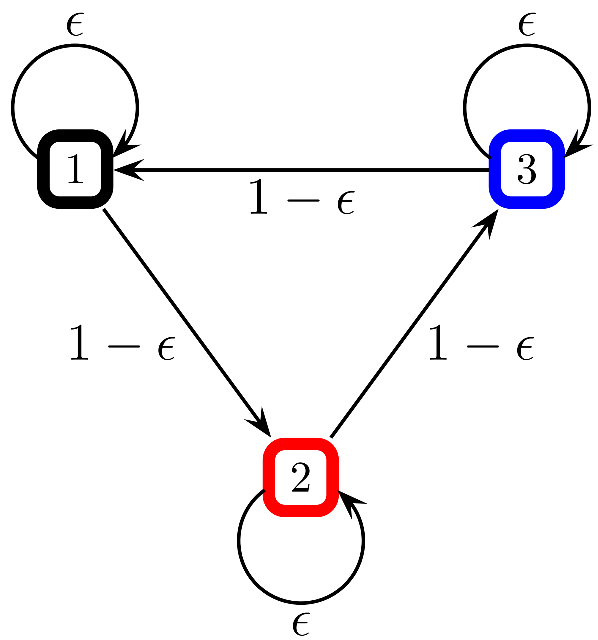
- $p^{(0)} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^\top$



- $p^{(0)} = \begin{pmatrix} 0.3 & 0.1 & 0.6 \end{pmatrix}^\top$



Example: Mixture



$$(1-\epsilon)\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} + \epsilon \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

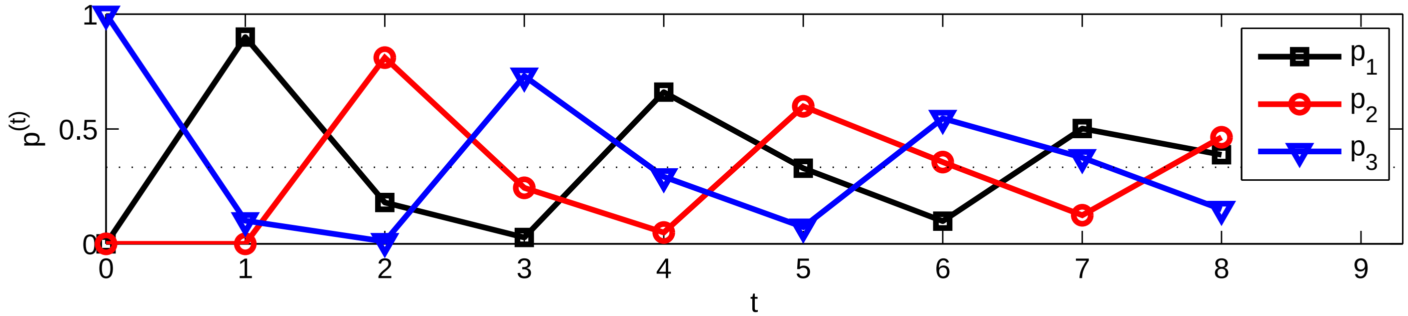
- All states communicate, not periodic

- Is there a unique stationary distribution?

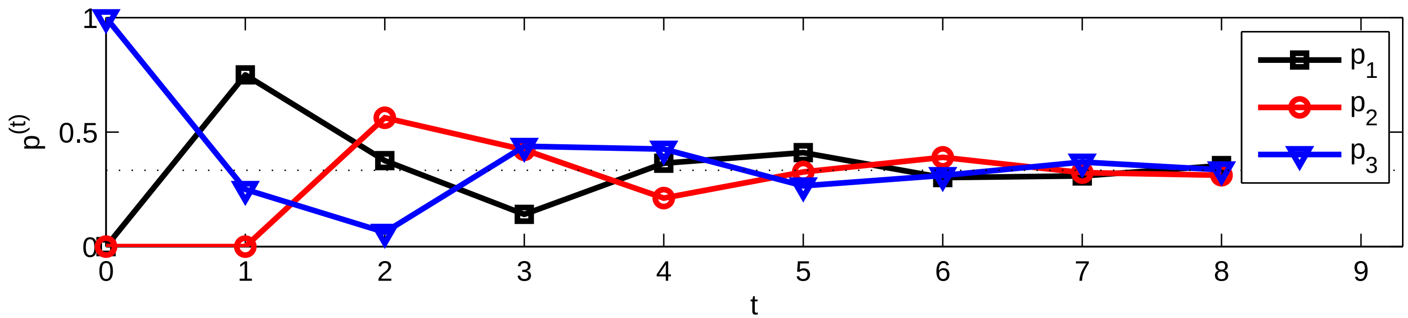
Example: Mixture

- There is a stationary distribution $p^{(\infty)} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}^\top$

- $\epsilon = 0.1$



- $\epsilon = 0.25$

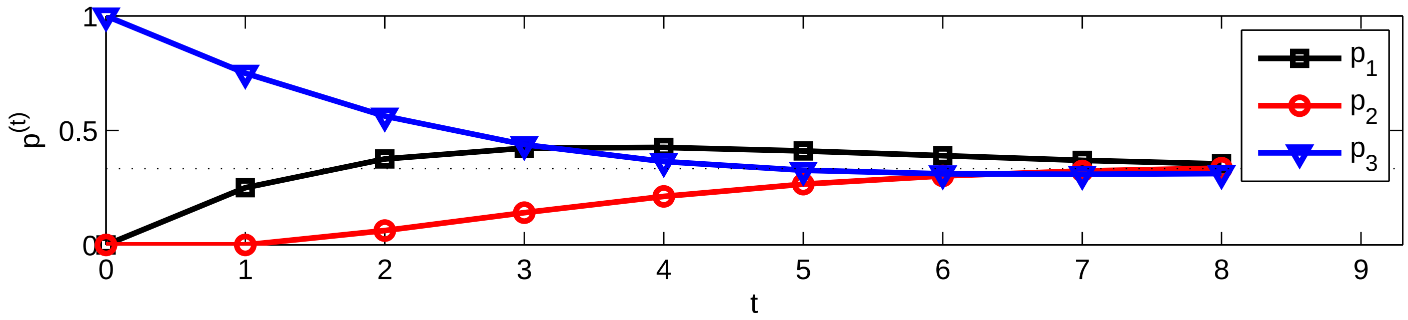


- Convergence rates are different

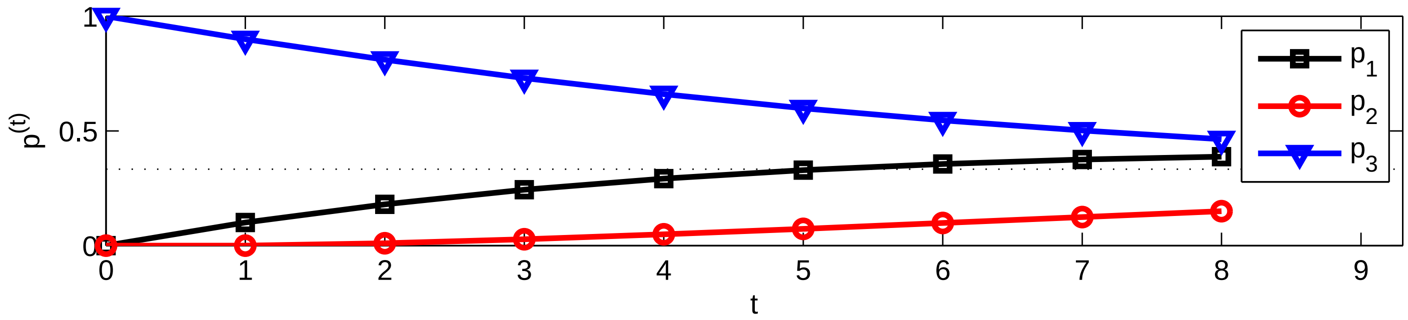
Example: Mixture

- There is a stationary distribution $p^{(\infty)} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}^\top$

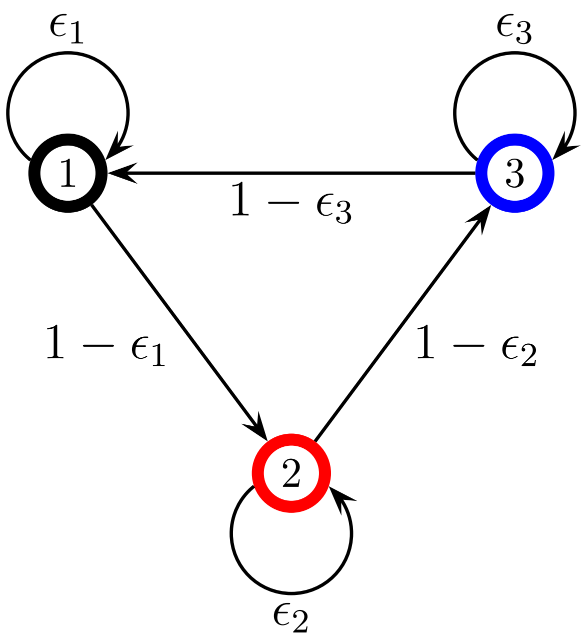
- $\epsilon = 0.75$



- $\epsilon = 0.9$



Example



$$\begin{pmatrix} \epsilon_1 & 0 & 1-\epsilon_3 \\ 1-\epsilon_1 & \epsilon_2 & 0 \\ 0 & 1-\epsilon_2 & \epsilon_3 \end{pmatrix}$$

- Self transition probabilities $\epsilon_1 > \epsilon_2 > \epsilon_3 \Rightarrow p_1^{(\infty)} > p_2^{(\infty)} > p_3^{(\infty)}$, but the exact relationship is not trivial
- How can we find the stationary distribution ? How fast is the convergence ?
- How can we design a chain that will converge to a given target distribution ?

Stationary Distribution

- We compute an eigendecomposition

$$\begin{aligned} \mathbf{T} &= B \Lambda B^{-1} \\ \Lambda &= \mathbf{diag}(1, \lambda_2, \dots, \lambda_K) \end{aligned}$$

- The stationary distribution is given by the limit

$$\begin{aligned} \lim_{t \to \infty} p^{(t)} &= \lim_{t \to \infty} \mathbf{T}^t p^{(0)} \\ \mathbf{T}^t &= B \Lambda B^{-1} B \Lambda \dots \Lambda B^{-1} = B \Lambda^t B^{-1} \end{aligned}$$

- It turns out since $\mathbf{T}$ is a conditional probability matrix (columns sum up to one), the eigenvalues satisfy

$$1 = \lambda_1 \geq |\lambda_2| \geq |\lambda_3| \geq \dots \leq |\lambda_K|$$

Stationary Distribution

- If and only if $|\lambda_2| < 1$

$$\mathbf{T}^t = B \begin{pmatrix} 1 & 0 & & 0 \\ 0 & \lambda_2^t & & 0 \\ & & \ddots & \\ 0 & & & \lambda_K^t \end{pmatrix} B^{-1} \xrightarrow{t \to \infty} \quad B \begin{pmatrix} 1 & 0 & & 0 \\ 0 & 0 & & 0 \\ & & \ddots & \\ 0 & & & 0 \end{pmatrix} B^{-1}$$

$$= \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_K \end{pmatrix} \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$$

- Geometric Convergence property, there exist $c > 0$ s.t.

$$\|\mathbf{T}^t p^{(0)} - \pi\|_{\text{var}} \leq c|\lambda_2|^t$$

- However, it is hard to show algebraically that $|\lambda_2| < 1$. Fortunately, there is a...

Convergence Theorem (for finite-state Markov Chains)

- Finite State space $\mathcal{X} = \{1, 2, \dots, K\}$
- $\mathbf{T}$ is irreducable and aperiodic, then there exist $0 < r < 1$ and $c > 0$ s.t.

$$\|\mathbf{T}^t p^{(0)} - \pi\|_{\text{var}} \leq c r^t$$

where π is the invariant distribution

$$\|P - Q\|_{\text{var}} \equiv \frac{1}{2} \sum_{s \in \mathcal{X}} |P(s) - Q(s)|$$

MCMC Equilibrium condition = Detailed Balance

$$T(\mathbf{s}|\mathbf{s}')\pi(\mathbf{s}') = T(\mathbf{s}'|\mathbf{s})\pi(\mathbf{s})$$

If detailed balance is satisfied then $\pi(\mathbf{s})$ is a stationary distribution

$$\pi(\mathbf{s}) = \int d\mathbf{s}' T(\mathbf{s}|\mathbf{s}')\pi(\mathbf{s}')$$

If the configuration space is discrete, we have

$$\pi(\mathbf{s}) = \sum_{\mathbf{s}'} T(\mathbf{s}|\mathbf{s}')\pi(\mathbf{s}')$$

$$\pi = T\pi$$

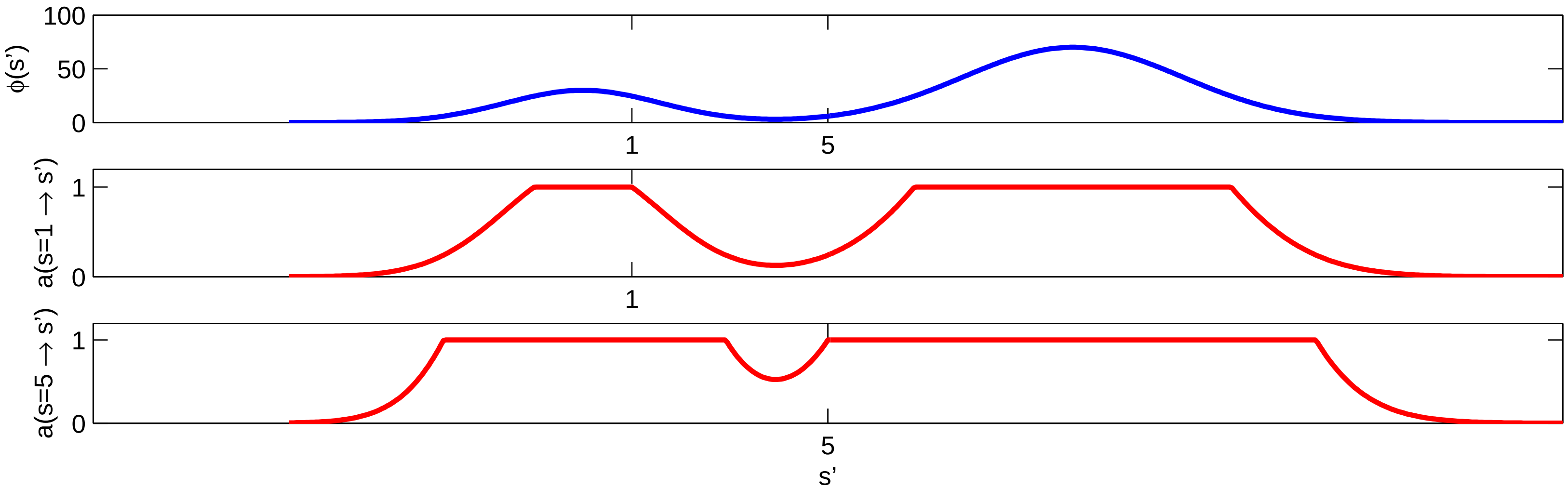
π has to be a (right) eigenvector of T.

Metropolis-Hastings Kernel

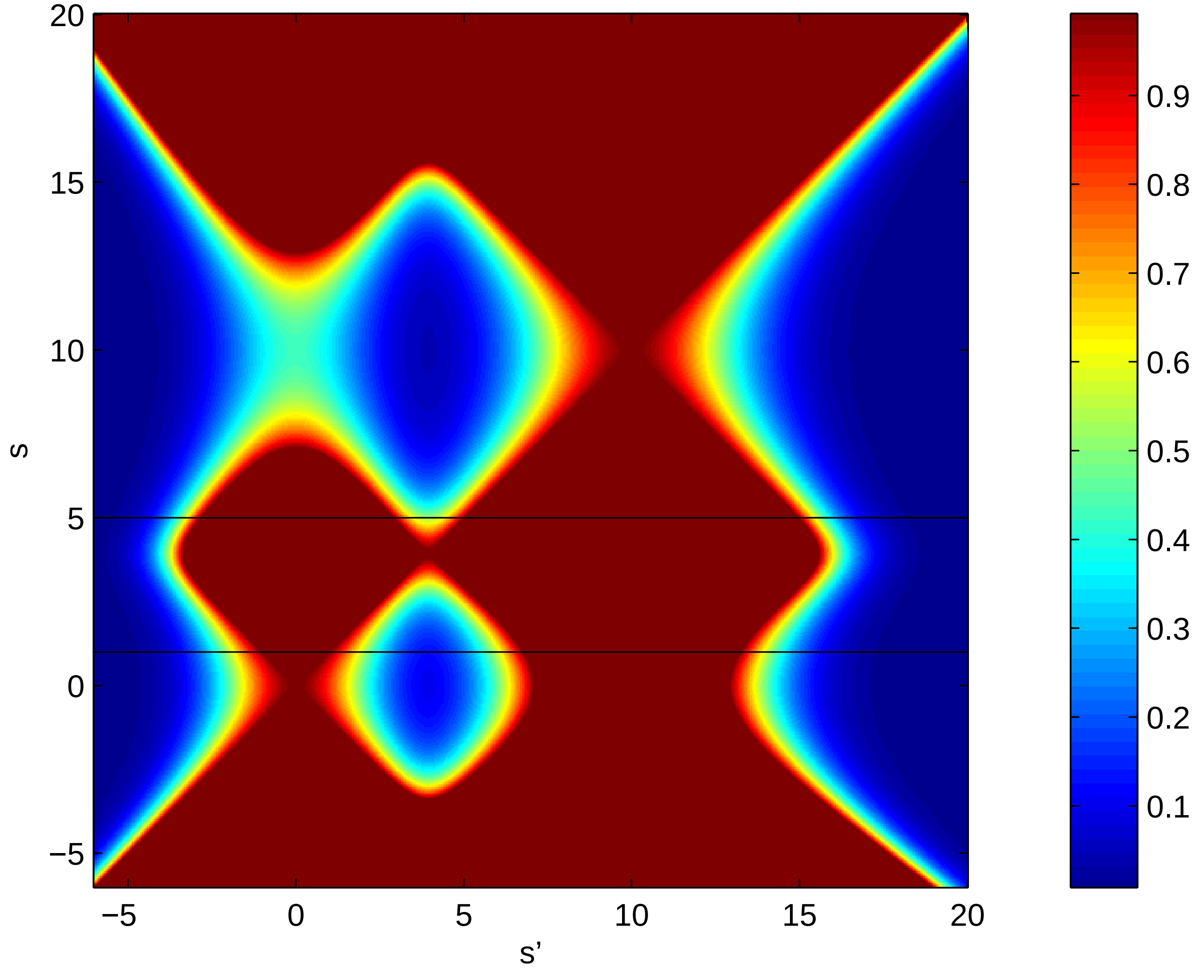
- We choose an arbitrary proposal distribution $q(s'|s)$ (that satisfies mild regularity conditions).
  (When q is symmetric, i.e., $q(s'|s) = q(s|s')$, we have a Metropolis algorithm.)

- We define the *acceptance probability* of a jump from s to s' as

$$a(s \to s') \equiv \min\{1, \frac{q(s|s')\pi(s')}{q(s'|s)\pi(s)}\}$$



Acceptance Probability $a(s \to s')$



Basic MCMC algorithm: Metropolis-Hastings

1. Initialize: $s^{(0)} \sim r(s)$

2. For $t = 1, 2, \dots$

   - Propose:

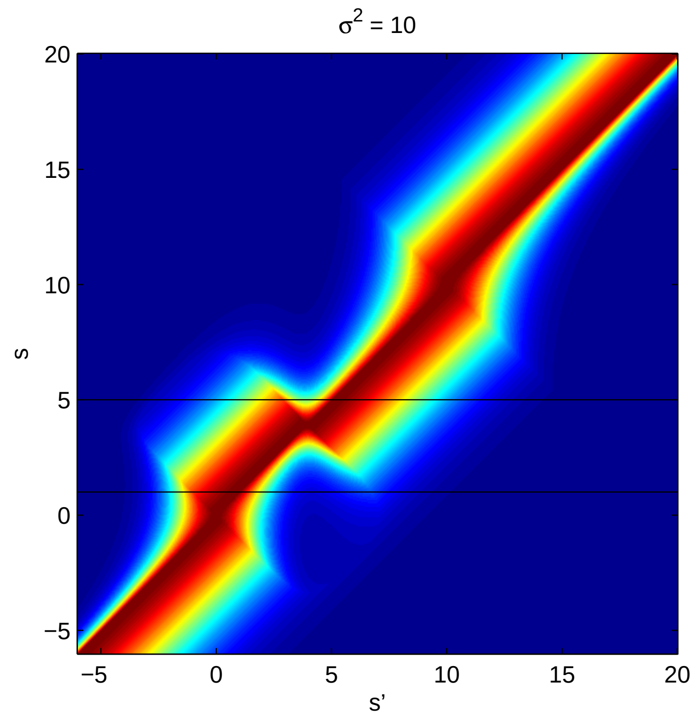
$$s' \sim q(s'|s^{(t-1)})$$

   - Evaluate Proposal: $u \sim$ Uniform$[0, 1]$

$$s^{(t)} \quad := \quad \begin{cases} s' & u < a(s^{(t-1)} \to s') \quad \text{Accept} \\ s^{(t-1)} & \text{otherwise} \quad \text{Reject} \end{cases}$$

Transition Kernel of the Metropolis-Hastings

$$T(s'|s) = \underbrace{q(s'|s)a(s \to s')}_{\text{Accept}} + \underbrace{\delta(s'-s) \int ds' q(s'|s)(1 - a(s \to s'))}_{\text{Reject}}$$



Only Accept part for visual convenience

Verification of detailed balance for Metropolis

$$
\begin{aligned}
\pi(s) &= \frac{1}{Z}\phi(s) \\
a(s \to s') &= \min\{1, \frac{\pi(s')}{\pi(s)}\} = \min\{1, \frac{\phi(s')}{\phi(s)}\} \qquad q(s|s') = q(s'|s)
\end{aligned}
$$

$$
\begin{aligned}
T(s'|s)\pi(s) &= q(s'|s)\min\{1, \frac{\phi(s')}{\phi(s)}\}\pi(s) \quad \{+\delta(s - s')\pi(s)\dots\} \\
&= q(s'|s)\min\{\frac{\phi(s)}{Z}, \frac{\phi(s')}{\phi(s)}\frac{\phi(s)}{Z}\} \\
&= q(s'|s)\min\{\frac{\phi(s)}{Z}, \frac{\phi(s')}{Z}\} \\
&= q(s|s')\frac{\phi(s')}{Z}\min\{\frac{\phi(s)/Z}{\phi(s')/Z}, 1\} = T(s|s')\pi(s')
\end{aligned}
$$

Verification of detailed balance for Metropolis-Hastings

$$
\begin{aligned}
\pi(s) &= \frac{1}{Z}\phi(s) \\
a(s \to s') &= \min\{1, \frac{q(s|s')\pi(s')}{q(s'|s)\pi(s)}\} = \min\{1, \frac{q(s|s')\phi(s')}{q(s'|s)\phi(s)}\}
\end{aligned}
$$

$$
\begin{aligned}
T(s'|s)\pi(s) &= q(s'|s)\min\{1, \frac{q(s|s')\phi(s')}{q(s'|s)\phi(s)}\}\frac{\phi(s)}{Z} \\
&= \min\{q(s'|s)\frac{\phi(s)}{Z}, \frac{q(s|s')\phi(s')}{Z}\} = T(s|s')\pi(s')
\end{aligned}
$$

Verification of detailed balance for Gibbs

- The transition kernel for Gibbs sampler is a product of transition kernels operating on a single coordinate i.

- The transition kernel for a deterministic scan Gibbs sampler is

$$T = \prod_i T_i$$

$$\begin{aligned} \pi(s_i, s_{-i}) &= \frac{1}{Z}\phi(s_i, s_{-i}) \\ q_i(s'_i, s'_{-i}|s_i, s_{-i}) &= \frac{1}{Z_i}\phi(s'_i|s_{-i})\delta(s_{-i} - s'_{-i}) \end{aligned}$$

The acceptance probability is

$$
\begin{aligned}
a(s \to s') &= \min\{1, \frac{q(s|s')\pi(s')}{q(s'|s)\pi(s)}\} \\
&= \min\{1, \frac{\frac{1}{Z_i}\phi(s_i|s'_{-i})\delta(s_{-i} - s'_{-i})\frac{1}{Z}\phi(s'_i, s'_{-i})}{\frac{1}{Z_i}\phi(s'_i|s_{-i})\delta(s_{-i} - s'_{-i})\frac{1}{Z}\phi(s_i, s_{-i})}\} \\
&= \min\{1, \frac{\frac{1}{Z_i}\phi(s_i|s_{-i})\frac{1}{Z}\phi(s'_i, s_{-i})}{\frac{1}{Z_i}\phi(s'_i|s_{-i})\frac{1}{Z}\phi(s_i, s_{-i})}\} \\
&= \min\{1, \frac{\frac{1}{Z_i}\phi(s_i|s_{-i})\frac{1}{Z}\phi(s'_i|s_{-i})\phi(s_{-i})}{\frac{1}{Z_i}\phi(s'_i|s_{-i})\frac{1}{Z}\phi(s_i|s_{-i})\phi(s_{-i})}\} = 1
\end{aligned}
$$

Hence all the moves are accepted by default.

Cascades and Mixtures of Transition Kernels

Let T_1 and T_2 have the same stationary distribution $p(s)$.

Then:

$$\begin{aligned} T_c &= T_1 T_2 \\ T_m &= \nu T_1 + (1-\nu) T_2 \quad 0 \leq \nu \leq 1 \end{aligned}$$

are also transition kernels with stationary distribution $p(s)$.

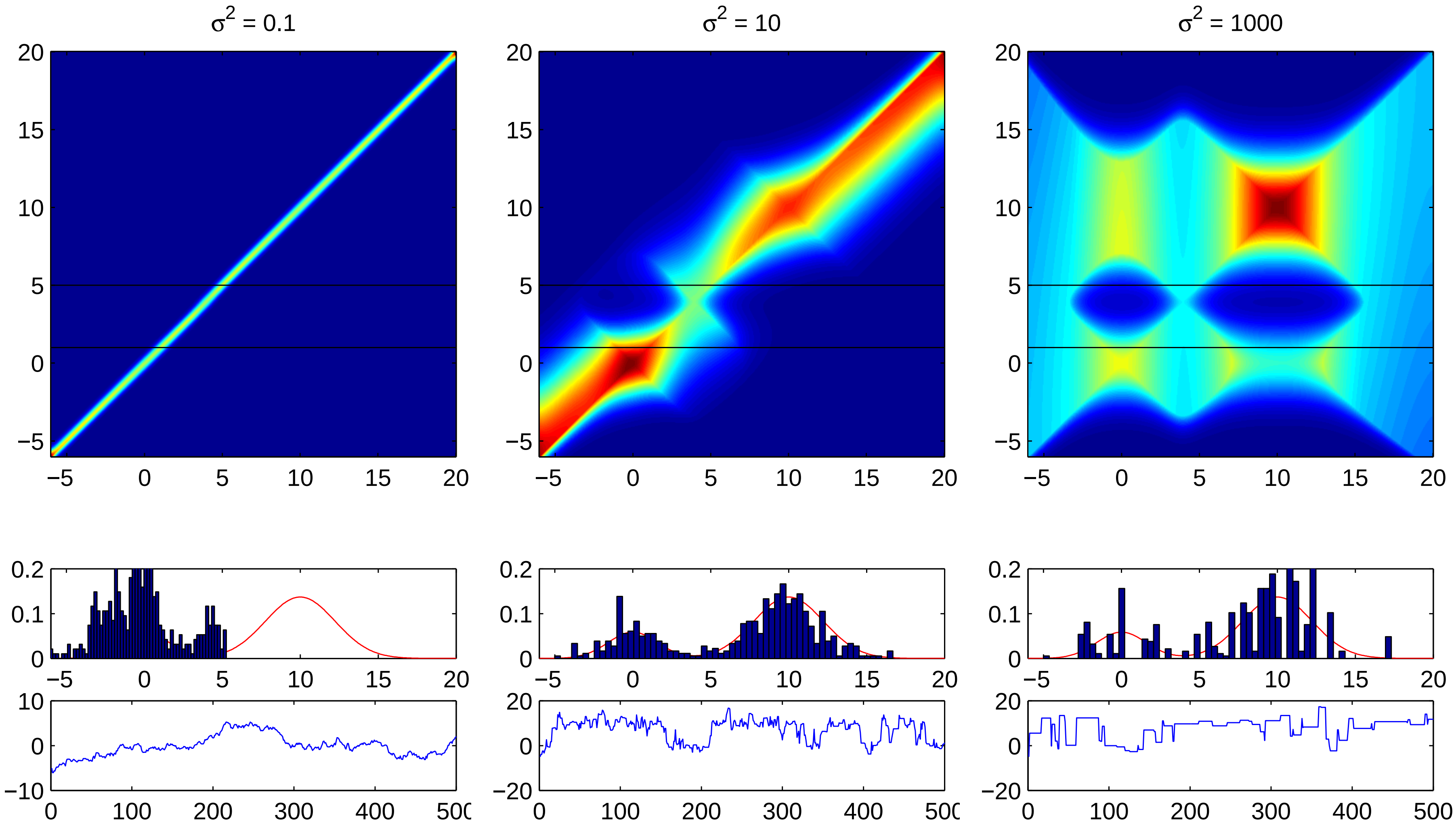
This opens up many possibilities to “tailor” application specific algorithms.

For example let

T_1 : global proposal (allows large “jumps”)
T_2 : local proposal (investigates locally)

We can use T_m and adjust ν as a function of rejection rate.

Various Kernels with the same stationary distribution



$$q(s'|s) = \mathcal{N}(s'; s, \sigma^2)$$

Optimization : Simulated Annealing and Iterative Improvement

For optimization, (e.g. to find a MAP solution)

$$s^* = \arg\max_{s \in \mathcal{S}} \pi(s)$$

The MCMC sampler may not visit s^*.

Simulated Annealing: We define the target distribution as

$$\pi(s)^{\tau_i}$$

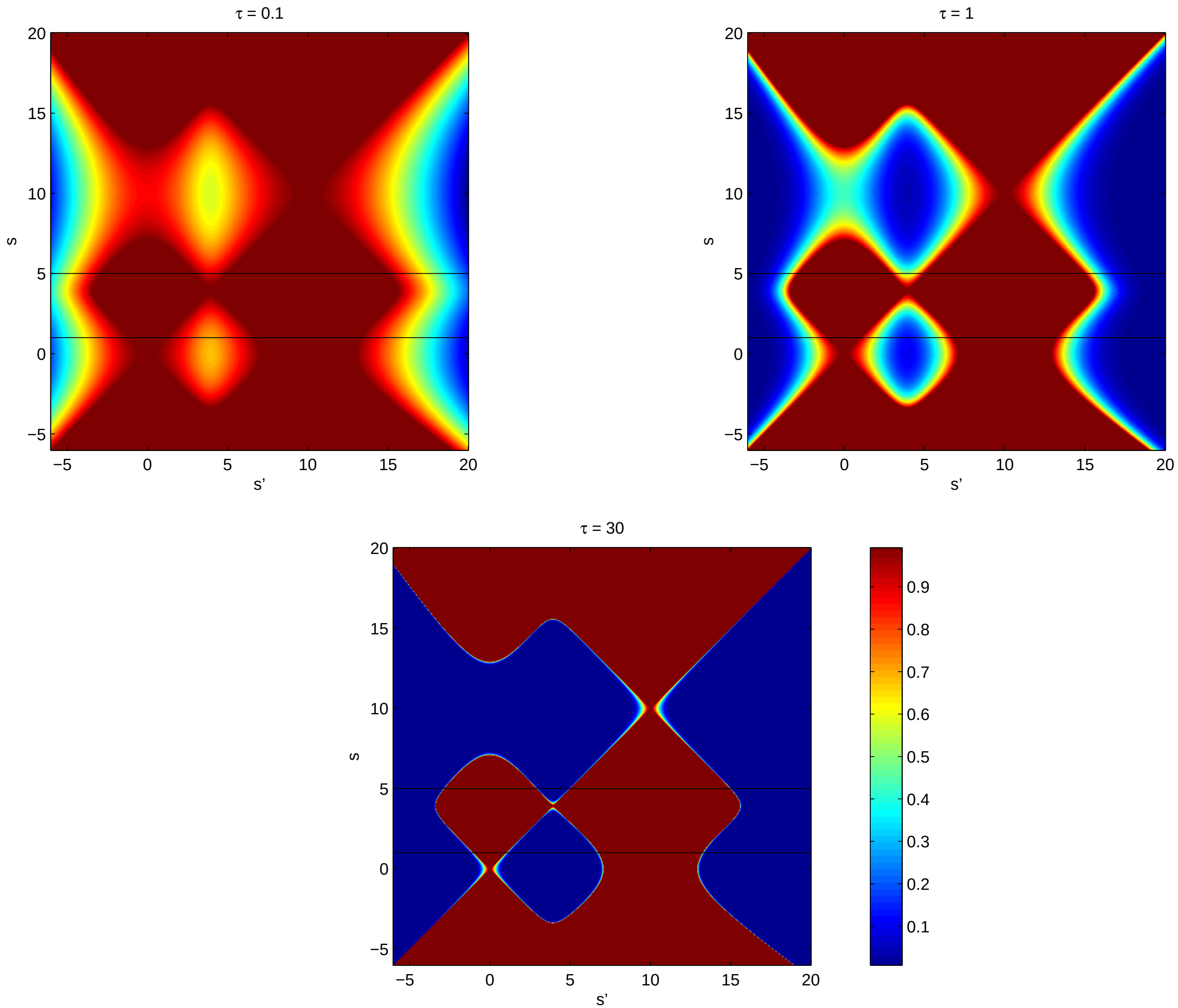
where τ_i is an annealing schedule. For example,

$$\tau_1 = 0.1, \dots, \tau_N = 10, \tau_{N+1} = \infty \dots$$

Iterative Improvement (greedy search) is a special case of SA

$$\tau_1 = \tau_2 = \dots = \tau_N = \infty$$

Acceptance probabilities $a(s \to s')$ at different τ



Summary

- Bayesian Inference,
- Probability models and Graphical model notation
  - Directed Graphical models, Factor Graphs
- The Gibbs sampler
- Metropolis-Hastings, MCMC Transition Kernels
- Sketch of convergence results
- Simulated annealing and iterative improvement

The End

Slides will be available online

`http://www-sigproc.eng.cam.ac.uk/~atc27/papers/5R1/`