CMPE 58N - Lecture 3
Monte Carlo methods

Markov Chains, MCMC, Metropolis-Hastings Algorithm

Department of Computer Engineering,

Boğaziçi University, Istanbul, Turkey

Instructor: A. Taylan Cemgil

Fall 2009

Outline

- Motivations
- Markov Chains
- Metropolis Hastings algorithm

Motivating example: Sampling uniformly from a set S

- Suppose we have a black box implementation of an indicator function $[x \in S]$
- How can we generate uniform samples from S?
- It turns out that the following algorithm works (in principle)

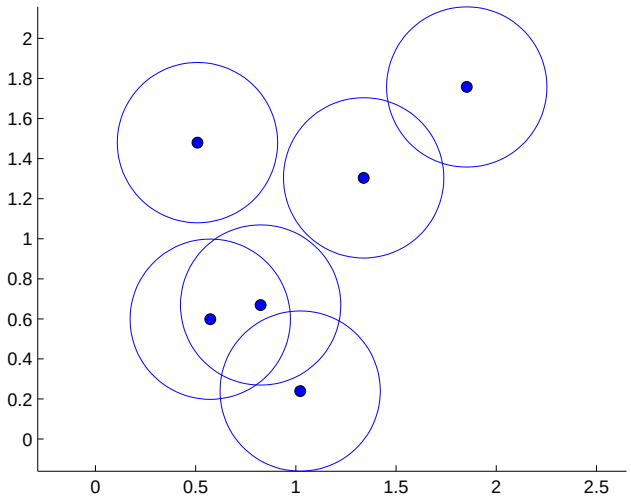
  Choose an arbitrary $x^{(0)} \in S$
  For $i = 1, 2, \ldots$

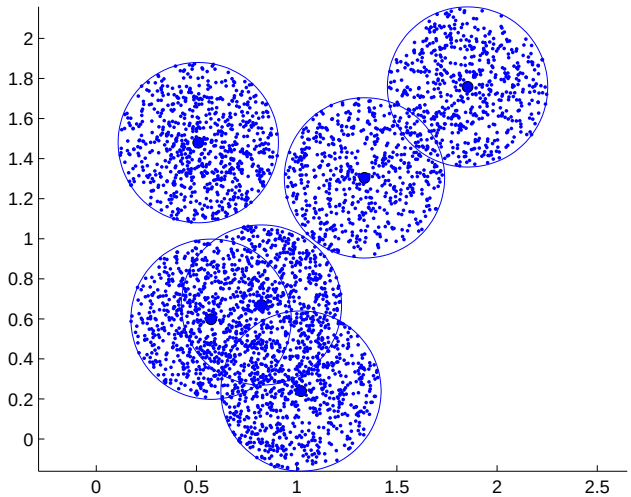
    Propose:
    $\epsilon_i \sim \mathcal{N}(0, V)$
    $x' \leftarrow x^{(i-1)} + \epsilon_i$
    Accept/Reject
    if $[x' \in S]$ then $x^{(i)} \leftarrow x'$ else $x^{(i)} \leftarrow x^{(i-1)}$ endif

  EndFor
  - $x^{(i)}$ are the desired samples! Why?

Motivating example: Sampling uniformly from a set S

$$S = \{x : \|c_i - x\| \leq \rho\}$$



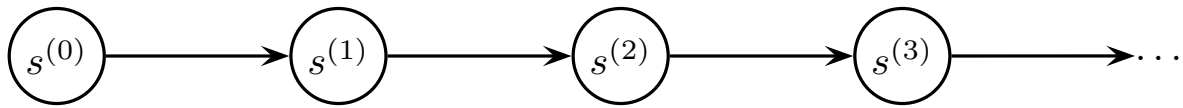


Markov Chains

- Markov Property

$$p(s^{(t)}|s^{(t-1)}, s^{(t-2)}, \ldots, s^{(0)}) = p(s^{(t)}|s^{(t-1)})$$

- Future and past are conditionally independent given current state

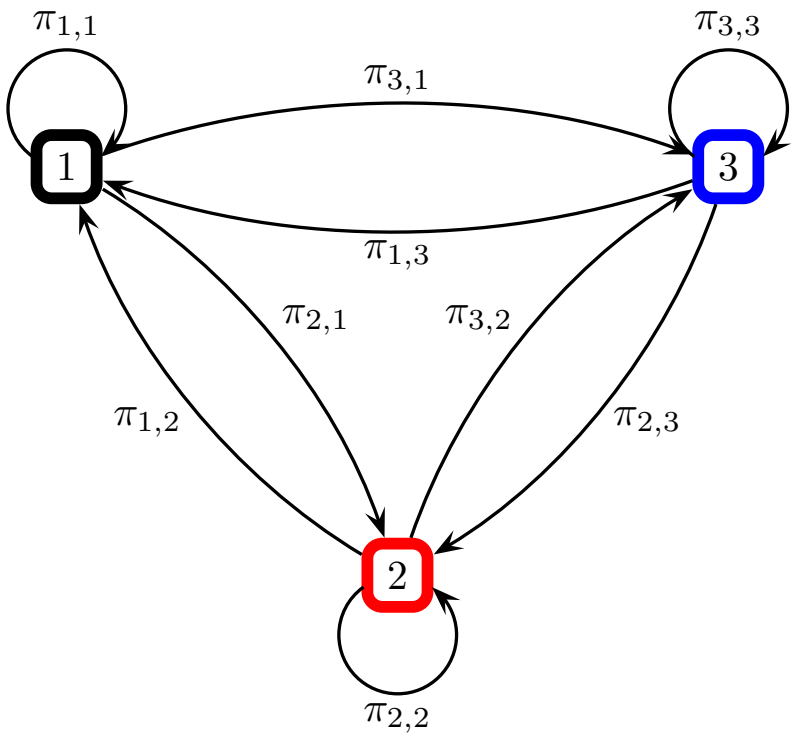


$$p(s^{(0)}, s^{(1)}, s^{(2)}, \ldots) = p(s^{(0)}) \prod_{t=1}^{\infty} p(s^{(t)}|s^{(t-1)})$$

Markov Chains

- Here, we look at Markov chains for analysis of inference algorithms, not for constructing a time series model
- Nature of the State space $\mathcal{X}$ matters ($s^{(t)} \in \mathcal{X}$)
  - $\mathcal{X}$ is Finite ($\leftarrow$ simplest, we look at this)
  - $\mathcal{X}$ is Countable
  - $\mathcal{X}$ is Uncountable ($\leftarrow$ actually needed but quite technical)

State Transition Diagram

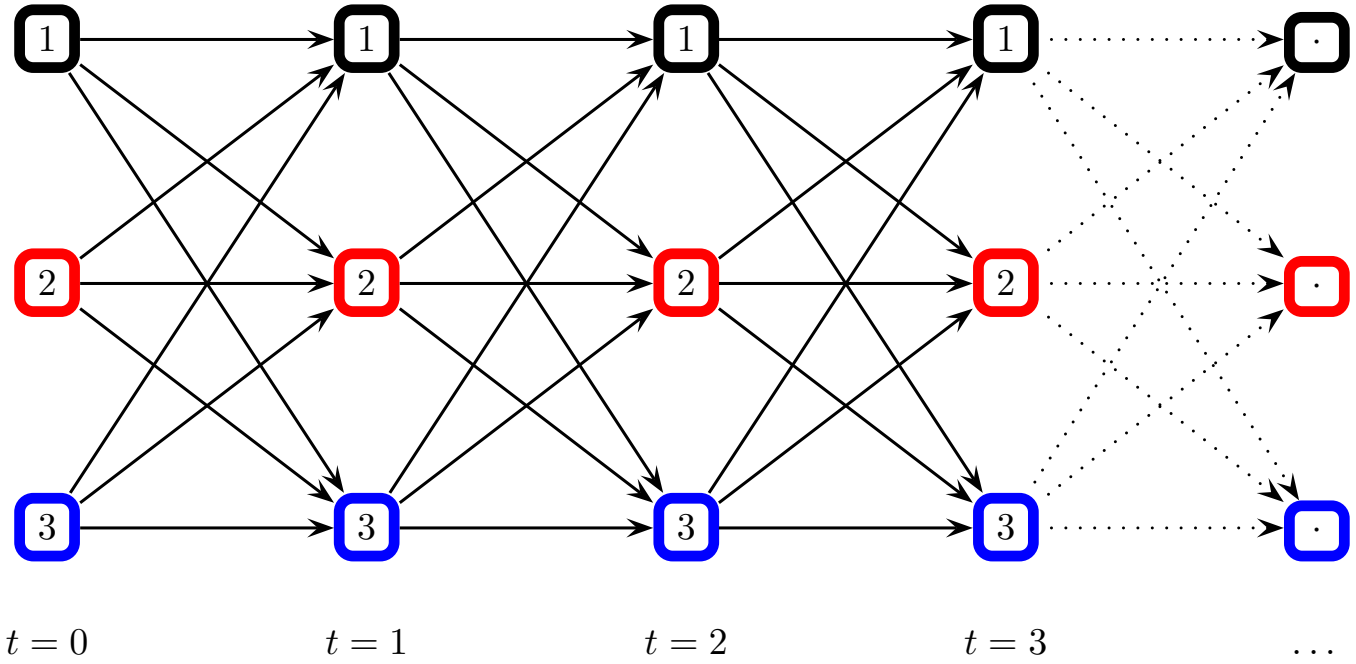


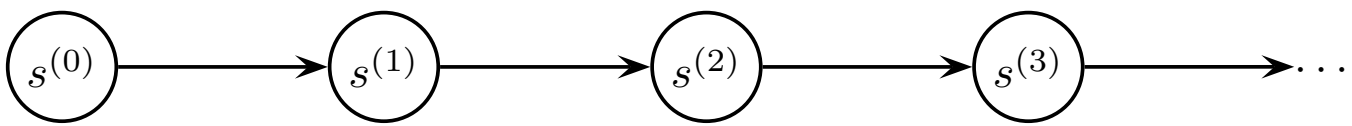
State Transition Diagram (cont.)

$$p(s^{(t)} = i | s^{(t-1)} = j) = \pi_{i,j}$$

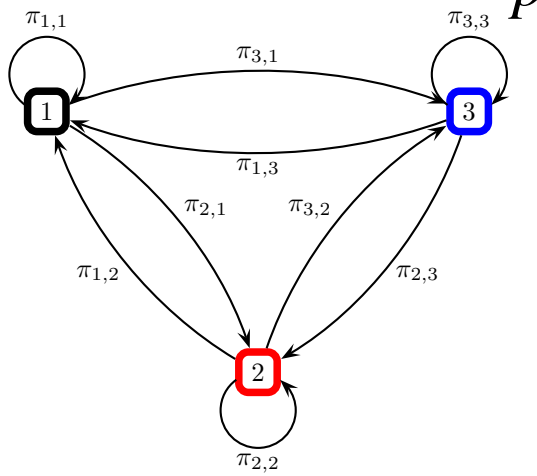
Caution: not a Graphical Model, a (nondeterministic) finite state machine

State Transition Diagram, alternative





Matrix Representation of the State Transition Diagram



$$p\left(s^{(t)}=i\middle|s^{(t-1)}=j\right)=\pi_{i,j}$$

$$p\left(s^{(t)}\middle|s^{(t-1)}\right)=\begin{pmatrix}\pi_{1,1} & \pi_{1,2} & \pi_{1,3}\\ \pi_{2,1} & \pi_{2,2} & \pi_{2,3}\\ \pi_{3,1} & \pi_{3,2} & \pi_{3,3}\end{pmatrix}\equiv\mathbf{T}$$

Computing Marginals

- Recursive

$$
\begin{aligned}
p(s^{(1)}) &= \sum_{s^{(0)}} p(s^{(1)}|s^{(0)})p(s^{(0)}) \qquad \left(\equiv p_i^{(1)} = \sum_j \pi_{i,j} p_j^{(0)} \right) \\
p(s^{(2)}) &= \sum_{s^{(1)}} p(s^{(2)}|s^{(1)})p(s^{(1)}) \\
&\vdots
\end{aligned}
$$

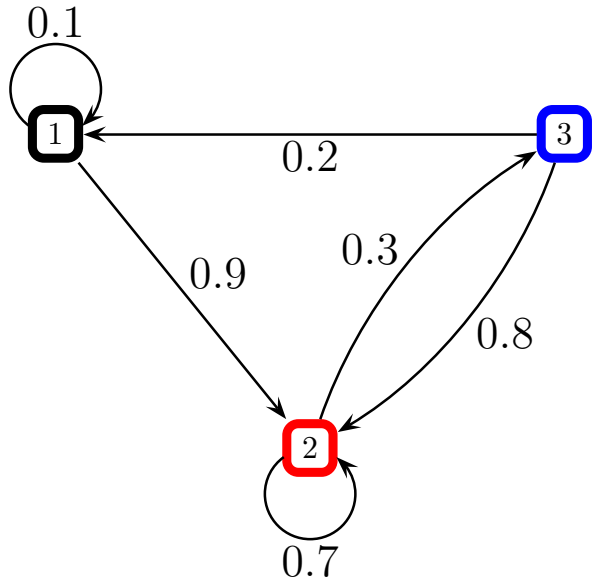
- Using matrix notation, by induction

$$
\begin{aligned}
p^{(1)} &= \mathbf{T}p^{(0)} \\
p^{(2)} &= \mathbf{T}p^{(1)} = \mathbf{T}^2 p^{(0)} \\
&\vdots \\
p^{(t)} &= \mathbf{T}^t p^{(0)}
\end{aligned}
$$

Chapman-Kolmogorov Equations

$$\mathbf{T}^{(n+m)} = \mathbf{T}^n \mathbf{T}^m$$

Numeric Example



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

- Suppose the inital state is 1, we have

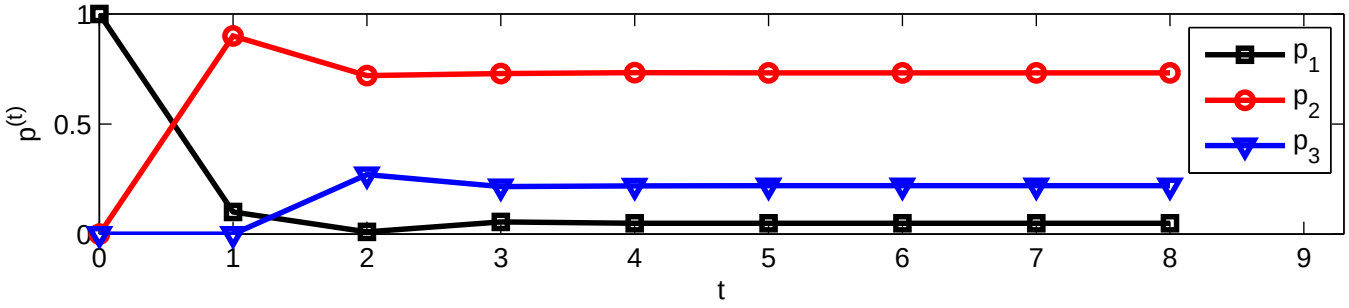
$$p^{(1)} = \mathbf{T}p^{(0)} = \begin{pmatrix} 0.1 & 0 & 0.2 \\ 0.9 & 0.7 & 0.8 \\ 0 & 0.3 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.9 \\ 0 \end{pmatrix}$$

Numeric Example

- Continue

$$p^{(2)} = \mathbf{T} \begin{pmatrix} 0.1 \\ 0.9 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.01 \\ 0.72 \\ 0.27 \end{pmatrix}$$

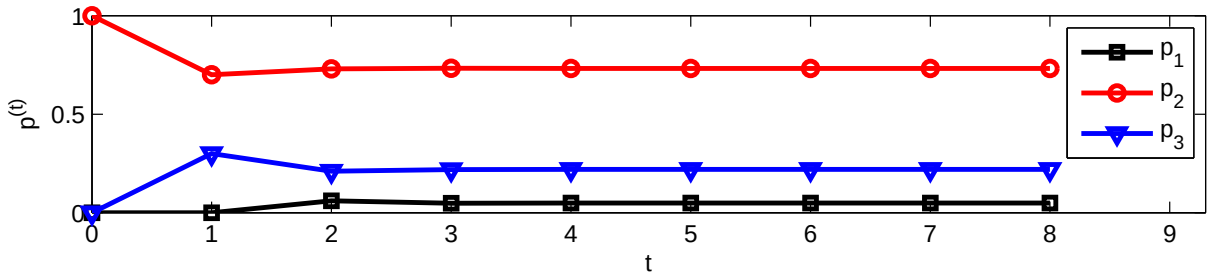
$$p^{(3)} = \mathbf{T} \begin{pmatrix} 0.01 \\ 0.72 \\ 0.27 \end{pmatrix} = \begin{pmatrix} 0.05 \\ 0.73 \\ 0.22 \end{pmatrix}$$



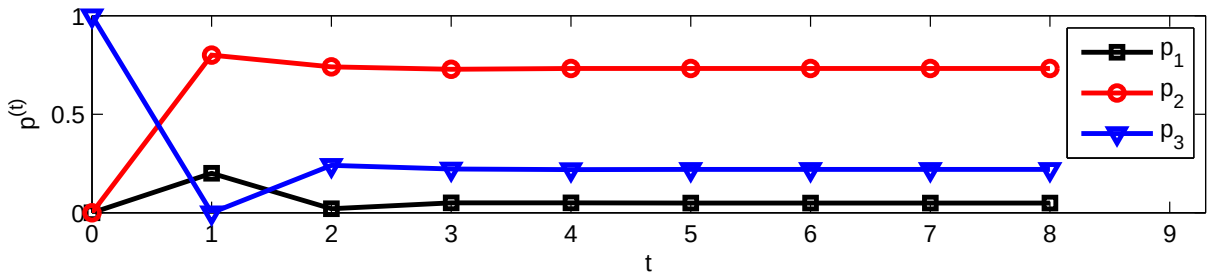
Convergence to a stationary distribution

Starting from other configurations does not alter the picture

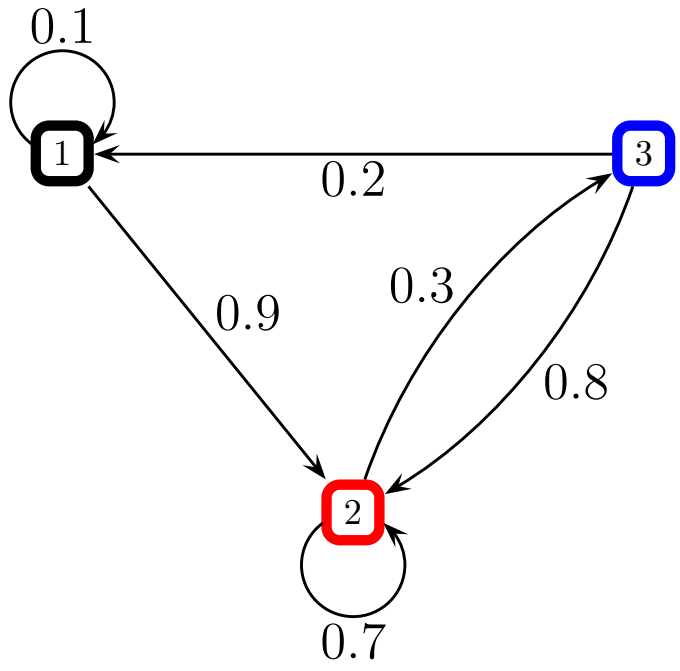
- $p^{(0)} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^\top$



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



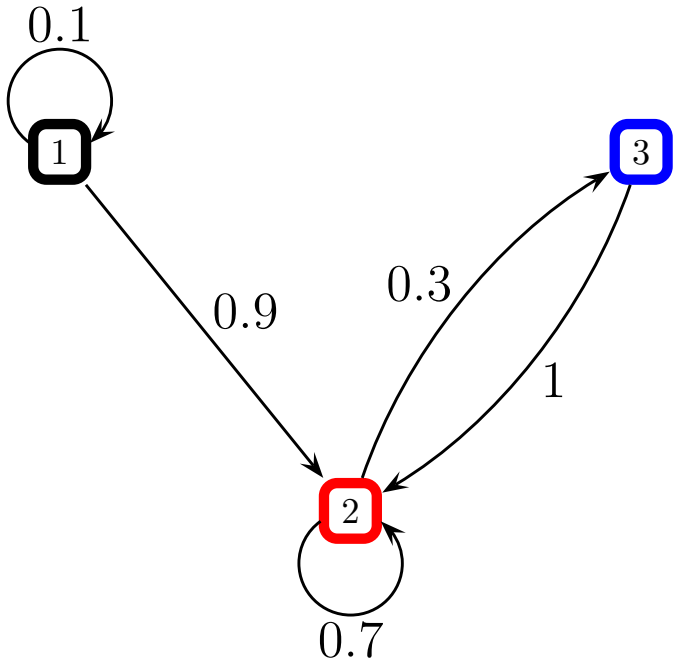
Examples: Irreducable chain



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

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

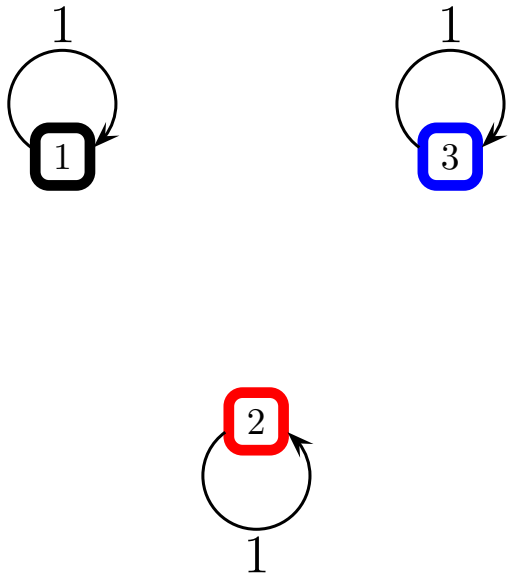
Examples: Transient states



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

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

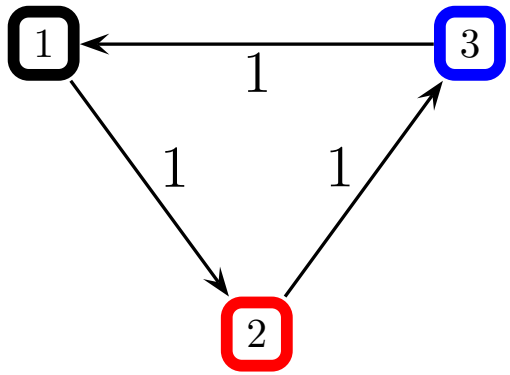
Examples: Reducable chains



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

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

Example: Periodic



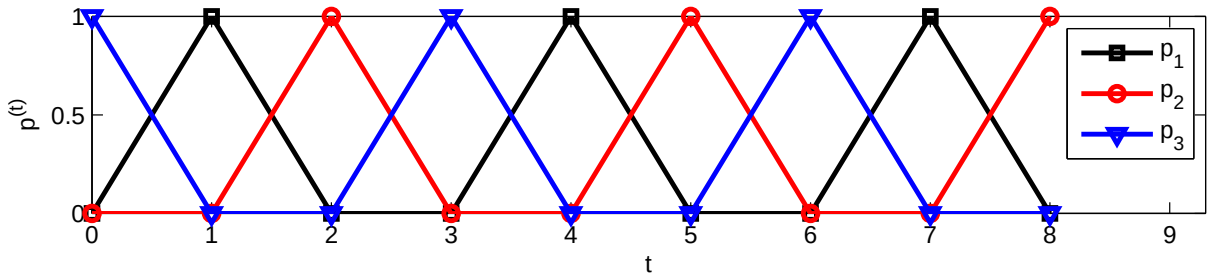
$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

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

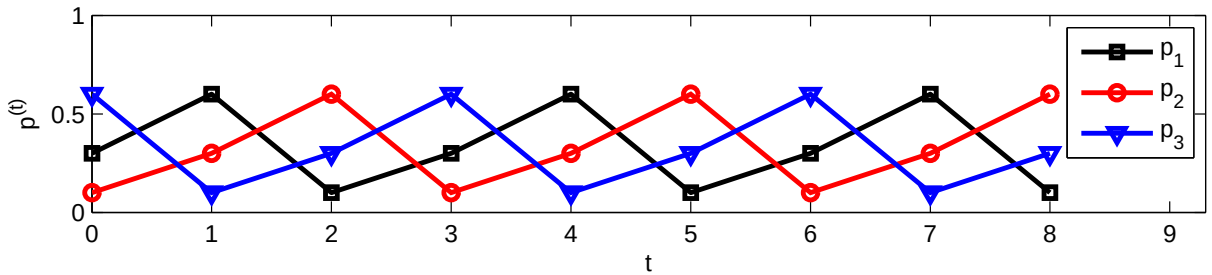
Example: Periodic

There is no stationary distribution

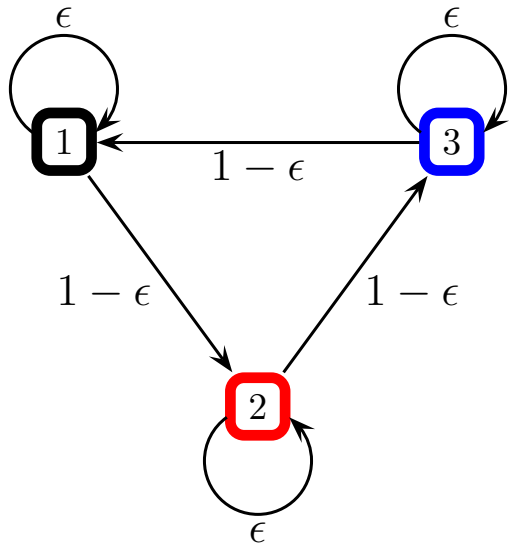
- $p^{(0)} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^\top$



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



Example: Mixture

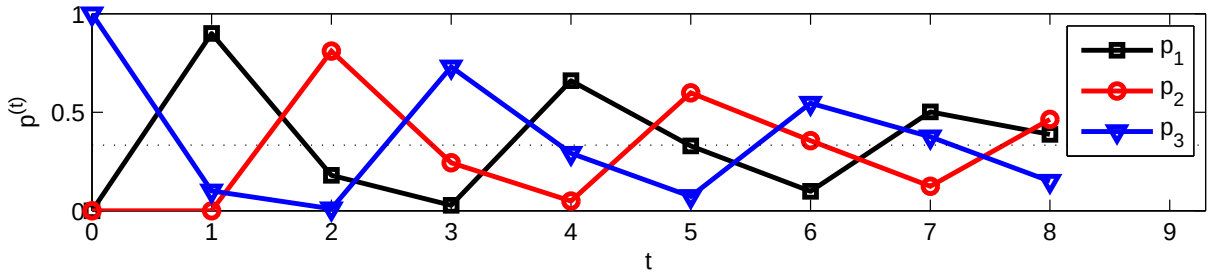


$$(1-\epsilon)\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} + \epsilon \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

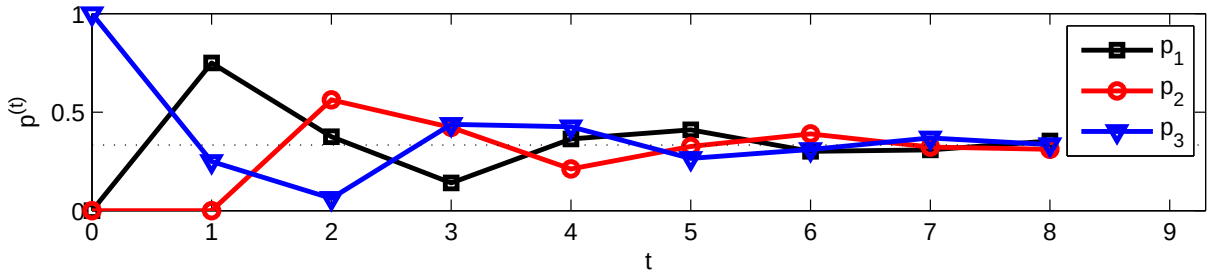
- All states communicate, not periodic
- Is there a unique stationary distribution?

Example: Mixture

- There is a stationary distribution
  $p^{(\infty)} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}^\top$
- $\epsilon = 0.1$



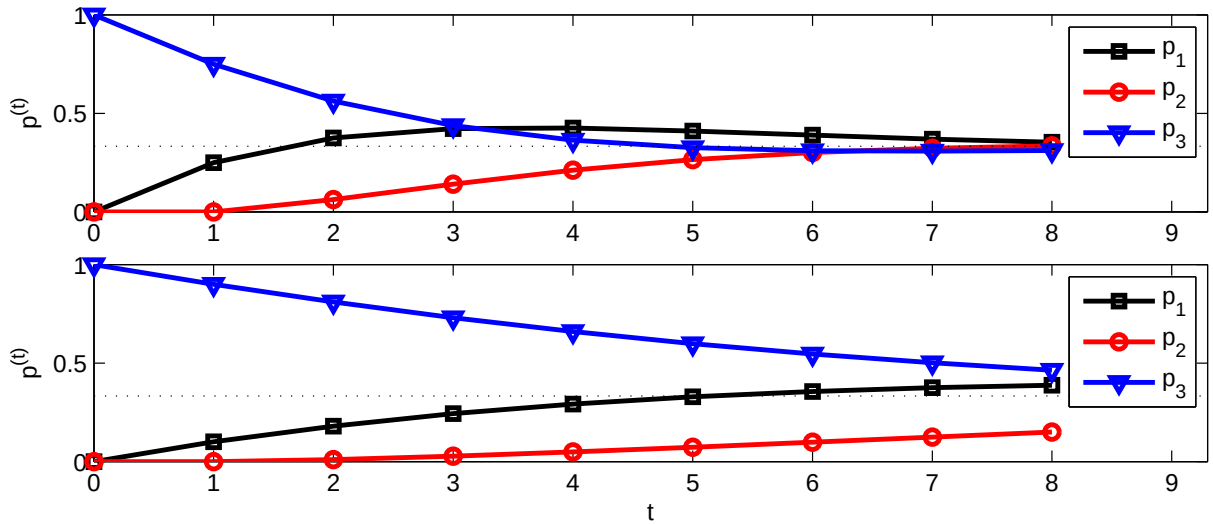
- $\epsilon = 0.25$



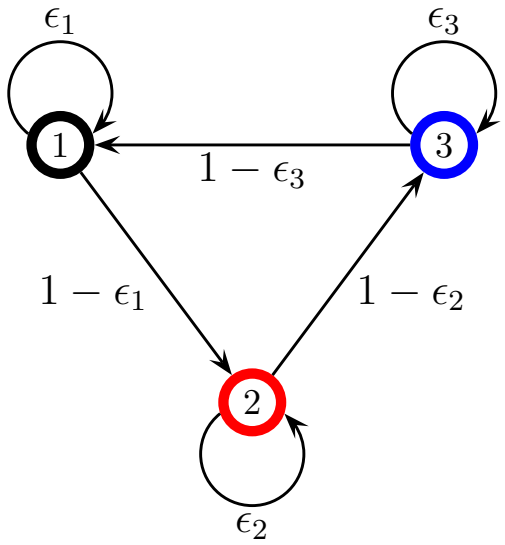
- Convergence rates are different

Example: Mixture

- There is a stationary distribution
  $p^{(\infty)} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}^\top$
- $\epsilon = 0.75$
- $\epsilon =$



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

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

- The stationary distribution is given by the limit

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

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

$$1 = \lambda_1 \geq |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_K|$$

Stationary Distribution

- If and only if $|\lambda_2| < 1$, $\mathbf{T}^t$ goes to

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

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

Matlab illustration

```
>> T = [0.1 0.9 0;0 0.7 0.3;0.2 0.8 0]';
>> [B L] = eig(T)
B =
   0.0637          -0.4714 - 0.3333i  -0.4714 + 0.3333i
   0.9559          -0.2357 + 0.3333i  -0.2357 - 0.3333i
   0.2868           0.7071             0.7071
L =
   1.0000                0                  0
        0          -0.1000 + 0.1414i        0
        0                0            -0.1000 - 0.1414i
```

Matlab illustration

```
>> inv(B)
ans =
   0.7655 + 0.0000i   0.7655 - 0.0000i   0.7655 + 0.0000i
  -0.1552 + 1.2073i  -0.1552 - 0.2927i   0.5519 + 0.7073i
  -0.1552 - 1.2073i  -0.1552 + 0.2927i   0.5519 - 0.7073i

>> B*L*inv(B)
ans =
   0.1000 + 0.0000i   0.0000 - 0.0000i   0.2000 + 0.0000i
   0.9000 + 0.0000i   0.7000 - 0.0000i   0.8000 + 0.0000i
  -0.0000 + 0.0000i   0.3000 + 0.0000i   0.0000

>> B*L^30*inv(B)
   0.0488 + 0.0000i   0.0488 - 0.0000i   0.0488 + 0.0000i
   0.7317 + 0.0000i   0.7317 - 0.0000i   0.7317 + 0.0000i
   0.2195 + 0.0000i   0.2195 - 0.0000i   0.2195 + 0.0000i
```

Rate of Convergence (for finite-state Markov Chains)

- 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 $\mathcal{X} = \{1, 2, \ldots, K\}$
- $\mathbf{T}$ 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

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

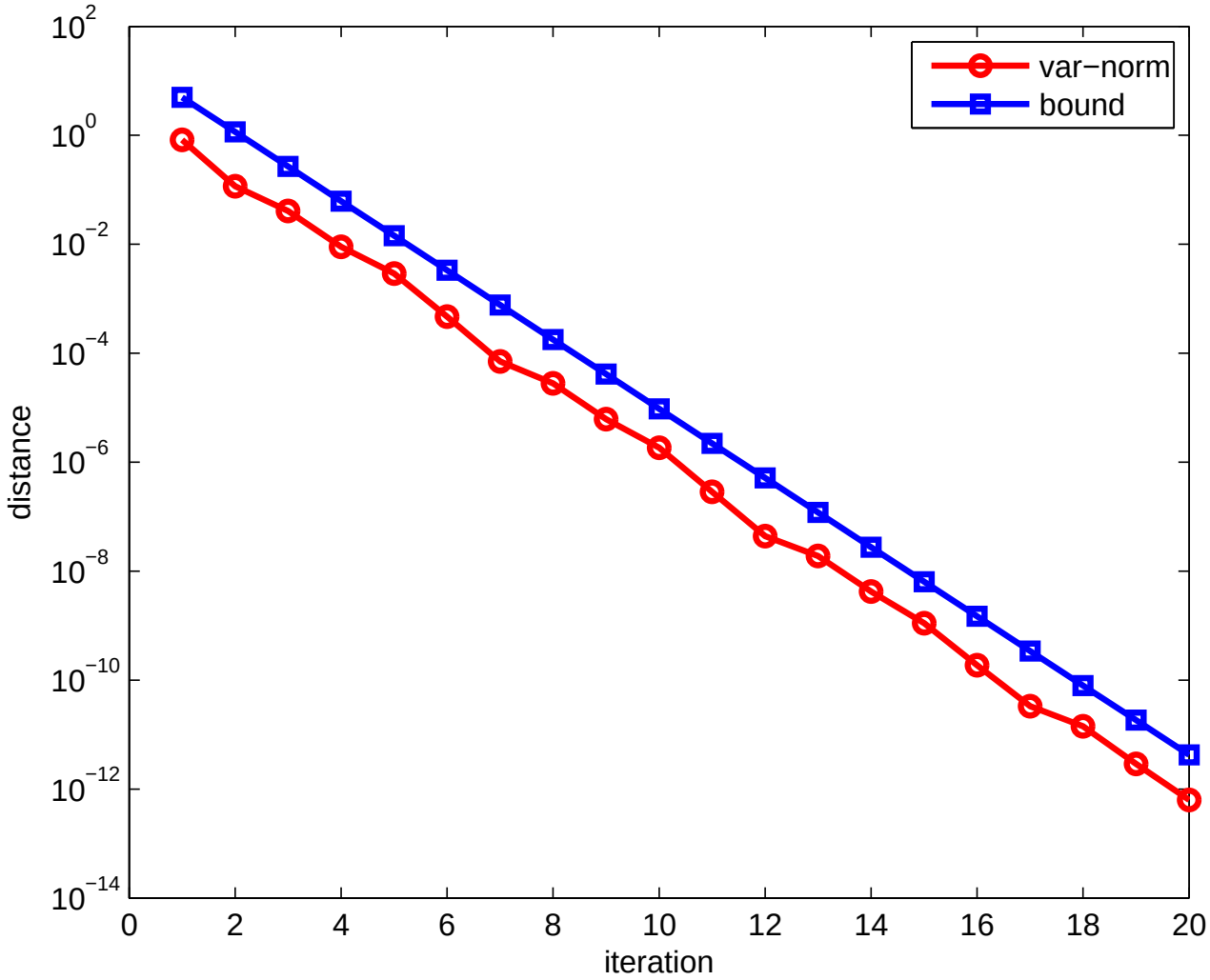
Example: Convergence in variation norm

```
N = 5;
% Generate a random transition matrix
T = rand(N); T = normalize(T,1);

% Compute the stationary distribution
% and second largest eigenvalue
[B L] = eig(T);
[dummy idx] = sort(abs(diag(L)), 'descend');
l2 = L(idx(2),idx(2));
sp = normalize(B(:,idx(1)));

% For the sake of the example
% choose the state with smallest probability
p = zeros(N,1); p(idx(end)) = 1;
```

Example: Convergence in variation norm

```
% iterate
EP = 20; d = zeros(1, EP); bound = zeros(1, EP);
for i=1:EP,
  d(i) = 0.5*sum(abs(p-sp));
  bound(i) = abs(N*l2^(i-1));
  p = T*p;
end;

% plot result
ln = semilogy(d, 'o-r'); set(ln, 'linew', 2);
hold on
ln = semilogy(bound, 's-b'); set(ln, 'linew', 2);
hold off
legend('var-norm', 'bound')
```

Example: Convergence in variation norm



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 $t = 1 \ldots \infty$, 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

Equilibrium condition = Detailed Balance

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

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

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

If the configuration space is discrete, we have

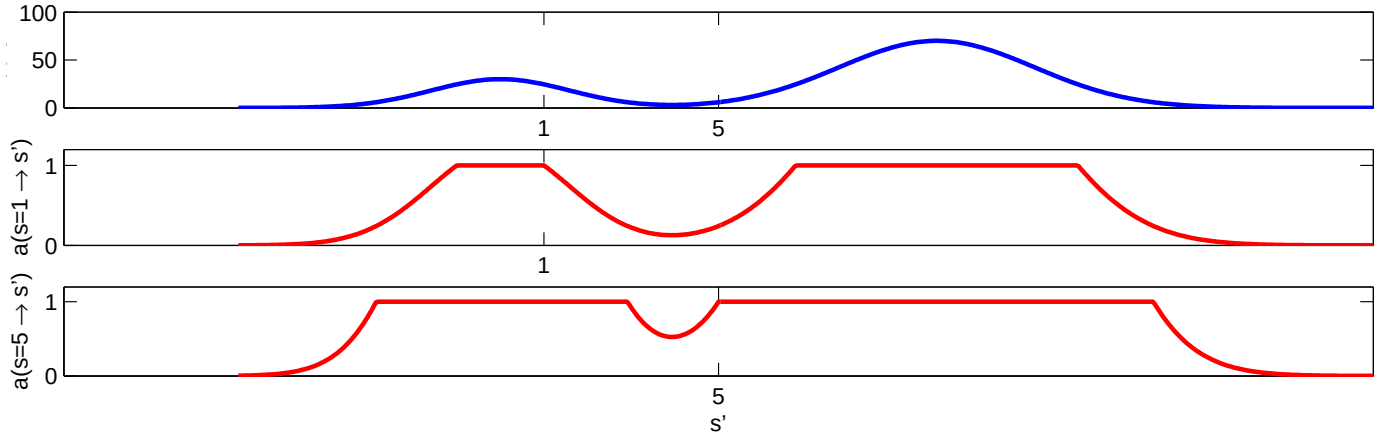
$$\begin{aligned} \pi(\mathbf{s}) &= \sum_{\mathbf{s}'} T(\mathbf{s}|\mathbf{s}')\pi(\mathbf{s}') \\ \pi &= T\pi \end{aligned}$$

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

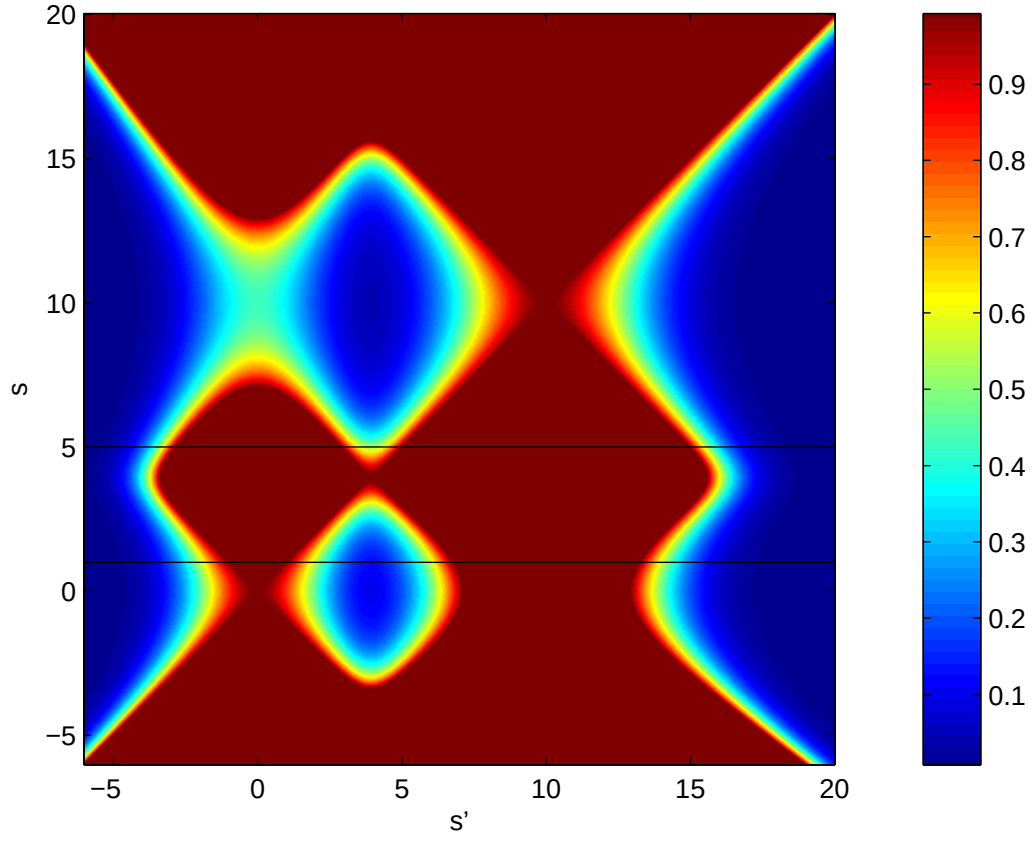
Metropolis-Hastings Kernel

- 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 $a(s \rightarrow s')$



Basic MCMC algorithm: Metropolis-Hastings

1. Initialize: $s^{(0)} \sim r(s)$
2. For $t = 1, 2, \ldots$
   - Propose:

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

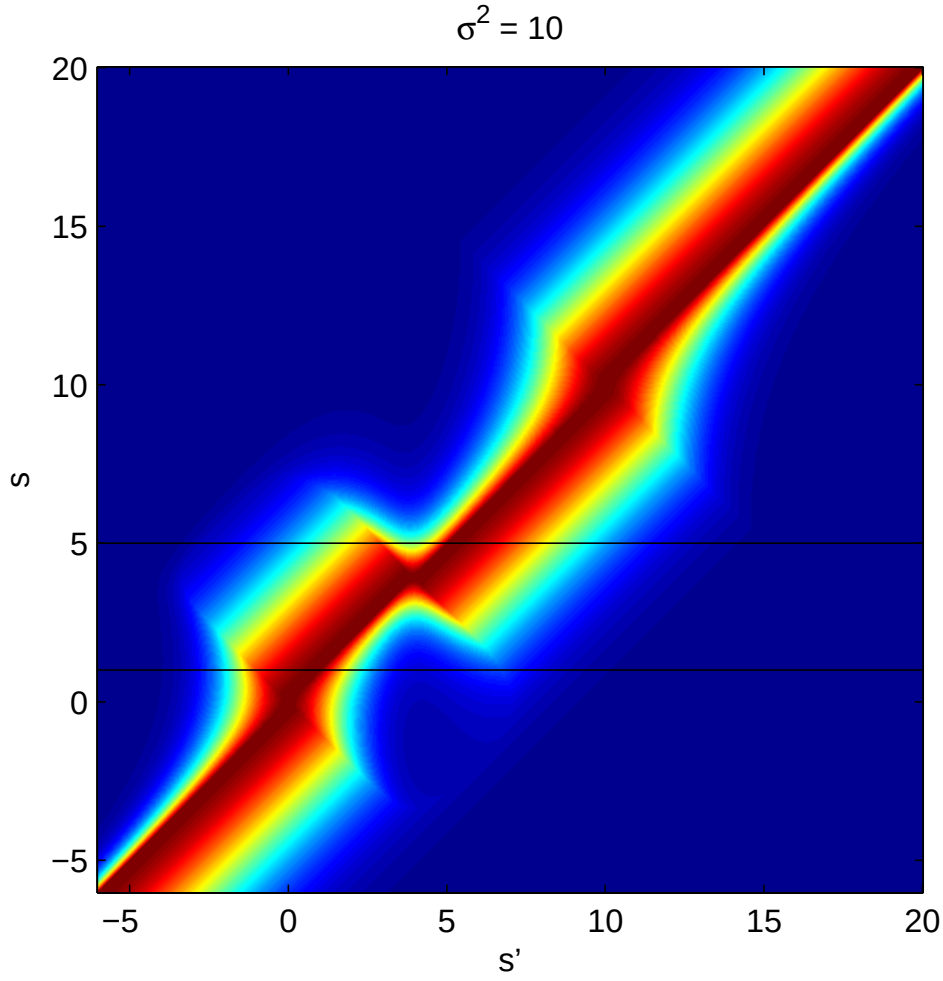
   - Evaluate Proposal: $u \sim \text{Uniform}[0, 1]$

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

Transition Kernel of the Metropolis-Hastings

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

Transition Kernel of the Metropolis-Hastings



Only Accept part for visual convenience

Verification of detailed balance for Metropolis

- Target Density

$$\pi(s) = \frac{1}{Z}\phi(s)$$

- Acceptance Probability

$$a(s \to s') = \min\{1, \frac{\pi(s')}{\pi(s)}\} = \min\{1, \frac{\phi(s')}{\phi(s)}\}$$

- Symmetric Proposal

$$q(s|s') = q(s'|s)$$

Verification of detailed balance for Metropolis

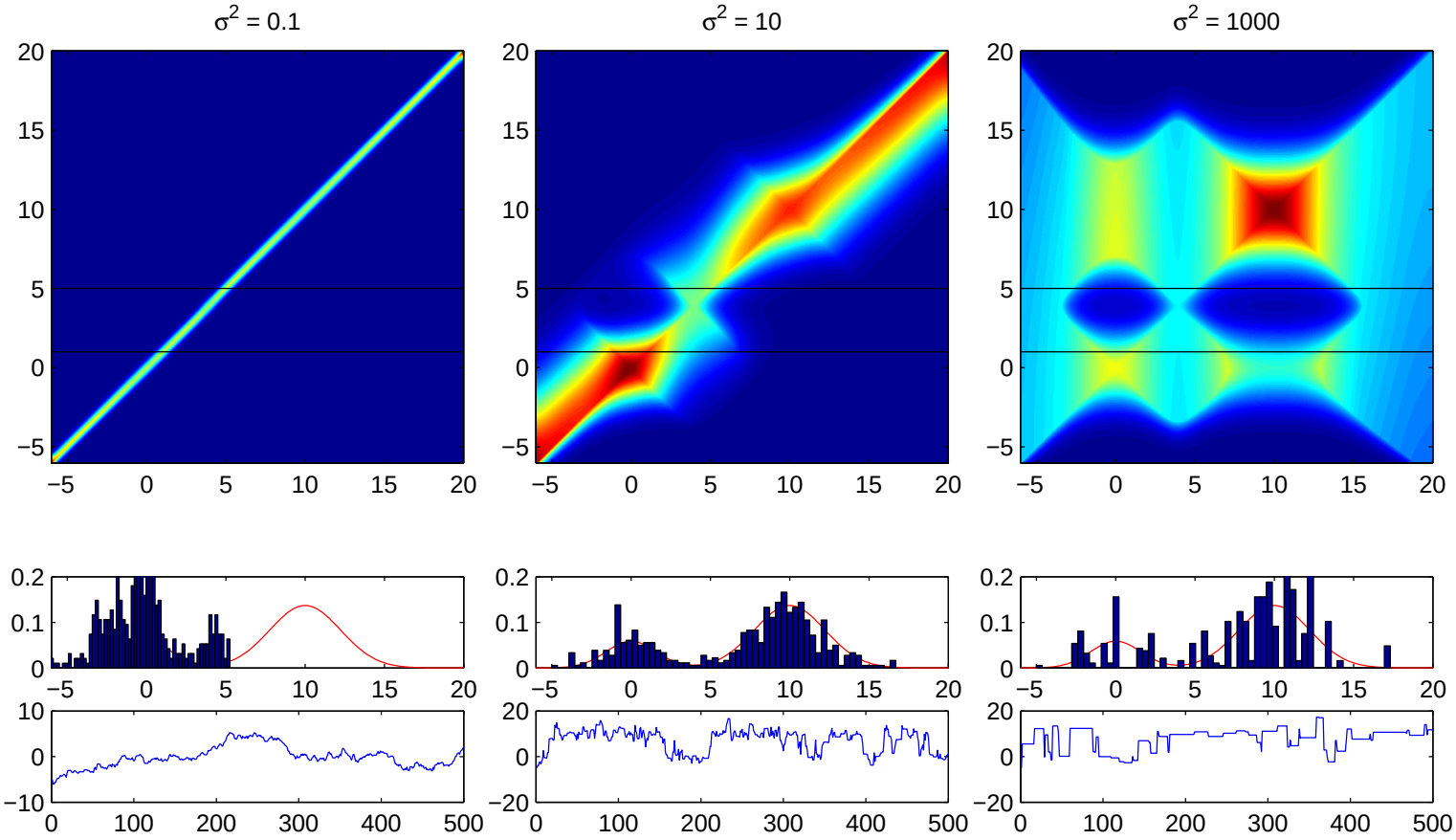
$$
\begin{aligned}
T(s'|s)\pi(s) &= q(s'|s)\min\{1, \frac{\phi(s')}{\phi(s)}\}\pi(s) \quad + \delta(s-s')\rho(s)\pi(s) \\
&= q(s'|s)\min\{\frac{\phi(s)}{Z}, \frac{\phi(s')}{\phi(s)}\frac{\phi(s)}{Z}\} + \delta(s-s')\rho(s)\pi(s) \\
&= q(s'|s)\min\{\frac{\phi(s)}{Z}, \frac{\phi(s')}{Z}\} + \pi(s')\rho(s')\delta(s'-s)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

Verification of detailed balance for Metropolis-Hastings

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

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

Various Kernels with the same stationary distribution



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

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.