Algorithmische Bioinformatik II
WS2004/05
Ralf Zimmer

Part III

Probabilistic Modeling IV

Bayesian Modeling: Algorithms, EM and MC Methods

HMMs

Markov Chain Monte Carlo Methods (MCMC)

- The rejection method is not very useful in high-dimensional problems as it s difficult to find a good bounding g-function
- MCMC is a class of algorithms for simulating random variables from a target distribution $\pi(x)$, given up to a normalizing constant
- This allows to draw from a high dimensional joint posterior distribution without knowing its denominator:

$$p(\theta, y_{mis}|y_{obs}) = \frac{p(y_{obs}, y_{mis}|\theta) f_0(\theta)}{\int\int p(y_{obs}, y'_{mis}|\theta') f_0(\theta') dy'_{mis} d\theta'}$$

- Idea: Design and simulate a Markov chain with equilibrium distribution $\pi(x)$
- There are two common methods for constructing such chains:
  - Metropolis-Hastings (variant: Simulated Annealing)
  - Gibbs Sampler

Definition: A Markov chain is a series of random variables $X^0, X^1, ...$, where the influence of $X^1, ..., X^n$ on the distribution of X^{n+1} is only mediated by X^n

$$P(x^{n+1}|x^n, \{x^t|t \in \mathcal{E}\}) = P(x^{n+1}|x^n)$$

for any subset $\mathcal{E} \subset \{0, ..., n-1\}$

t: successive times

X^t have a common range, the state space of the Markov chain

A Markov chain can completely be specified by

1. the initial probabilities p_0 for X^0 and

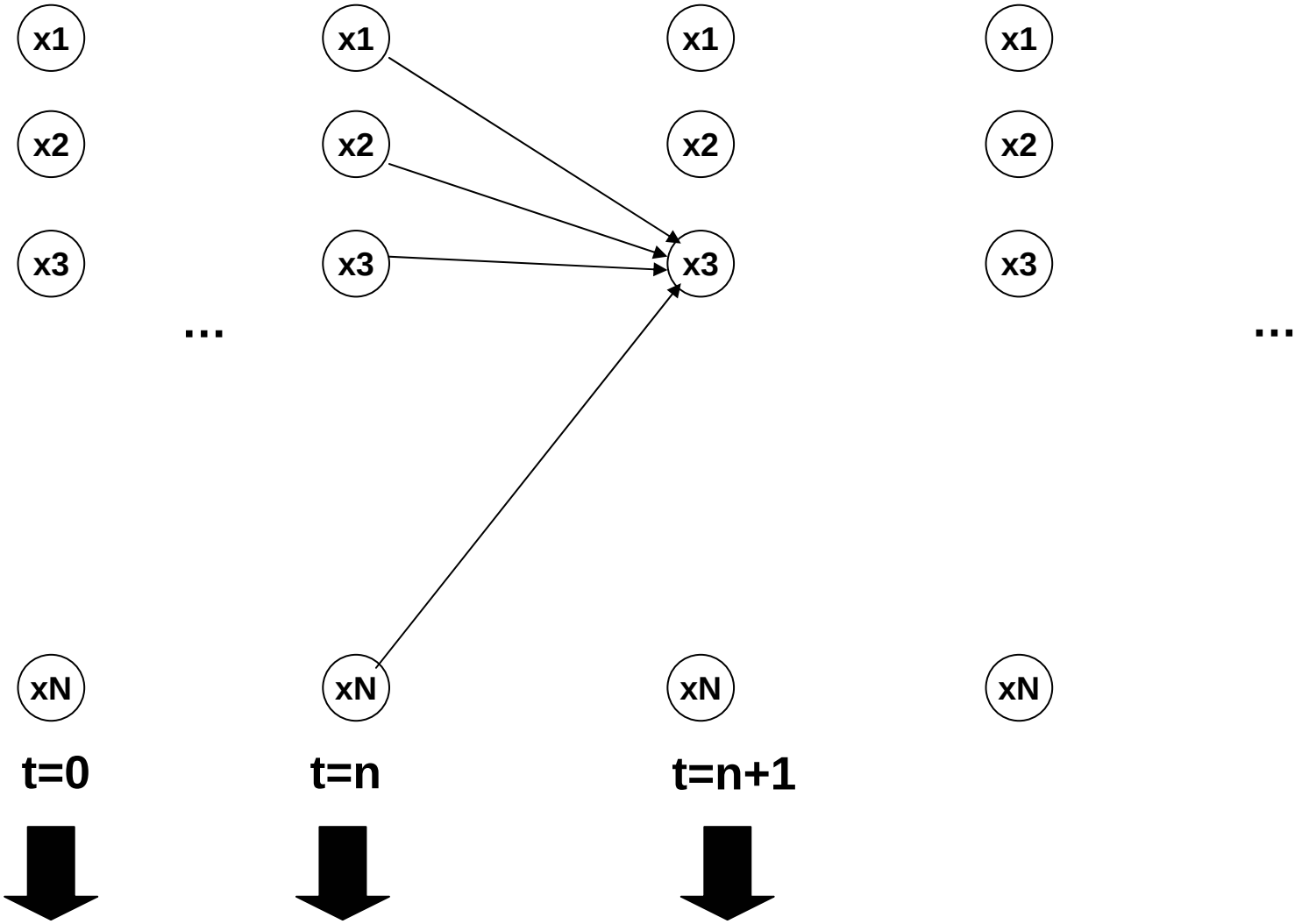
2. the transition probability $T_n(x, x')$ for state x' at time $n+1$ to follow state x at time n

i.e., the probability of a state x at time $n+1$ is:

$$p_{n+1}(x) = \sum_y p_n(y) T_n(y, x)$$

This determines the chain for all times!

Markov chains

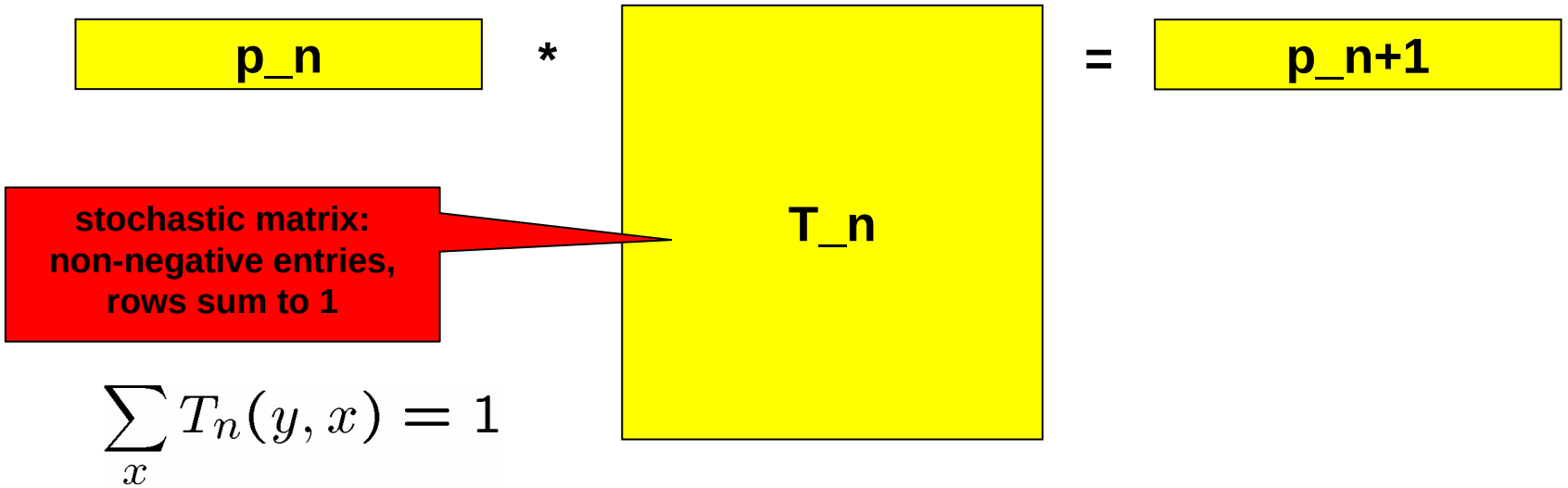


Definition:

A Markov chain is *homogeneous* or *stationary*

$:\Leftrightarrow \forall_n : T_n = T$

$$P_{n+1}(x) = \sum_y p_n(y) T_n(y,x)$$



$$\sum_x T_n(y,x) = 1$$

$$p_n = p_0 T_0 T_1 T_2 ... T_{n-1} \qquad p_n = p_0 T^n$$

Invariant distributions of a Markov Chain

The distribution π is invariant with respect to the Markov chain with transition probabilities T_i iff for all n

$$\forall_x : \pi(x) = \sum_y \pi(y) T_n(y, x)$$

π persists forever once it is reached!

$$\forall_n : \pi = \pi T_n$$

π invariant of a stationary Markov chain with T iff

$$\pi = \pi T$$

Invariant distributions of a Markov Chain

- A Markov chain can have more than one invariant distribution
- Example ?
- If T= I (identity matrix)
- … then any distribution is invariant !
- A finite Markov chain always has at least one invariant distribution
- Goal: Construct a Markov chain for which the target distribution to be sampled is invariant

Detailed balance

- **A time reversible homogeneous Markov chain satisfies the detailed balance condition :⇔ for all x, y**

$$\pi(x)T(x,y) = \pi(y)T(y,x)$$

- **i.e. a transition occurs from a state picked according to probability π, then the probability of that transition from state x to state y is the same as from state y to x**

- **Detailed balance implies that π is an invariant distribution**

$$\sum_x \pi(x)T(x,y) = \sum_x \pi(y)T(y,x) = \pi(y)\left(\sum_x T(y,x)\right) = \pi(y)$$

=1

Detailed balance

- **Example: states = {0,1,2), T(0,1) = T(1,2) = T(2,0) = 1**
- **The uniform distribution (1/3,1/3,1/3) is**
  - **invariant** $\pi = \pi T$

$$(1/3, 1/3, 1/3) * \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = (1/3, 1/3, 1/3)$$

  - **but detailed balance:**

$$\pi(0)T(0,1) = 1/3 * 1 \neq \pi(1)T(1,0) = 1/3 * 0$$

  - **does not hold**

Ergodic Markov Chains

- **Invariant distributions are not enough**
- **For sampling we require that the invariant distribution is reached and is unique independent from the initial probabilities**
- **A Markov chain is *ergodic* iff the probabilities p_n (for n -> ∞) converge to the invariant distribution (then called the *equilibrium distribution*)**
- **Some Markov chains "converge" against a cycle of distributions (and therefore are not ergodic according to the above definition)**

Ergodic Markov Chains: Fundamental Theorem

FUNDAMENTAL THEOREM. *If a homogeneous Markov chain on a finite state space with transition probabilities $T(x, x')$ has π as an invariant distribution and*

$$\nu \quad = \quad \min_{x} \min_{x' : \pi(x') > 0} T(x, x')/\pi(x') \quad > \quad 0 \tag{3.12}$$

then the Markov chain is ergodic, i.e., regardless of the initial probabilities, $p_0(x)$

$$\lim_{n \to \infty} p_n(x) \quad = \quad \pi(x) \tag{3.13}$$

for all x. A bound on the rate of convergence is given by

$$|\pi(x) - p_n(x)| \quad \leq \quad (1 - \nu)^n \tag{3.14}$$

Furthermore, if $a(x)$ is any real-valued function of the state, then the expectation of a with respect to the distribution p_n, written $E_n[a]$, converges to its expectation with respect to π, written $\langle a \rangle$, with

$$|\langle a \rangle - E_n[a]| \quad \leq \quad (1 - \nu)^n \max_{x,\, x'} |a(x) - a(x')| \tag{3.15}$$

Ergodic Markov Chains: Proof of the Fundamental Theorem

$$p_n(x) = [1-(1-\nu)^n]\,\pi(x) + (1-\nu)^n\, r_n(x) \tag{3.16}$$

rn is a valid probability distribution
v <= 1, since we cannot have: $\pi(x') < T(x,x')$ for all x'

3.16 is true for n=0: just set r0(x) = p0(x)

$$p_{\bar{n}+1}(x) = \sum_{\tilde{x}} p_{\bar{n}}(\tilde{x})\,T(\tilde{x},x) \tag{3.17}$$

transition probability of MC

3.16

$$= [1-(1-\nu)^{\bar{n}}] \sum_{\tilde{x}} \pi(\tilde{x})\,T(\tilde{x},x) + (1-\nu)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,T(\tilde{x},x) \tag{3.18}$$

invariant

$$= [1-(1-\nu)^{\bar{n}}]\,\pi(x) + (1-\nu)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,[T(\tilde{x},x) - \nu\pi(x) + \nu\pi(x)] \tag{3.19}$$

$$= [1-(1-\nu)^{\bar{n}}]\,\pi(x) + (1-\nu)^{\bar{n}}\,\nu\,\pi(x) + (1-\nu)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,[T(\tilde{x},x) - \nu\pi(x)] \tag{3.20}$$

$$= [1-(1-\nu)^{\bar{n}+1}]\,\pi(x) + (1-\nu)^{\bar{n}+1} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,\frac{T(\tilde{x},x) - \nu\pi(x)}{1-\nu} \tag{3.21}$$

$$= [1-(1-\nu)^{\bar{n}+1}]\,\pi(x) + (1-\nu)^{\bar{n}+1}\, r_{\bar{n}+1}(x) \tag{3.22}$$

where $r_{\bar{n}+1}(x) = \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,[T(\tilde{x},x) - \nu\pi(x)]/(1-\nu)$

Ergodic Markov Chains: Proof of the Fundamental Theorem

$$p_n(x) = [1-(1-\nu)^n]\,\pi(x) + (1-\nu)^n\, r_n(x) \qquad (3.16)$$

rn is a valid probability distribution

v <= 1, since we cannot have: $\pi(x') < T(x,x')$ for all x'

3.16 is true for n=0: just set r0(x) = p0(x)

$$p_{\bar{n}+1}(x) = \sum_{\tilde{x}} p_{\bar{n}}(\tilde{x})\,T(\tilde{x},x) \qquad (3.17)$$

transition probability of MC

3.16

$$= [1-(1-\nu)^{\bar{n}}] \sum_{\tilde{x}} \pi(\tilde{x})\,T(\tilde{x},x) + (1-\nu)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,T(\tilde{x},x) \qquad (3.18)$$

invariant

$$= [1-(1-\nu)^{\bar{n}}]\,\pi(x) + (1-\nu)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,[T(\tilde{x},x) - \nu\pi(x) + \nu\pi(x)] \qquad (3.19)$$

$$= [1-(1-\nu)^{\bar{n}}]\,\pi(x) + (1-\nu)^{\bar{n}}\,\nu\,\pi(x)$$

$$+ (1-\nu)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,[T(\tilde{x},x) - \nu \qquad 0)$$

$$= [1-(1-\nu)^{\bar{n}+1}]\,\pi(x) + (1-\nu) \qquad 1)$$

$$= [1-(1-\nu)^{\bar{n}+1}]\,\pi(x) + (1-\nu)^{\bar{n}+1}\, r_{\bar{n}+1}(x) \qquad (3.22)$$

$$(1-(1-v)^n) + (1-v)^n v$$
$$= 1-(1-v)^n + (1-v)^n v$$
$$= 1-((1-v)^n - (1-v)^n v)$$
$$= 1-((1-v)^n(1-v))$$
$$= 1-(1-v)^{n+1}$$

where $r_{\bar{n}+1}(x) = \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x})\,[T(\tilde{x},x) - \nu\pi(x)]/(1-\nu)$

Ergodic Markov Chains: Proof of the Fundamental Theorem

r_{n+1} is a probability distribution, i.e.

1. $r_{n+1}(x) \geq 0$ and

2. $\sum_x r_{n+1}(x) = 1$

Ergodic Markov Chains: Proof of the Fundamental Theorem

r_{n+1} is a probability distribution, i.e.

1. $r_{n+1}(x) \geq 0$

   Why? $r_{n+1}(x) = \sum_y r_n(y) \frac{(T(y,x) - \nu\pi(x))}{(1-\nu)}$

   $(1-\nu) \geq 0$, as $\nu \leq 1$ (see above)

   $T(y,x) - \nu\pi(x) \geq 0 \Leftrightarrow \frac{T(y,x)}{\pi(x)} \geq \nu$

   (as $\nu = min_y min_{x,\pi(x)>0} \frac{T(y,x)}{\pi(x)}$!)

2. $\sum_x r_{n+1}(x) = 1$

Ergodic Markov Chains: Proof of the Fundamental Theorem

r_{n+1} is a probability distribution, i.e.

1. $r_{n+1}(x) \geq 0$

2. $\sum_x r_{n+1}(x) = 1$

Why?

$$
\begin{aligned}
\sum_x r_{n+1}(x) &= \sum_x \sum_y r_n(y) \frac{T(y,x) - \nu\pi(x)}{1-\nu} \\
&= \frac{1}{1-\nu} \sum_y r_n(y) \sum_x (T(y,x) - \nu\pi(x)) \\
&= \frac{1}{1-\nu} \sum_y r_n(y) \left(\sum_x (T(y,x)) - \sum_x \nu\pi(x) \right) \\
&= \frac{1}{1-\nu} \sum_y r_n(y) (1 - \nu) \\
&= \sum_y r_n(y) \\
&= 1
\end{aligned}
$$

Ergodic Markov Chains: Proof of the Fundamental Theorem

Using (3.16), we can now show that (3.14) holds:

$$
\begin{aligned}
|\pi(x) - p_n(x)| &= |\,\pi(x) - [1-(1-\nu)^n]\,\pi(x) - (1-\nu)^n\, r_n(x)\,| && (3.23)\\
&= |\,(1-\nu)^n\,\pi(x) - (1-\nu)^n\, r_n(x)\,| && (3.24)\\
&= (1-\nu)^n\,|\pi(x) - r_n(x)| && (3.25)\\
&\leq (1-\nu)^n && (3.26)
\end{aligned}
$$

Ergodic Markov Chains: Proof of the Fundamental Theorem

$$p_n(x) \;=\; [1-(1-\nu)^n]\,\pi(x) \;+\; (1-\nu)^n\, r_n(x)$$

We can show (3.15) similarly:

$$\begin{aligned}
|\langle a\rangle - E_n[a]| &= \Big|\textstyle\sum_{\tilde{x}} a(\tilde{x})\,\pi(\tilde{x}) \;-\; \sum_{\tilde{x}} a(\tilde{x})\,p_n(\tilde{x})\Big| && (3.27)\\
&= \Big|\textstyle\sum_{\tilde{x}} a(\tilde{x})\,[(1-\nu)^n\pi(\tilde{x}) \;-\; (1-\nu)^n r_n(\tilde{x})]\Big| && (3.28)\\
&= (1-\nu)^n\,\Big|\textstyle\sum_{\tilde{x}} a(\tilde{x})\,\pi(\tilde{x}) \;-\; \sum_{\tilde{x}} a(\tilde{x})\,r_n(\tilde{x})\Big| && (3.29)\\
&\le (1-\nu)^n \max_{x,\,x'} |a(x)-a(x')| && (3.30)
\end{aligned}$$

This completes the proof.

Fundamental Theorem

- **Theorem applies to homogeneous Markov chains**
- **Cyclic Markov chains (transition matrices repeat after some period)**

$$T_n = T_{n+d}$$

=> homogeneous Markov chain with transition matrix

$$T = T_0 T_1 \dots T_{d-1}$$

- **If for a homogeneous chain v>0 does not hold for its T, it could probably hold for some T^k, thus, the theorem holds with rate bounds with the exponent n replaced by n/k**

- **Theorem: at large times the distribution will be close to the invariant distribution**
- **The state of an ergodic Markov chain at time m will be nearly independent of the state at time n when m >> n. Starting at any state at n will lead to the invariant distribution at m.**
- **This motivates to approximate the expectation via the mean of a series of values taken from an ergodic Markov chain with the appropriate invariant distribution.**

Markov Chains

- **Computational effort for a good Monte Carlo estimate from the states of a Markov chain depend on**
  - **time to simulate each transition**
  - **time to converge to the equilibrium distribution (giving the states to be discarded before convergence)**
  - **number of transitions needed to move between two independent states from the equilibrium distribution**

- **Bounds in the theorem often not very practical**
  - **Many theoretical analyses have been tried (e.g. via Eigenvalues)**
  - **or empirical testing is required**

Analysis of Markov Chains via Eigenvalues

- **Definition (left eigenvector of a matrix M)**

  A left eigenvector of a matrix M is a vector $v \neq 0$ s.t.

  $$vM = \lambda v$$

  for some possibly complex number λ, the eigenvalue associated with eigenvector v

- **Thus, obviously, the invariant distribution π of a homogeneous Markov chain is an eigenvector of its transition matrix T for eigenvalue 1**

$$\pi T = \pi$$

Analysis of Markov Chains via Eigenvalues

- **For most stochastic (transition) matrices T one can find a set of linearly independent eigenvectors.**
  The eigenvectors are
  1. **associated with eigenvalue 1 and with real positive values summing to 1**
  2. **other eigenvectors associated with eigenvalue 1**
  3. **eigenvectors with elements summing to 0 associated with eigenvalues of magnitude < 1**
- **class (1) contain the invariant distributions**
- **class (2) eigenvectors indicate that the chain is periodic**
- **Thus, ergodic Markov chains have one eigenvector of class (1) and possibly several of class (3)** $1 > |\lambda_2| \geq |\lambda_3| \geq \ldots$ **with eigenvectors** $\pi, v_2, v_3, \ldots$

Analysis of Markov Chains via Eigenvalues

- **Represent the initial probability** $p_0 = \pi + a_2 v_2 + a_3 v_3 + \ldots$
- **then the distribution after n steps of the chain is**

$$p_n = p_0 T^n = \pi T^n + a_2 v_2 T^n + a_3 v_3 T^n + \ldots$$
$$= \pi + \lambda_2^n a_2 v_2 + \lambda_3^n a_3 v_3 + \ldots$$

- **Thus, with increasing n, p_n will converge to π with the asymptotic rate of convergence given by the magnitude of the second-largest eigenvalue λ_2**
- **As the eigenvalues of a reversible Markov chain are real, we have a bound on the time for convergence (independent of the initial distribution)**

$$max_x \frac{p_n(x) - \pi(x)}{\pi(x)} \leq \frac{\lambda_2}{min_x \pi(x)}$$

- **If T(x,x) >= ½ for all x, all eigenvalues are non-negative and λ_2 is the largest eigenvalue less than one**