ε.δε.μ²
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Stochastic Methods Rooted in Statistical Mechanics
Markov Chains: Definitions of States & Transitions
Stationarity & Ergodic Probabilities
Global & Detailed Balance Equations

Prof. Vasilis Maglaris
maglaris@netmode.ntua.gr
www.netmode.ntua.gr
Room 002, New ECE Building

Tuesday March 4, 2025

STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Statistical Mechanics & Machine Learning

- Long-term statistics of sample element $\mathbf{x}(n)$ are related to macroscopic concepts of **mechanical physics** reaching **dynamic equilibrium** under a given temperature. There is an analogy of terms such as **entropy** (disorder) with states and control parameters of **Machine Learning** systems

- **Statistical Mechanics** offers **inference** models based on measured (and assumed) statistics of input elements $\mathbf{x}(n)$ that enable self-organization algorithms (e.g. via **Unsupervised Learning** for data compression, classification into clusters, forming context-aware maps, correction of corrupted data) and **generation** of elements distributed according to probabilistic assumptions of the sample space (**Statistical Sampling** - **GenAI**)

- The features of sample elements are encoded into **random variables** that constitute the m coordinates of input **sample vectors** selected from the environment (sample space). Analogies with statistical mechanics may lead to better understanding the impact of the m **features** of a large number of **vectors** used to train and validate a machine learning system

- Ground-braking application: **Boltzmann Machine** (*Hinton – Sejnowski*, 1983) used for processing and generation of images inspired by statistical mechanics models (named after physicist and philosopher *Ludwig Boltzmann*, 1844-1906)

STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Statistical Mechanics: Gibbs Distribution, Partition Function, Entropy

Thermal Equilibrium of Physical System with many Degrees of Freedom

- A physical system with many degrees of freedom reaches dynamic equilibrium at absolute temperature T in state i of energy E_i with probability p_i (frequency of occurrence of i) given by the ***Gibbs Distribution*** (1902) or ***Boltzmann*** (1868):

$$p_i \propto \exp\left(-\frac{E_i}{T}\right), \quad \frac{p_i}{p_j} = \exp\left(-\frac{E_i - E_j}{T}\right), \qquad p_i \geq 0, \ \sum_i p_i = 1$$

- States i of low energy E_i tend to occur more often in equilibrium with probabilities:

$$p_i = \frac{1}{Z}\exp\left(-\frac{E_i}{T}\right), \ Z = \sum_i \exp\left(-\frac{E_i}{T}\right)$$

Normalization Constant (*Zustadsumme*) Z: *Partition Function*

- The ***energy*** of state i is $E_i = -T\log(Zp_i)$ with an average $< E > \ = \sum_i p_i E_i$

- The total ***Free Energy*** F (***Helmholtz Free Energy***) is:

$$F = -T\log Z \Rightarrow < E > -F = -T\sum_i p_i \log p_i$$

- The ***entropy*** of the system is:

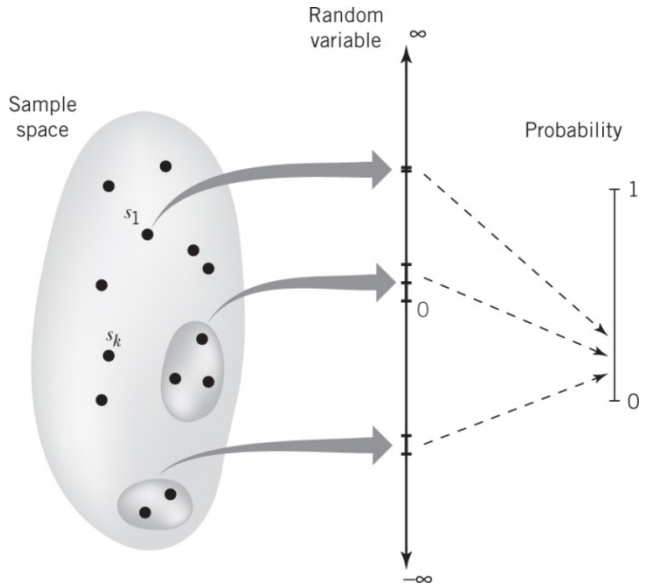
$$H \triangleq -\sum_i p_i \log p_i \Rightarrow \ < E > -F = TH \text{ ή } F = < E > -TH$$

Principle of Minimal Free Energy (***Landau*** & ***Lifshitz***, 1980)

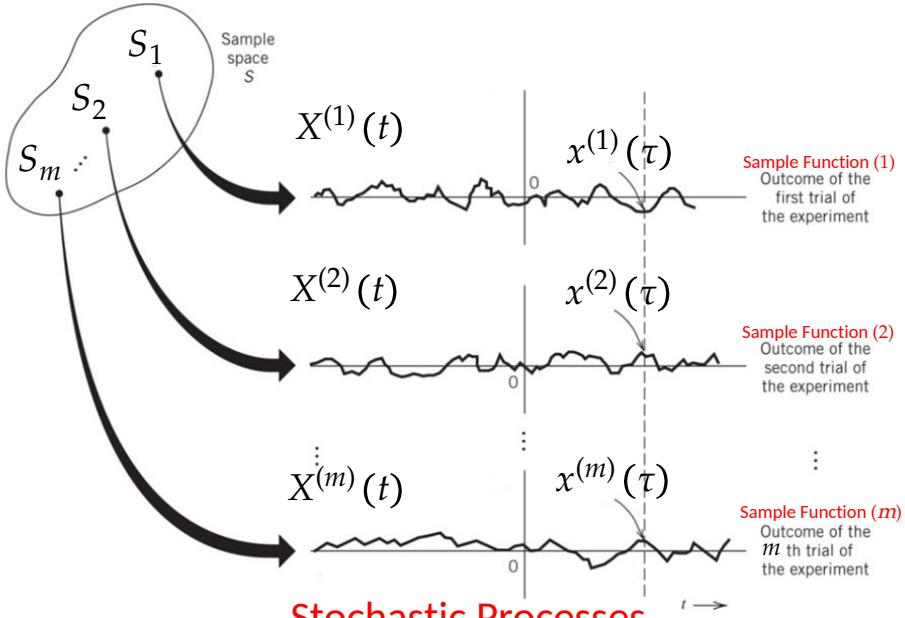
*In thermal equilibrium, H tends to its maximum, F to its minimum and states follow the **Gibbs Distribution***

STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Stochastic Processes – Time Series, Markov Property



Random Variables



Stochastic Processes

- ***Stochastic Process of State*** $X(t)$ with transitions from time τ to time t of the *same sample element* (*time series outcome, trajectory*) with probability $P\{[X(t)=a]\,|[X(\tau)=b]\}$
- ***Stochastic Process in Discrete Transition Times of State*** $X_n \triangleq X(n\times\Delta t)$ with transitions from $\tau=(k\times\Delta t)$ to $t=(n\times\Delta t)$ of the *same sample element* (*discrete time series outcome*, *trajectory*) with probability $P(X_n=a|X_k=b\}$

Markov Property in Discrete Time Transition Processes

A discrete time Stochastic Process exhibits the Markov property if transition probabilities from state $X_n \rightarrow X_{n+1}$ do not depend on its past state evolution $\{X_1,X_2,\ldots,X_{n-1}\}$

$$P\left(X_{n+1}=x_{n+1}|X_n=x_n,\ldots,X_1=x_1\right\}=P\left(X_{n+1}=x_{n+1}|X_n=x_n\ \right\}$$

STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Markov Processes of Discrete State - Markov Chains

Consider a Markov Process of ***Discrete States*** $X_n = i$, referred to as ***Markov Chain***, with ***Discrete Transition Times*** for $(X_n = i) \rightarrow (X_{n+1} = j)$ at instant n, independent of the past

Assume that transition probabilities in one step are constant and independent of instant n:

$$p_{ij} = P\left(X_{n+1} = j \middle| X_n = i\right) \geq 0, \quad \sum_j p_{ij} = 1 \ \ \forall i$$

If $i \leq K$ the transition probabilities are elements of the transition matrix $\mathbf{P}$ with row elements adding to one: $\sum_j p_{ij} = 1$ (***stochastic matrix***):

$$\mathbf{P} = \begin{bmatrix} p_{11} & \cdots & p_{1K} \\ \vdots & \ddots & \vdots \\ p_{K1} & \cdots & p_{KK} \end{bmatrix}$$

The transition probabilities in m steps are $p_{ij}^{(m)} = P\left(X_{n+m} = j \middle| X_n = i\right)$, $m = 1,2,\ldots$ are given by the ***Chapman-Kolmogorov*** identity:

$$p_{ij}^{(m+1)} = \sum_k p_{ik}^{(m)} p_{kj} \quad \text{και} \quad p_{ij}^{(m+n)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)}, \quad m,n = 1,2,\ldots$$

STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

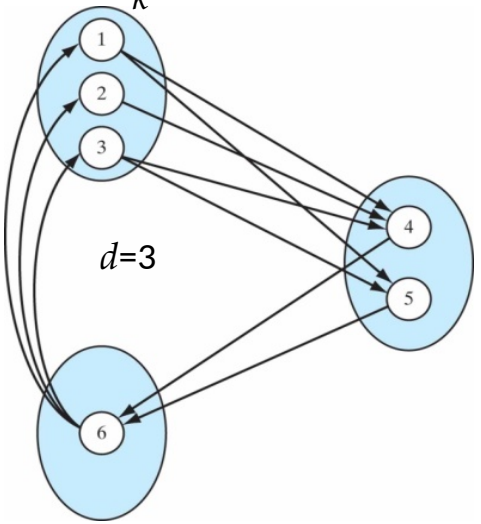
Definitions of Markov Chain States

- **Recurrent States**: The process visits these states infinitely often in an infinite time-horizon

- **Transient States**: After a finite number of transition steps, the process stops visiting these states

- **Periodicity**: If all recurrent states are grouped in d disjoint subsets $S_1, S_2, \ldots, S_d$ with transitions allowed from a given subset to a next, then visits into subsets S_k occur periodically, every d transitions:

$$\text{If } i \in S_k,\ p_{ij} > 0 \ \Rightarrow \begin{cases} j \in S_{k+1} & \text{for } k = 1,\ldots,d-1 \\ j \in S_1 & \text{for } k = d \end{cases}$$

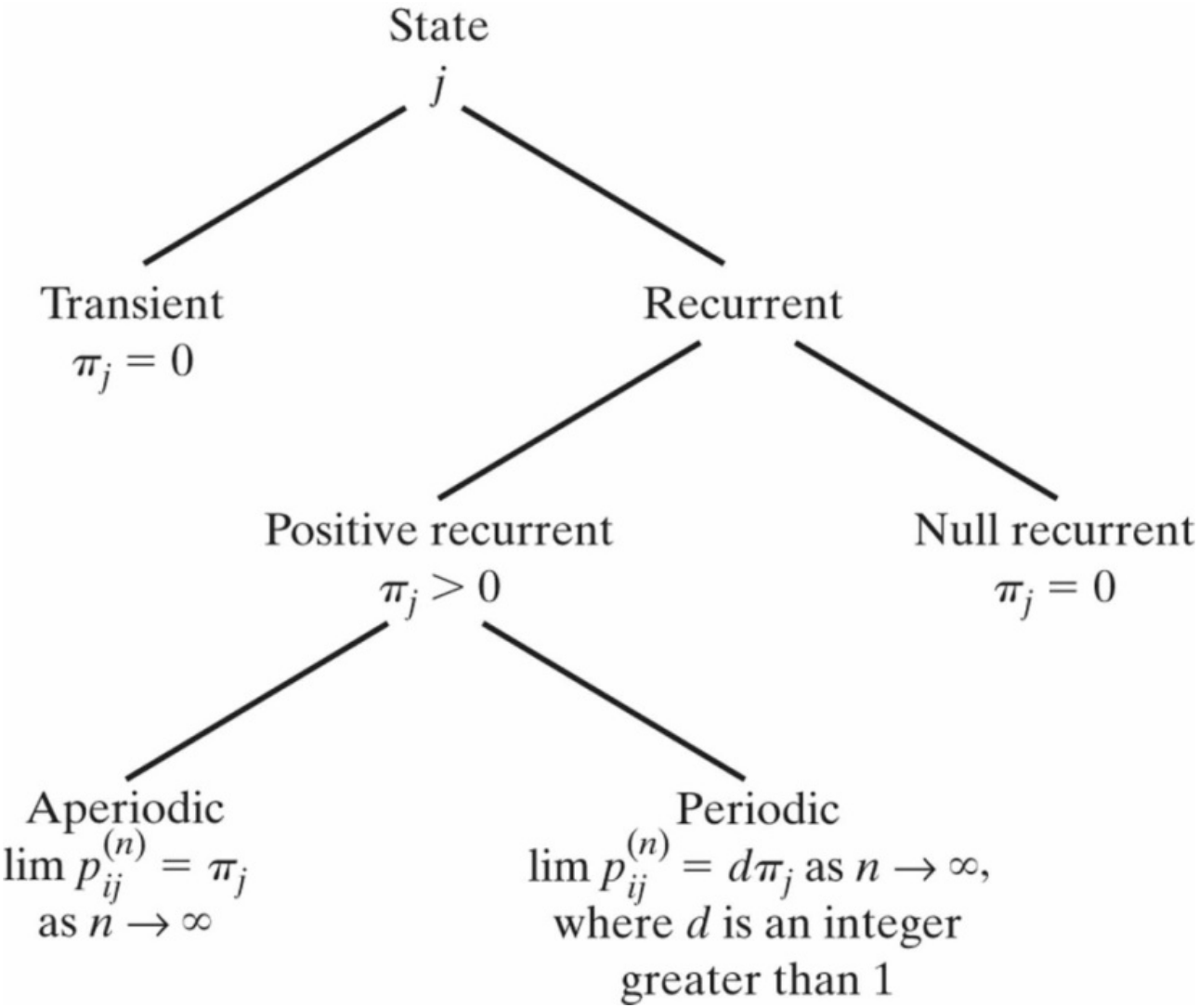


- **Irreducible Markov Chains**:
  Two states ***communicate*** $i \leftrightarrow j$ if the probability of reaching each other in finite number of steps is non-zero
  - If $i \leftrightarrow j$ & $i \leftrightarrow k \Rightarrow i \leftrightarrow k$ (***transitivity***)
  - If all states communicate the chain is ***Irreducible***

- **Classes**: States can be classified in subsets. ***Open classes*** allow exits to a different class. ***Closed classes*** do not allow external transitions and (possible after a transient interval) the process is limited within this subset. If there is a single closed calss, the chain is ***Irreducible***

STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Summary of Markov Chain State Classification



STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Transition Properties of Markov Chain (1/2)

- **Mean Recurrence Time**

  In a Markov Chain (***MC***) the ***Mean Recurrence Time*** $\mathrm{E}[T_i(k)]$ is defined as the average number of transitions (steps) for a ***Recurrent State*** i to ***cycle*** (return to itself), if it accomplished $k-1$ cycles from i to i. The relative occurrence of state i is proportional to the ***Steady State Probabilities***:

  $$\pi_i = \frac{1}{\mathrm{E}[T_i(k)]}$$

  If $\mathrm{E}[T_i(k)] < \infty$ then $\pi_i > 0$ and i is ***Positive Recurrent***. Else $\pi_i = 0$ and i is ***Null Recurrent***

- **Steady State Probabilities, Ergodicity**

  After l cycles into a ***Positive Recurrent State*** i, the proportion of steps (time) the ***MC*** resides in state i (***Sojourn Time***) is:

  $$\nu_i(l) = \frac{l}{\sum_{k=1}^{l} T_i(k)}$$

  The ***Recurrence Times*** $T_i(k)$ form a series of ***Independent Identically Distributed*** (***IID***) random variables. For $l \to \infty$ the proportions of transition steps the ***MC*** resides in state i approach to the ***Steady State Probabilities*** π_i:

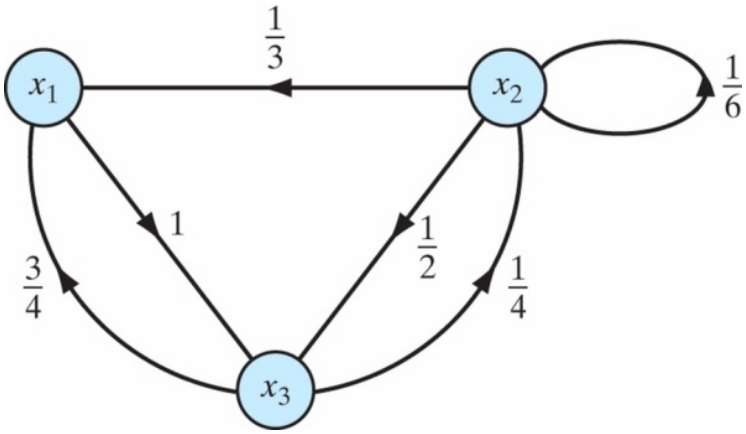
  $$\lim_{l\to\infty} \nu_i(l) = \pi_i,\ i = 1,2,\ldots,K$$

  The limit formula defines ***Ergodicity*** of state i (Ensemble Average = Time Average)

STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

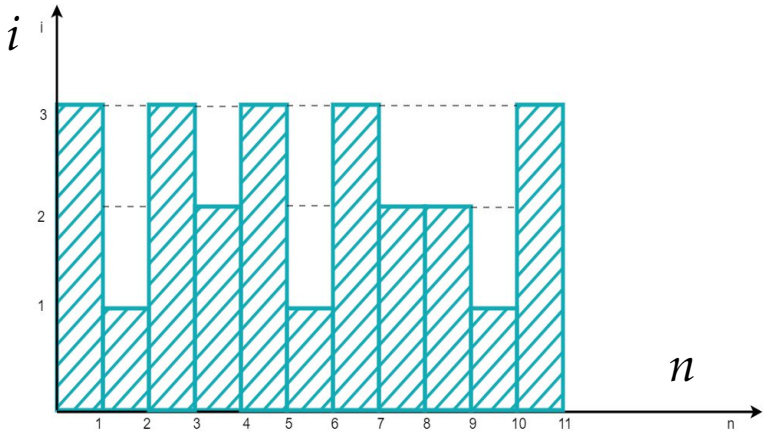
Transition Properties of Markov Chain (2/2)

MC State Transition Diagram: 3 State Example



$$\pi_1 = 0.3953 \quad \pi_2 = 0.1395 \quad \pi_3 = 0.4652$$

Sample State Trajectory $x_i(n) = i$



For State x_3, $i = 3$: $T_3(k)$, $v_3(l)$, $k \in \{1,2,\ldots,5\}$, $l = 5$

- $T_3(1) = 0$
- $T_3(2) = 2$
- $T_3(3) = 2$
- $T_3(4) = 2$
- $T_3(5) = 4$

- $v_3(5) = \frac{5}{0+2+2+2+4} = \frac{5}{10} = 0.5 \rightarrow \pi_3 = 0.4652$

STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Stationary Distribution of MC - Ergodic Probabilities

- Steady-state probabilities π_i of a ***MC*** are evaluated by measuring the relative frequency that state i occurs in an infinite (very large) repetition of sample outcomes
- Equivalently and for ***irreducible*** states steady-state probabilities can be deduced (measured) from the ***trajectory*** of a **single** outcome (time series) as the fraction of time that the process spends in i over an infinite (very lengthy) horizon.
- Such ***MC*** is defined as ***ergodic*** and the steady-state probabilities π_i as ***ergodic probabilities.*** An ***irreducible*** non-periodic ***MC*** is always ergodic

Convergence of State Probabilities to Ergodic Distributions

MC state probabilities of X_i, $i = 1,2,...,K$ at transition step $n = 0,1,2...$ define a $(1 \times K)$ vector $\boldsymbol{\pi}^{(n)}$ evolving according to the $(K \times K)$ ***stochastic transition matrix*** $\mathbf{P}$ starting from an initial conditions $\boldsymbol{\pi}^{(0)}$:

$$\boldsymbol{\pi}^{(n)} = \left[\pi_1^{(n)}\ \pi_2^{(n)} ... \pi_K^{(n)}\right],\ \boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(n-1)}\mathbf{P} = \boldsymbol{\pi}^{(n-2)}\mathbf{P}^2 = ... = \boldsymbol{\pi}^{(0)}\mathbf{P}^n$$

The ***ergodic probabilities*** $\boldsymbol{\pi} = [\pi_1\ \pi_2\ ...\pi_K]$ are defined as the limit $\boldsymbol{\pi}^{(n)}$ with $n \to \infty$:

$$\lim_{n\to\infty} \boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} \times \lim_{n\to\infty} \mathbf{P}^n = \boldsymbol{\pi}^{(0)} \begin{bmatrix} \pi_1 & ... & \pi_K \\ \vdots & \ddots & \vdots \\ \pi_1 & ... & \pi_K \end{bmatrix} = \boldsymbol{\pi}^{(0)} \begin{bmatrix} \boldsymbol{\pi} \\ \vdots \\ \boldsymbol{\pi} \end{bmatrix} = \sum_{j=1}^{K} \pi_j^{(0)} \times \boldsymbol{\pi} = 1 \times \boldsymbol{\pi} = \boldsymbol{\pi}$$

Hence $\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\pi}^{(n)}$ is independent of the initial condition $\boldsymbol{\pi}^{(0)}$ and can be evaluated by solving the linear system of $K-1$ ***Global Equilibrium Equations***, with an additional normalization linear equation to enforce ***linear independence*** of K equations:

$$\pi_j = \sum_{i=1}^{K} \pi_i p_{ij} \quad j = 1,2,...,K \text{ or } \boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P} \text{ and } \sum_{j=1}^{K} \pi_j = 1$$

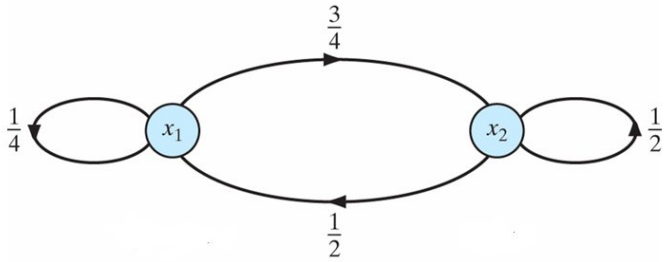
STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Examples of Ergodic Markov Chains

State Transition Diagrams

States are denoted as circles $x_1, x_2, \ldots$ and transitions $x_i \rightarrow x_j$ as arrows with probabilities p_{ij}

Calculation of Ergodic Probabilities



$$\mathbf{P} = \begin{bmatrix} 1/4 & 3/4 \\ 1/2 & 1/2 \end{bmatrix}, \quad \boldsymbol{\pi}^{(0)} = [1/6 \quad 5/6]$$

$$\boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}^{(0)}\mathbf{P} = [11/24 \quad 13/24]$$

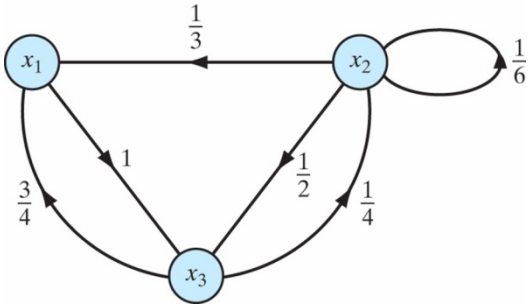
$$\mathbf{P}^2 = \begin{bmatrix} 0.4375 & 0.5625 \\ 0.3750 & 0.6250 \end{bmatrix}$$

$$\mathbf{P}^3 = \begin{bmatrix} 0.4001 & 0.5999 \\ 0.3999 & 0.6001 \end{bmatrix}$$

$$\mathbf{P}^4 = \begin{bmatrix} 0.4000 & 0.6000 \\ 0.4000 & 0.6000 \end{bmatrix}$$

$$\pi_1 = 0.4, \qquad \pi_2 = 0.6$$

Convergence in 4 iterations (steps)



$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ 1/3 & 1/6 & 1/2 \\ 3/4 & 1/4 & 0 \end{bmatrix}$$

Steady-state transition linear system

$$\pi_1 = \frac{1}{3}\pi_2 + \frac{3}{4}\pi_3$$

$$\pi_2 = \frac{1}{6}\pi_2 + \frac{1}{4}\pi_3$$

$$\pi_3 = \pi_1 + \frac{1}{2}\pi_2$$

$$\pi_1 + \pi_2 + \pi_3 = 1 \text{ (normalization)}$$

$$\pi_1 = 0.3953, \qquad \pi_2 = 0.1395, \quad \pi_3 = 0.4652$$

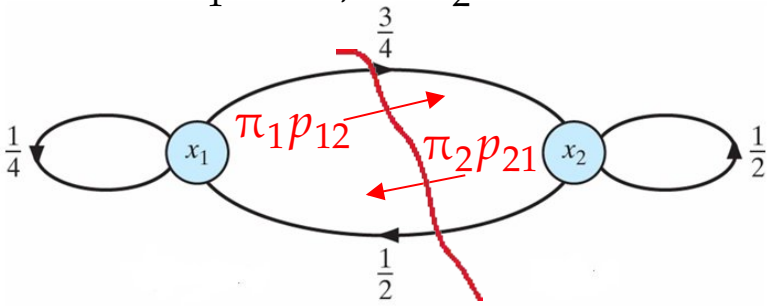
STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

Detailed Balanced Equations, Time Reversibility

- In ***thermal equilibrium*** a system that reaches state probabilities ***Gibbs*** $\pi_i = \frac{1}{Z} exp\left(-\frac{E_i}{T}\right)$ with Z the ***Partition Function***, state transitions $i \rightarrow j$ are ***balanced*** in the long term by reverse transitions $j \rightarrow i$

- In this case ***Detailed Balance Equations*** hold with the ergodic probabilities π_i satisfying ***Global Balance Equations*** as expected for all Markov Chains (***MC***):

$$\pi_1 = 0.4, \quad \pi_2 = 0.6$$



$$\pi_i p_{ij} = \pi_j p_{ji}$$

$$\sum_{i=1}^{K} \pi_i p_{ij} = \sum_{i=1}^{K} \left(\frac{\pi_i}{\pi_j} p_{ij}\right) \pi_j = \sum_{i=1}^{K} p_{ji} \pi_i = \pi_i$$

- The ***Detailed Balance Equations*** (unlike the ***Global Balance Equations***) do not hold for all ***MC***'s. An ***MC*** that abides by them is referred to as ***Time Reversible*** with forward transition probabilities p_{ij} and backwards with transition probabilities $\hat{p}_{ji} = \frac{\pi_i}{\pi_j} p_{ij}$

- The ergodic state probabilities π_i are much simpler to analyze if time-reversibility holds (e.g. for queuing networks under conditions yielding ***product-form*** state probabilities). Unfortunately, this is not true in many realistic models but holds for systems in thermal equilibrium that converge to ***Gibbs*** statistics (as in many ***Machine Learning*** systems)