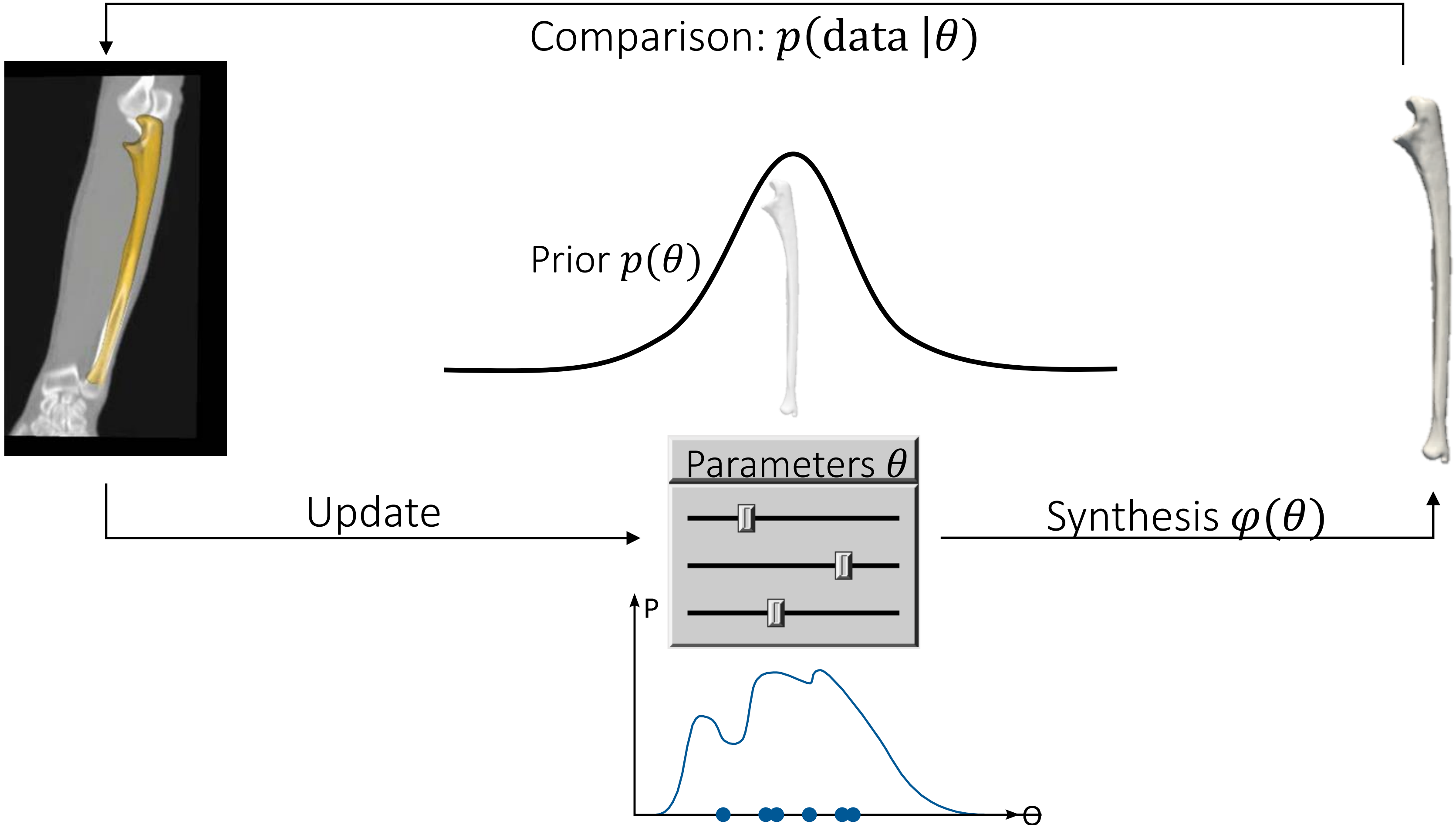
Understanding MCMC

Marcel Lüthi,

University of Basel

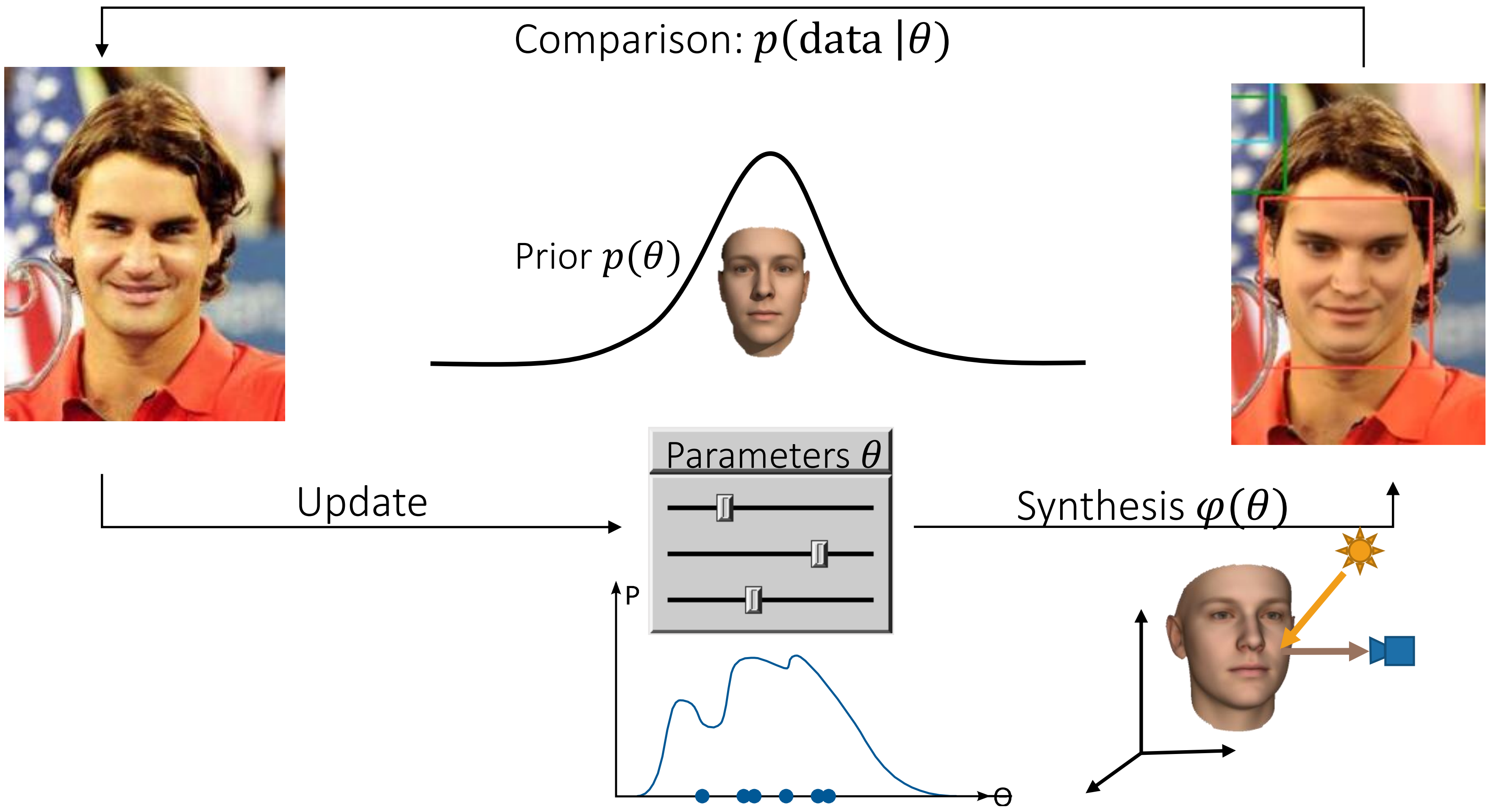
Slides based on presentation by Sandro Schönborn

Reminder: Analysis by synthesis



Goal: Estimate posterior $p(\theta|\text{data})$

Reminder: Analysis by synthesis



Goal: Estimate posterior $p(\theta|\text{data})$

Reminder: The Metropolis-Hastings Algorithm

Requirements:
- Proposal distribution $Q(\boldsymbol{x}'|\boldsymbol{x})$ – *must generate samples*
- Target distribution $P(\boldsymbol{x})$ – *with point-wise evaluation*

Tuning "knob" – influences convergence

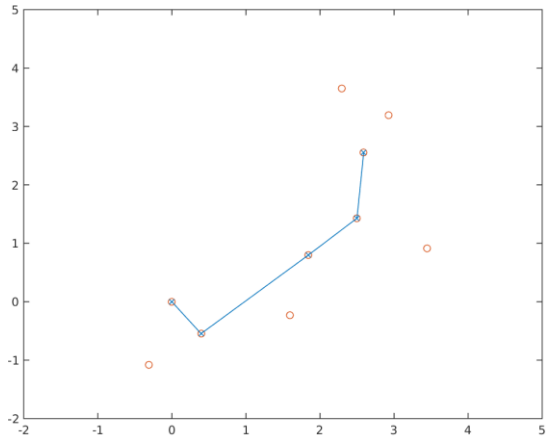
Target distribution in our problem= $p(\theta|data)$

Result:
- Stream of samples approximately from $P(\boldsymbol{x})$

- Initialize with sample $\boldsymbol{x}$
- Generate next sample, with current sample $\boldsymbol{x}$
  1. Draw a sample $\boldsymbol{x}'$ from $Q(\boldsymbol{x}'|\boldsymbol{x})$ ("proposal")
  2. With *probability* $\alpha = \min\left\{\frac{P(x')}{P(x)}\frac{Q(x|x')}{Q(x'|x)}, 1\right\}$ accept $\boldsymbol{x}'$ as new state $\boldsymbol{x}$
  3. Emit current state $\boldsymbol{x}$ as sample

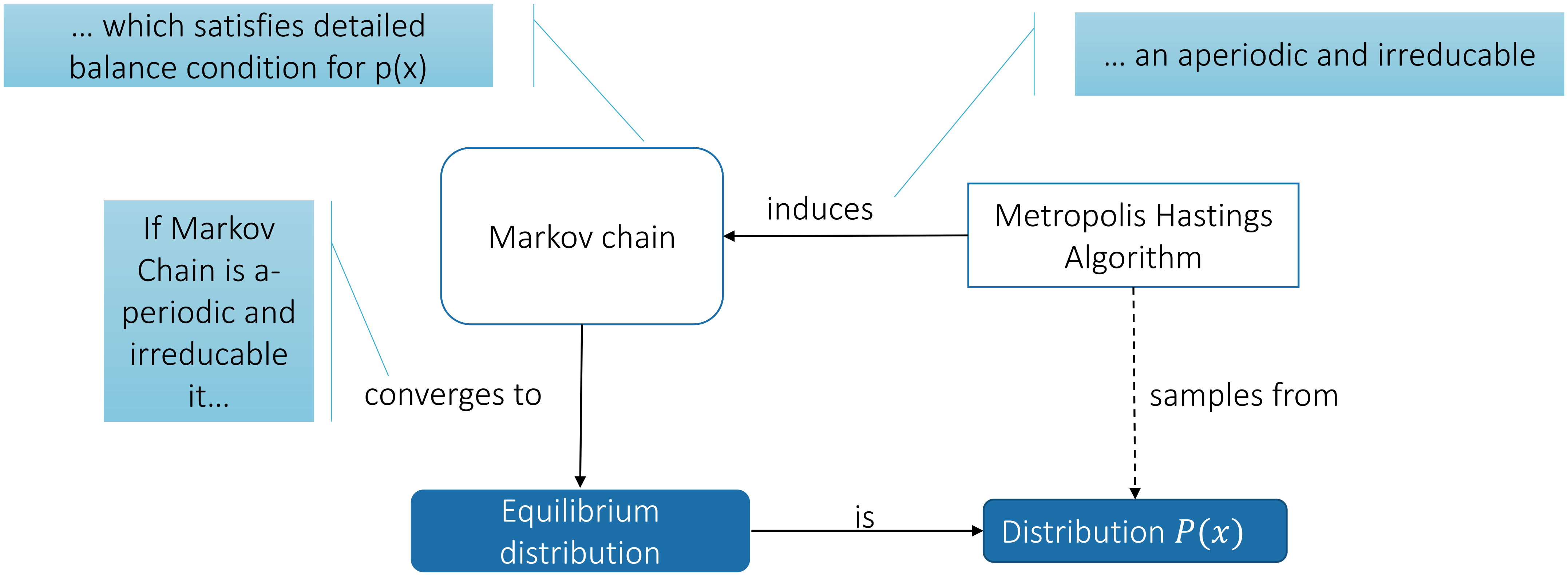
Reminder: The Metropolis-Hastings Algorithm

- Target: $P(\boldsymbol{x}) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$
- Proposal: $Q(\boldsymbol{x}'|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}'\,|\,\boldsymbol{x}, \sigma^2 I_2)$



Each sample is a random variable $X_i \sim P_i(\boldsymbol{x})$
Convergence: Distribution of X_i becomes $P(\boldsymbol{x})$ if $i \to \infty$

The big picture



Understanding Markov Chains

Markov Chain

State space

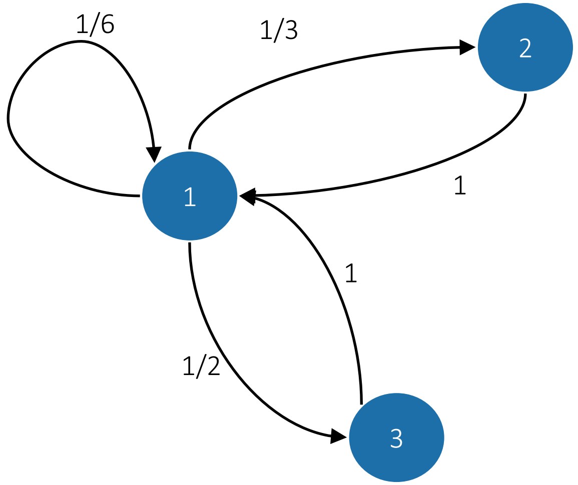
- Sequence of random variables $\{X_i\}_{i=1}^{N}$, $X_i \in S$ with Markov Property

$$P(X_i | X_1, X_2, \dots, X_{i-1}) = P(X_i | X_{i-1})$$

Transition probability

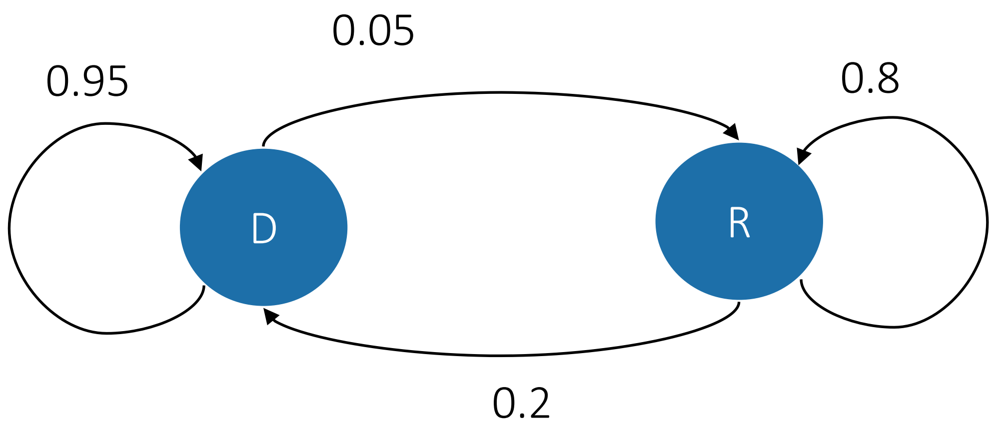
- Simplifications: (for our analysis)
  - Discrete state space: $S = \{1, 2, \dots, K\}$
  - Homogeneous Chain: $P(X_i = l | X_{i-1} = m) = T_{lm}$

Automatically true if we use computers (e.g. 32 bit floats)



- Can be simulated, for any given initial distribution X_1

Example: Markov Chain



- Simple weather model: *dry* (D) or *rainy* (R) hour
  - Condition in next hour? X_{t+1}
  - State space $S = \{D, R\}$
  - Stochastic: $P(X_{t+1}|X_t)$
  - Depends only on *current* condition X_t
- Draw samples from chain:
  - Initial: $X_0 = D$
  - Evolution: $P(X_{t+1}|X_t)$
- Long-term Behavior
  - Does it converge? *Average* probability of rain?
  - Dynamics? How *quickly* will it converge?

DDDDDDDDRRRRRRRRRRRDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
DDDDDDDDDDRDD...

Discrete Homogeneous Markov Chain

Formally linear algebra:

- Distribution (vector):

$$P(X_i):\ \boldsymbol{p_i} = \begin{bmatrix} P(X_i = 1) \\ \vdots \\ P(X_i = K) \end{bmatrix}$$

- Transition probability (transition matrix):

$$P(X_i | X_{i-1}):\ T = \begin{bmatrix} P(1 \leftarrow 1) & \cdots & P(1 \leftarrow K) \\ \vdots & \ddots & \vdots \\ P(K \leftarrow 1) & \cdots & P(K \leftarrow K) \end{bmatrix}$$

$$T_{lm} = P(l \leftarrow m) = P(X_i = l | X_{i-1} = m)$$

Evolution of the Initial Distribution

- Evolution of $P(X_1) \to P(X_2)$:

$$P(X_2 = l) = \sum_{m \in S} P(l \leftarrow m) P(X_1 = m)$$

$$\boldsymbol{p}_2 = T\boldsymbol{p}_1$$

- Evolution of n steps:

$$\boldsymbol{p}_{n+1} = T^n \boldsymbol{p}_1$$

- Is there a *stable* distribution $\boldsymbol{p}^*$? (steady-state)

$$\boldsymbol{p}^* = T\boldsymbol{p}^*$$

A stable distribution is an *eigenvector* of T with eigenvalue $\lambda = 1$

Steady-State Distribution: $\boldsymbol{p}^*$

- It exists:
  - T subject to normalization constraint: *left* eigenvector to eigenvalue 1

$$\sum_l T_{lm} = 1 \quad \Leftrightarrow \quad [1 \quad \dots \quad 1]T = [1 \quad \dots \quad 1]$$

  - T has eigenvalue $\lambda = 1$ (left-/right eigenvalues are the same)
  - Steady-state distribution as corresponding right eigenvector

$$T\boldsymbol{p}^* = \boldsymbol{p}^*$$

- Does *any* arbitrary initial distribution *evolve* to $\boldsymbol{p}^*$?
  - Convergence?
  - Uniqueness?

Equilibrium Distribution: $\boldsymbol{p}^*$

- Additional requirement for T: $(\mathrm{T}^{\mathrm{n}})_{lm} > 0$ for all $n > N_0$

  The chain is called *irreducible* and *aperiodic* (implies *ergodic*)

    - All states are connected using at most N_0 steps
    - Return intervals to a certain state are irregular

- *Perron-Frobenius* theorem for positive matrices:
  - PF1: $\lambda_1 = 1$ is a simple eigenvalue with 1d eigenspace (*uniqueness*)
  - PF2: $\lambda_1 = 1$ is dominant, all $|\lambda_i| < 1, \ \ i \neq 1$ (*convergence*)

- $\boldsymbol{p}^*$ is a stable attractor, called *equilibrium distribution*

$$T\boldsymbol{p}^* = \boldsymbol{p}^*$$

Convergence

- Time evolution of arbitrary distribution $\boldsymbol{p}_0$

$$\boldsymbol{p}_n = T^n \boldsymbol{p}_0$$

- Expand $\boldsymbol{p}_0$ in Eigen basis of T:

$$T\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i, \qquad |\lambda_i| < \lambda_1 = 1, \qquad |\lambda_k| \geq |\lambda_{k+1}|$$

$$\boldsymbol{p}_0 = \sum_i^K c_i \boldsymbol{e}_i$$

$$T\boldsymbol{p}_0 = \sum_i^K c_i \lambda_i \boldsymbol{e}_i$$

$$T^n \boldsymbol{p}_0 = \sum_i^K c_i \lambda_i^n \boldsymbol{e}_i = c_1 \boldsymbol{e}_1 + \lambda_2^n c_2 \boldsymbol{e}_2 + \lambda_3^n c_3 \boldsymbol{e}_3 + \cdots$$

Convergence (II)

$$T^n \boldsymbol{p}_0 = \sum_i^K c_i \lambda_i^n \boldsymbol{e}_i = c_1 \boldsymbol{e}_1 + \lambda_2^n c_2 \boldsymbol{e}_2 + \lambda_3^n c_3 \boldsymbol{e}_3 + \cdots$$

$$(n \gg 1) \qquad \approx \boldsymbol{p}^* + \lambda_2^n c_2 \boldsymbol{e}_2$$

$c_1 \boldsymbol{e}_1 = \boldsymbol{p}^*$

Normalizations:
$\|\boldsymbol{e}_1\| = 1$
$\sum_i p_i^* = 1$

- We have *convergence*:

$$T^n \boldsymbol{p}_0 \xrightarrow{n \to \infty} \boldsymbol{p}^*$$

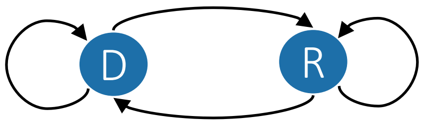
- *Rate* of convergence:

$$\|\boldsymbol{p}_n - \boldsymbol{p}^*\| \approx \|\lambda_2^n c_2 \boldsymbol{e}_2\| = |\lambda_2|^n |c_2|$$

Example: Weather Dynamics

Rain forecast for stable versus mixed weather:

stable $W_s = \begin{bmatrix} 0.95 & 0.2 \\ 0.05 & 0.8 \end{bmatrix}$



mixed $W_m = \begin{bmatrix} 0.85 & 0.6 \\ 0.15 & 0.4 \end{bmatrix}$

$\boldsymbol{p}^* = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}$

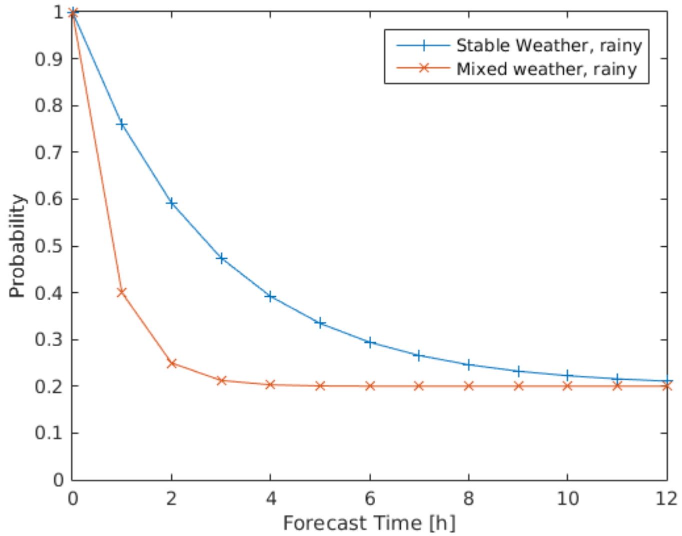
Long-term average probability of rain: **20%**

$\boldsymbol{p}^* = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}$

Eigenvalues: 1, **0.75**

Rainy now, next hours?

$RRRRDDDDDDDDDDDDD$
$DDDDDDDDDDDDDD...$



Eigenvalues: 1, **0.25**

Rainy now, next hours?

$RDDDDDDDDDDDDDDDDD$
$RDDDRDDDDDDDDD...$

Markov Chain: First Results

- *Aperiodic* and *irreducible* chains are *ergodic*:
  (every state reachable after $> N$ steps, irregular return time)
  - Convergence towards a unique *equilibrium distribution* $\boldsymbol{p}^*$
- Equilibrium distribution $\boldsymbol{p}^*$
  - Eigenvector of T with eigenvalue $\lambda = 1$:

$$T\boldsymbol{p}^* = \boldsymbol{p}^*$$

  - Rate of convergence:

Exponential decay with second largest eigenvalue $\propto |\lambda_2|^n$

Only useful if we can design chain with desired equilibrium distribution!

Detailed Balance

- Special property of some Markov chains

  Distribution p satisfies *detailed balance* if the total flow of probability between every pair of states is equal, (we have a local equilibrium):

$$P(l \leftarrow m)p(m) = P(m \leftarrow l)p(l)$$

- Detailed balance implies: p is the equilibrium distribution

$$(T\boldsymbol{p})_l = \sum_m T_{lm} p_m = \sum_m T_{ml} p_l = p_l$$

- Most MCMC methods construct chains which satisfies detailed balance.

The Metropolis-Hastings Algorithm

MCMC to draw samples from an arbitrary distribution

Idea of Metropolis Hastings algorithm

- Design a Markov Chain, which satisfies the detailed balance condition

$$T_{MH}(x' \leftarrow x)P(x) = T_{MH}(x \leftarrow x')P(x')$$

- *Ergodicity ensures that chain converges to this distribution*

Attempt 1: A simple algorithm

- Initialize with sample $\boldsymbol{x}$
- Generate next sample, with current sample $\boldsymbol{x}$
    1. Draw a sample $\boldsymbol{x}'$ from $Q(\boldsymbol{x}'|\boldsymbol{x})$ ("proposal")
    2. Emit current state $\boldsymbol{x}$ as sample

- It's a Markov chain
- Need to choose Q for every P to satisfy detailed balance

$$Q(x' \leftarrow x)P(x) = Q(x \leftarrow x')P(x')$$

Attempt 2: More general solution

- Initialize with sample $\boldsymbol{x}$
- Generate next sample, with current sample $\boldsymbol{x}$
  1. Draw a sample $\boldsymbol{x}'$ from $Q(\boldsymbol{x}'|\boldsymbol{x})$ ("proposal")
  2. With *probability* $\alpha(\mathrm{x}, \mathrm{x}')$ emit $\boldsymbol{x}'$ as new sample
  3. With *probability* $1 - \alpha(\mathrm{x}, \mathrm{x}')$ emit x as new sample

- It's a Markov chain
- Decouples Q from P through acceptance rule a
  - How to choose a?

What is the acceptance function a?

$$T_{MH}(x' \leftarrow x)P(x) = T_{MH}(x \leftarrow x')P(x')$$
$$a(x'|x)Q(x'|x)P(x) = a(x|x')Q(x|x')P(x')$$

Case A: $x' = x$

- Detailed balance trivially satisfied for every a(x’,x)

Case B: $x' \neq x$

- We have the following requirement

$$\frac{a(x'|x)}{a(x|x')} = \frac{Q(x|x')P(x')}{Q(x'|x)P(x)}$$

What is the acceptance function a?

Requirement: Choose $a(x'|x)$ *such that*

$$\frac{a(x'|x)}{a(x|x')} = \frac{Q(x|x')P(x')}{Q(x'|x)P(x)}$$

- $a(x|x')$ is probability distribution $a(x|x') \leq 1$ and $a(x|x') \geq 0$
- Easy to check that:

$$a(x'|x) = \min\left(1, \frac{Q(x|x')P(x')}{Q(x'|x)P(x)}\right)$$

  satisfies this property.

What is the acceptance function a?

Case 1:

$\frac{Q(x'|x)P(x)}{Q(x|x')P(x')} > 1$

$$\frac{a(x'|x)}{a(x|x')} = \frac{\min\left(1, \frac{Q(x|x')P(x')}{Q(x'|x)P(x)}\right)}{\min\left(1, \frac{Q(x'|x)P(x)}{Q(x|x')P(x')}\right)} = \frac{\frac{Q(x|x')P(x')}{Q(x'|x)P(x)}}{1} = \frac{Q(x|x')P(x')}{Q(x'|x)P(x)}$$

Case 2:

$\frac{Q(x|x')P(x')}{Q(x'|x)P(x)} > 1$

$$\frac{a(x'|x)}{a(x|x')} = \frac{\min\left(1, \frac{Q(x|x')P(x')}{Q(x'|x)P(x)}\right)}{\min\left(1, \frac{Q(x'|x)P(x)}{Q(x|x')P(x')}\right)} = \frac{1}{\frac{Q(x'|x)P(x)}{Q(x|x')P(x')}} = \frac{Q(x|x')P(x')}{Q(x'|x)P(x)}$$

The big picture