Markov Chain Monte Carlo Methods

Michel Bierlaire

michel.bierlaire@epfl.ch

Transport and Mobility Laboratory

Markov Chains

Andrey Markov, 1856–1922, Russian mathematician.

Markov Chains

Glossary:

- Stochastic process: X_t, $t = 0, 1, \ldots,$, collection of r.v. with same support, or *states space* $\{1, \ldots, i, \ldots, J\}$.
- Markov process: (short memory)

$$\Pr(X_t = i | X_0, \ldots, X_{t-1}) = \Pr(X_t = i | X_{t-1})$$

- Homogeneous Markov process:

$$\Pr(X_t = j | X_{t-1} = i) = \Pr(X_{t+k} = j | X_{t-1+k} = i) = P_{ij} \;\; \forall t \geq 1, k \geq 0.$$

- Transition matrix: $P \in \mathbb{R}^{J \times J}$.
- Properties:

$$\sum_{j=1}^{J} P_{ij} = 1, \; i = 1, \ldots, J, \;\; P_{ij} \geq 0, \; \forall i, j,$$

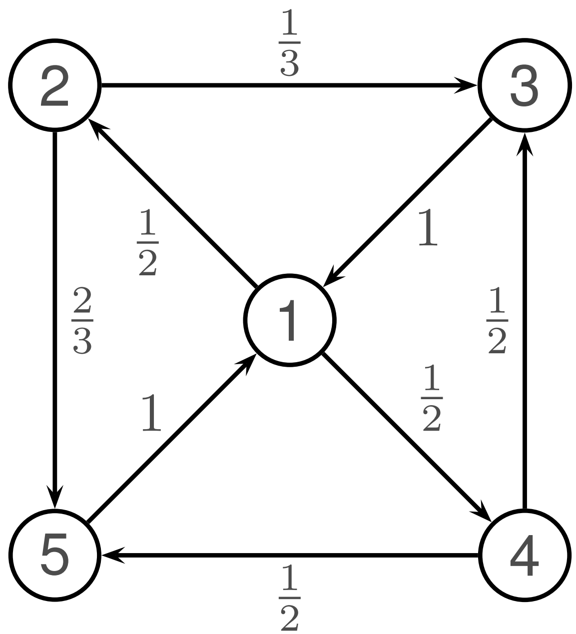
Markov Chains

- If state j can be reached from state i with non zero probability, we say that i *communicates* with j.
- Two states that communicate belong to the same *class*.
- A Markov chain is *irreducible* or *ergodic* if it contains only one class.
- With an ergodic chain, it is possible to go to every state from any state.

Markov Chains

- P^t_{ij} is the probability that the process reaches state j from i after t steps.
- Consider all t such that $P^t_{ii} > 0$. The largest common divisor d is called the *period* of state i.
- A state with period 1 is *aperiodic*.
- If $P_{ii} > 0$, state i is aperiodic.
- The period is the same for all states in the same class.
- Therefore, if the chain is irreducible, if one state is aperiodic, they all are.

A periodic chain



$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad d = 3.$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Markov Chains

$$\Pr(j) = \sum_{i=1}^{J} \Pr(j|i) \Pr(i)$$

- Stationary probabilities: unique solution of the system

$$\pi_j = \sum_{i=1}^{J} P_{ij} \pi_i, \quad \forall j = 1, \ldots, J. \tag{1}$$

$$\sum_{j=1}^{J} \pi_j = 1.$$

- Solution exists for any irreducible chain.

Markov Chains

- Consider the following system of equations:

$$x_i P_{ij} = x_j P_{ji}, \quad i \neq j, \quad \sum_{i=1}^{J} x_i = 1 \tag{2}$$

- We sum over i:

$$\sum_{i=1}^{J} x_i P_{ij} = x_j \sum_{i=1}^{J} P_{ji} = x_j.$$

- If (2) has a solution, it is also a solution of (1). As π is the unique solution of (1) then $x = \pi$.

$$\pi_i P_{ij} = \pi_j P_{ji}, \quad i \neq j$$

- The chain is said *time reversible*

Example

- A machine can be in 4 states with respect to wear
    - perfect condition,
    - partially damaged,
    - seriously damaged,
    - completely useless.
- The degradation process can be modeled by an irreducible aperiodic homogeneous Markov process, with the following transition matrix:

$$P = \begin{pmatrix} 0.95 & 0.04 & 0.01 & 0.0 \\ 0.0 & 0.90 & 0.05 & 0.05 \\ 0.0 & 0.0 & 0.80 & 0.20 \\ 1.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}$$

Example

Stationary distribution: $\left(\frac{5}{8}, \frac{1}{4}, \frac{3}{32}, \frac{1}{32}\right)$

$$\left(\frac{5}{8}, \frac{1}{4}, \frac{3}{32}, \frac{1}{32}\right) \begin{pmatrix} 0.95 & 0.04 & 0.01 & 0.0 \\ 0.0 & 0.90 & 0.05 & 0.05 \\ 0.0 & 0.0 & 0.80 & 0.20 \\ 1.0 & 0.0 & 0.0 & 0.0 \end{pmatrix} = \left(\frac{5}{8}, \frac{1}{4}, \frac{3}{32}, \frac{1}{32}\right)$$

- Machine in perfect condition 5 days out of 8, in average.
- Repair occurs in average every 32 days

From now on: Markov process = irreducible aperiodic homogeneous Markov process

Stationary distributions

- Property:

$$\pi_j = \lim_{t \to \infty} \Pr(X_t = j) \quad j = 1, \ldots, J.$$

- Ergodicity:
  - Let f be any function on the state space.
  - Then, with probability 1,

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \sum_{j=1}^{J} \pi_j f(j).$$

  - Computing the expectation of a function of the stationary states is the same as to take the average of the values along a trajectory of the process.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Simulation

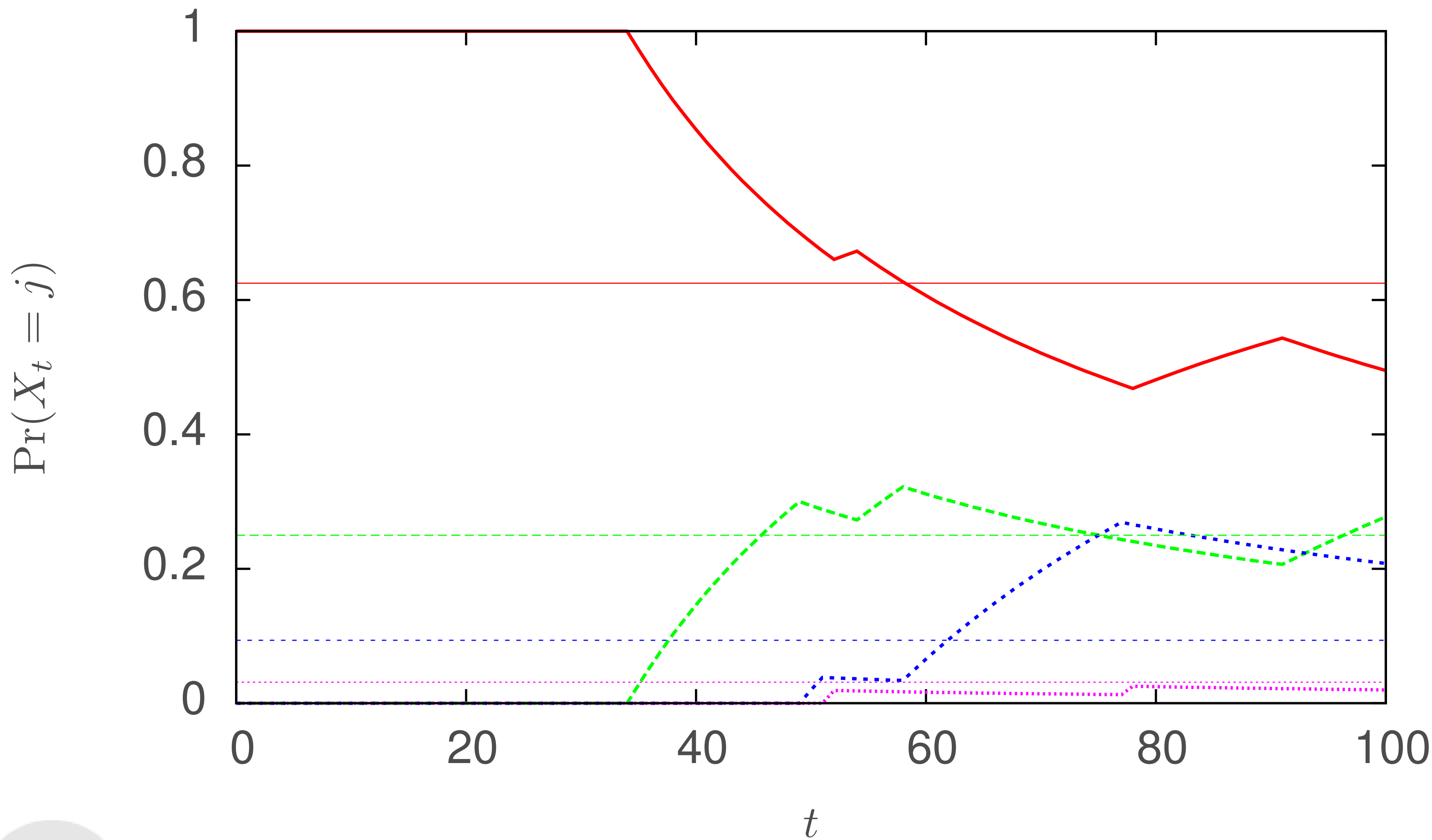
- We want to simulate a r.v. X with pmf

$$\Pr(X = j) = p_j.$$

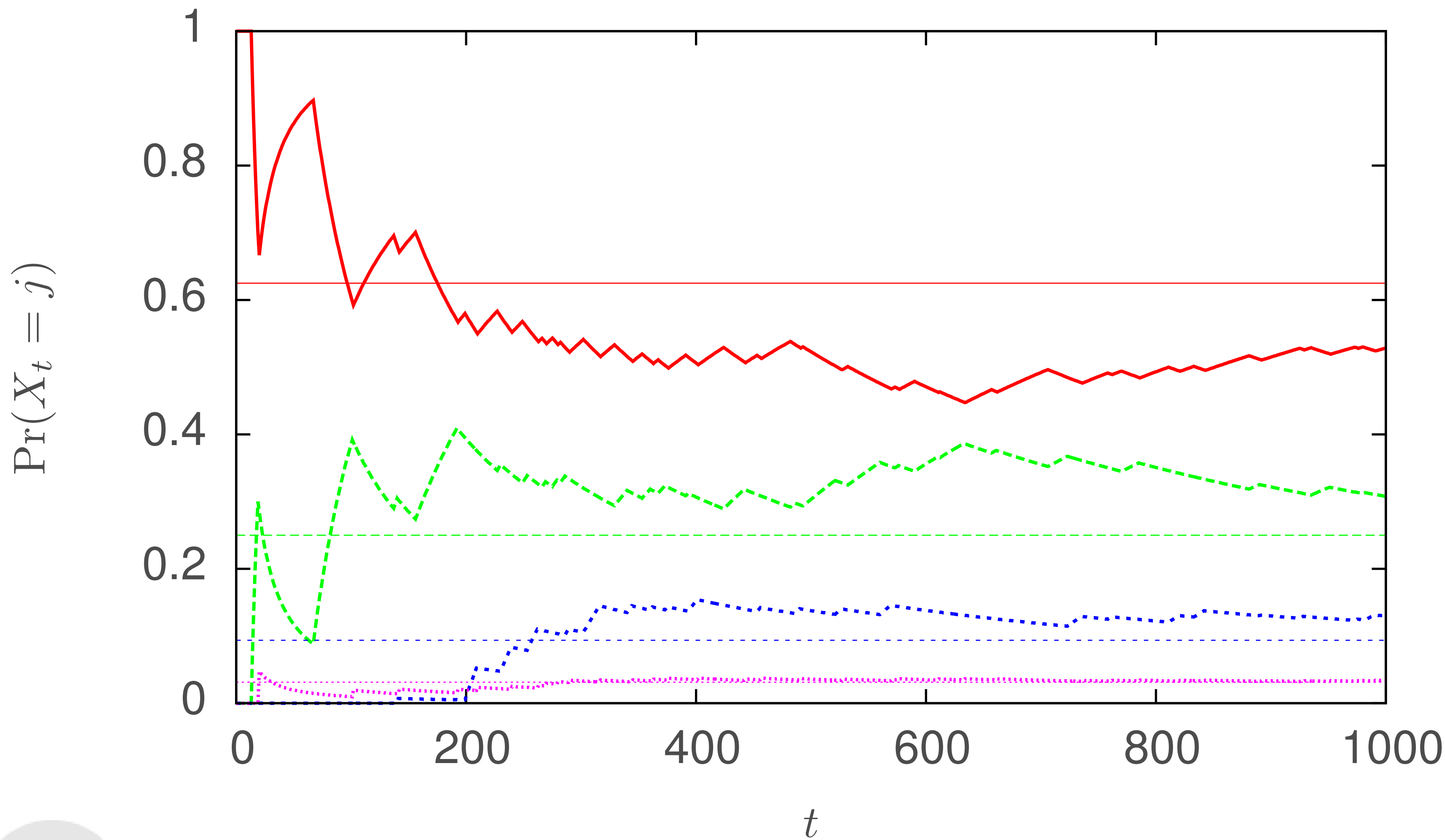
- We generate a Markov process with limiting probabilities p_j (how?)
- We simulate the evolution of the process.

$$p_j = \pi_j = \lim_{t \to \infty} \Pr(X_t = j) \quad j = 1, \ldots, J.$$

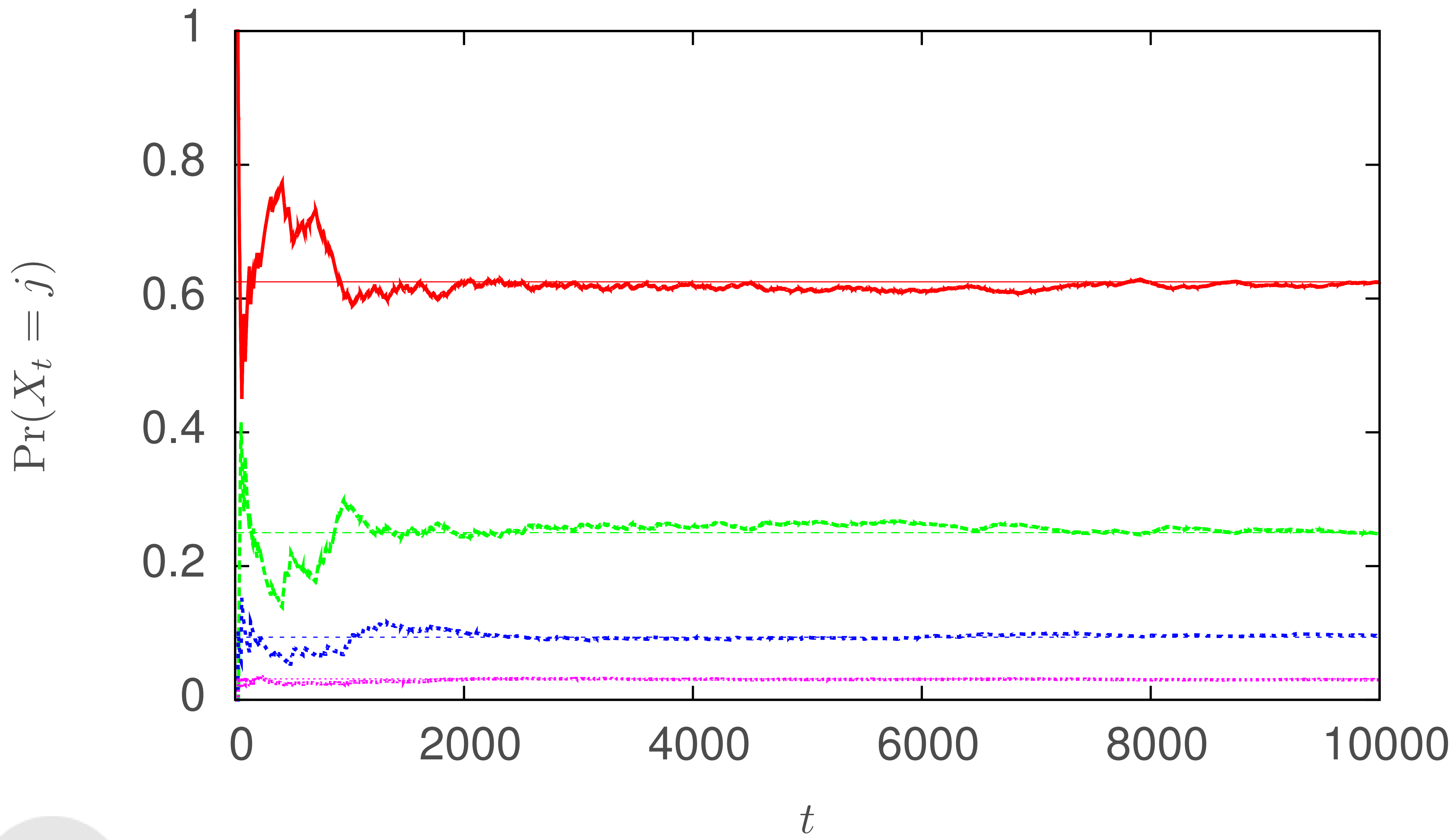
Example: $T = 100$



Example: $T = 1000$



Example: $T = 10000$



Simulation

- Assume that we are interested in simulating

$$\mathrm{E}[f(X)] = \sum_{j=1}^{J} f(j) p_j .$$

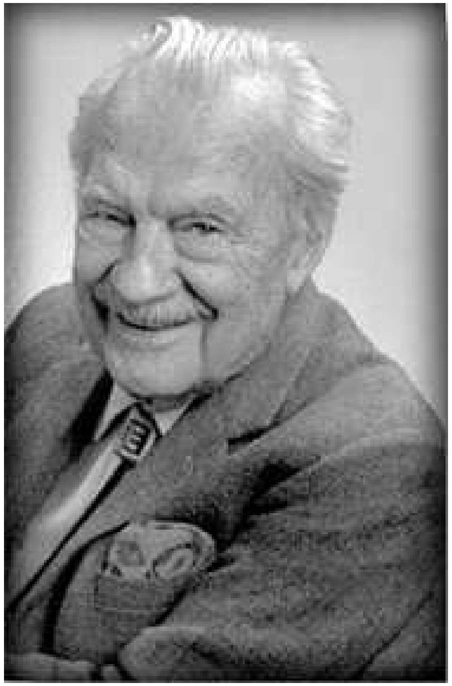
- We use ergodicity to estimate it with

$$\frac{1}{T} \sum_{t=1}^{T} f(X_t).$$

- We should drop early states (see above example). Better estimate:

$$\frac{1}{T} \sum_{t=1+k}^{T+k} f(X_t).$$

Metropolis-Hastings



Nicholas Metropolis
1915 – 1999

W. Keith Hastings
1930 –

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Metropolis-Hastings

- Let b_j, $j = 1, \ldots, J$ be positive numbers.
- Let $B = \sum_j b_j$. If J is huge, B cannot be computed.
- Let $\pi_j = b_j / B$.
- We want to simulate a r.v. with pmf π_j.
- Consider a Markov process on $\{1, \ldots, J\}$ with transition probability Q.
- Define another Markov process with the same states in the following way:
  - Assume the process is in state i, that is $X_t = i$,
  - Simulate the (candidate) next state j according to Q.
  - Define

$$X_{t+1} = \begin{cases} j & \text{with probability } \alpha_{ij} \\ i & \text{with probability } 1 - \alpha_{ij} \end{cases}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Metropolis-Hastings

- Transition probability P:

$$\begin{array}{rcll} P_{ij} & = & Q_{ij}\alpha_{ij} & \text{if } i \neq j \\ P_{ii} & = & Q_{ii}\alpha_{ii} + \sum_{\ell\neq i} Q_{i\ell}(1-\alpha_{i\ell}) & \text{otherwise} \end{array}$$

- Must verify the property:

$$\begin{array}{rcl} 1 = \sum_j P_{ij} & = & P_{ii} + \sum_{j\neq i} P_{ij} \\ & = & Q_{ii}\alpha_{ii} + \sum_{\ell\neq i} Q_{i\ell}(1-\alpha_{i\ell}) + \sum_{j\neq i} Q_{ij}\alpha_{ij} \\ & = & Q_{ii}\alpha_{ii} + \sum_{\ell\neq i} Q_{i\ell} - \sum_{\ell\neq i} Q_{i\ell}\alpha_{i\ell} + \sum_{j\neq i} Q_{ij}\alpha_{ij} \\ & = & Q_{ii}\alpha_{ii} + \sum_{\ell\neq i} Q_{i\ell} \end{array}$$

As $\sum_j Q_{ij} = 1$, we have $\alpha_{ii} = 1$.

Metropolis-Hastings

- Stationary distribution and time reversibility:

$$\pi_i P_{ij} = \pi_j P_{ji}, \quad i \neq j$$

- that is

$$\pi_i Q_{ij} \alpha_{ij} = \pi_j Q_{ji} \alpha_{ji}, \quad i \neq j$$

- It is satisfied if

$$\alpha_{ij} = \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}} \text{ and } \alpha_{ji} = 1$$

or

$$\frac{\pi_i Q_{ij}}{\pi_j Q_{ji}} = \alpha_{ji} \text{ and } \alpha_{ij} = 1$$

Metropolis-Hastings

- Therefore

$$\alpha_{ij} = \min\left(\frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}, 1\right)$$

- Remember: $\pi_j = b_j/B$. Therefore

$$\alpha_{ij} = \min\left(\frac{b_j B Q_{ji}}{b_i B Q_{ij}}, 1\right) = \min\left(\frac{b_j Q_{ji}}{b_i Q_{ij}}, 1\right)$$

- The normalization constant B does not play a role in the computation of α_{ij}.
- In summary:
  - Given Q and b_j
  - defining α as above
  - creates a Markov process characterized by P
  - with stationary distribution π.

Metropolis-Hastings

Algorithm:

1. Choose a Markov process characterized by Q.
2. Initialize the chain with a state i: $t = 0$, $X_0 = i$.
3. Simulate the (candidate) next state j based on Q.
4. Let r be a draw from $U[0,1[$.
5. Compare r with $\alpha_{ij} = \min\left(\frac{b_j Q_{ji}}{b_i Q_{ij}}, 1\right)$. If

$$r < \frac{b_j Q_{ji}}{b_i Q_{ij}}$$

then $X_{t+1} = j$, else $X_{t+1} = i$.

6. Increase t by one.
7. Goto step 3.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Example

$$\begin{aligned} b &= (20, 8, 3, 1) \\ \pi &= \left(\frac{5}{8}, \frac{1}{4}, \frac{3}{32}, \frac{1}{32}\right) \end{aligned}$$

$$Q = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

Run MH for 10000 iterations. Collect statistics after 1000.

- Accept: [2488, 1532, 801, 283]
- Reject: [0, 952, 1705, 2239]
- Simulated: [0.627, 0.250, 0.095, 0.028]
- Target: [0.625, 0.250, 0.09375, 0.03125]

Gibbs sampling

- Let $X = (X^1, X^2, \ldots, X^n)$ be a random vector with pmf (or pdf) $p(x)$.
- Assume we can draw from the marginals:

$$\Pr(X^i | X^j = x^j,\ j \neq i),\ i = 1, \ldots, n.$$

- Markov process. Assume current state is x.
  - Draw randomly (equal probability) a coordinate i.
  - Draw r from the ith marginal.
  - New state: $y = (x^1, \ldots, x^{i-1}, r, x^{i+1}, \ldots, x^n)$.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Gibbs sampling

- Transition probability:

$$Q_{xy} = \frac{1}{n}\Pr(X^i = r | X^j = x^j,\ j \neq i) = \frac{p(y)}{n\Pr(X^j = x^j,\ j \neq i)}$$

- The denominator is independent of X_i.
- So Q_{xy} is proportional to $p(y)$.
- Metropolis-Hastings:

$$\alpha_{xy} = \min\left(\frac{p(y)Q_{yx}}{p(x)Q_{xy}}, 1\right) = \min\left(\frac{p(y)p(x)}{p(x)p(y)}, 1\right) = 1$$

- The candidate state is always accepted.

Example: bivariate normal distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$$

Marginal distribution:

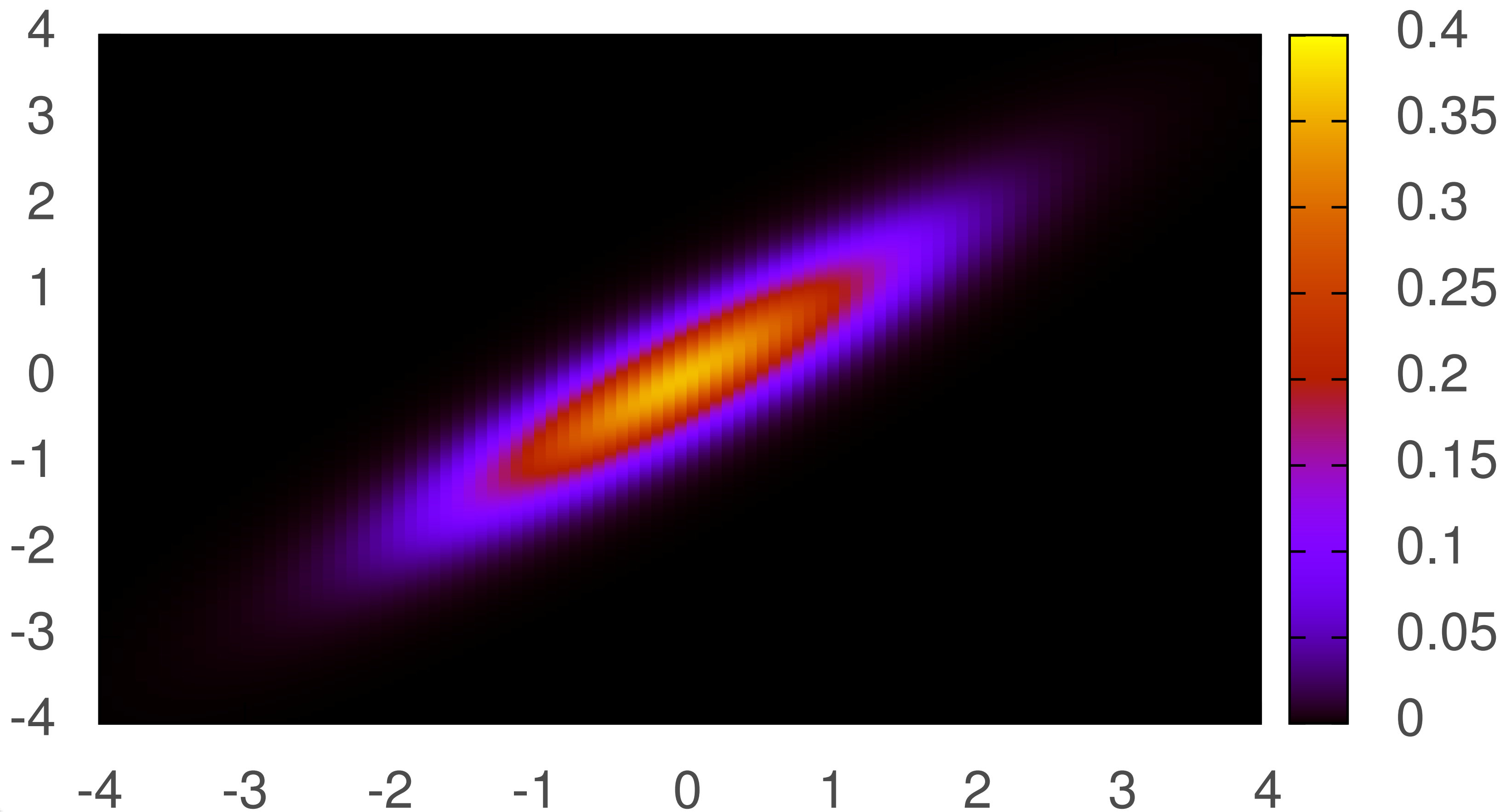
$$Y|(X = x) \sim N\left(\mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x - \mu_X), (1 - \rho^2)\sigma_Y^2\right)$$

Apply Gibbs sampling to draw from:

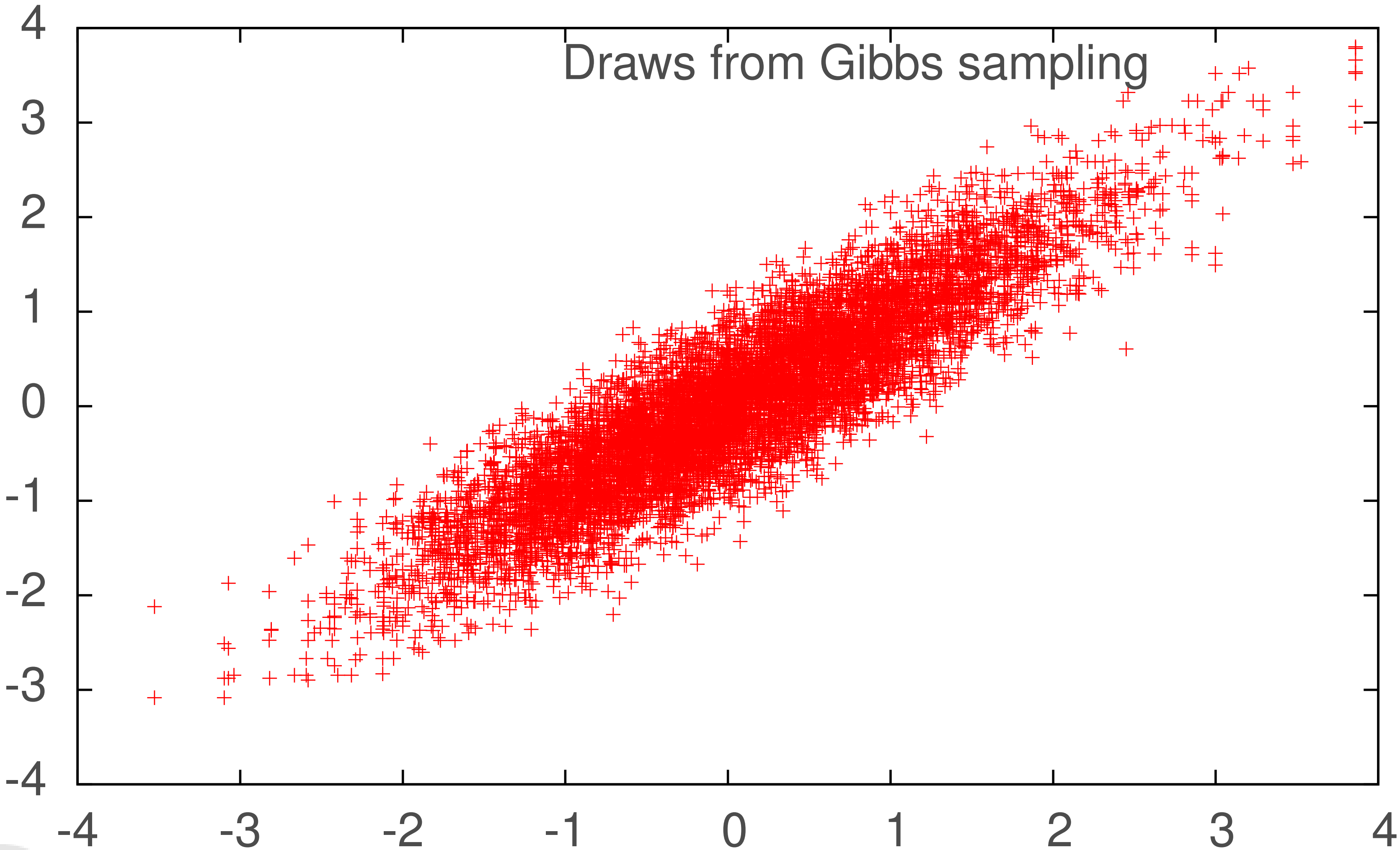
$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right)$$

Note: just for illustration. Should use Cholesky factor.

Example: pdf



Example: draws from Gibbs sampling



Simulated annealing

- Application of the Metropolis-Hastings algorithm to optimization.
- Name comes from analogy with annealing in metallurgy, involving heating and controlled cooling of a material to reduce its defects.
- Optimization problem:
$$\min_{x \in \mathcal{F}} f(x)$$
where the feasible set $\mathcal{F}$ is a finite set of vectors.
- Let $\mathcal{X}^*$ be the set of optimal solutions, that is
$$\mathcal{X}^* = \{x \in \mathcal{F} | f(x) \leq f(y), \ \forall y \in \mathcal{F}\} \text{ and } f(x^*) = f^*, \ \forall x^* \in \mathcal{X}^*.$$
- Consider the pmf on $\mathcal{F}$
$$p_\lambda(x) = \frac{e^{-\lambda f(x)}}{\sum_{y \in \mathcal{F}} e^{-\lambda f(y)}}, \ \lambda > 0.$$

Simulated annealing

$$p_\lambda(x) = \frac{e^{-\lambda f(x)}}{\sum_{y \in \mathcal{F}} e^{-\lambda f(y)}}$$

- Equivalently

$$p_\lambda(x) = \frac{e^{\lambda(f^* - f(x))}}{\sum_{y \in \mathcal{F}} e^{\lambda(f^* - f(y))}}$$

- As $f^* - f(x) \leq 0$, when $\lambda \to \infty$, we have

$$\lim_{\lambda \to \infty} p_\lambda(x) = \frac{\delta(x \in \mathcal{X}^*)}{|\mathcal{X}^*|},$$

where

$$\delta(x \in \mathcal{X}^*) = \begin{cases} 1 & \text{if } x \in \mathcal{X}^* \\ 0 & \text{otherwise.} \end{cases}$$

Example

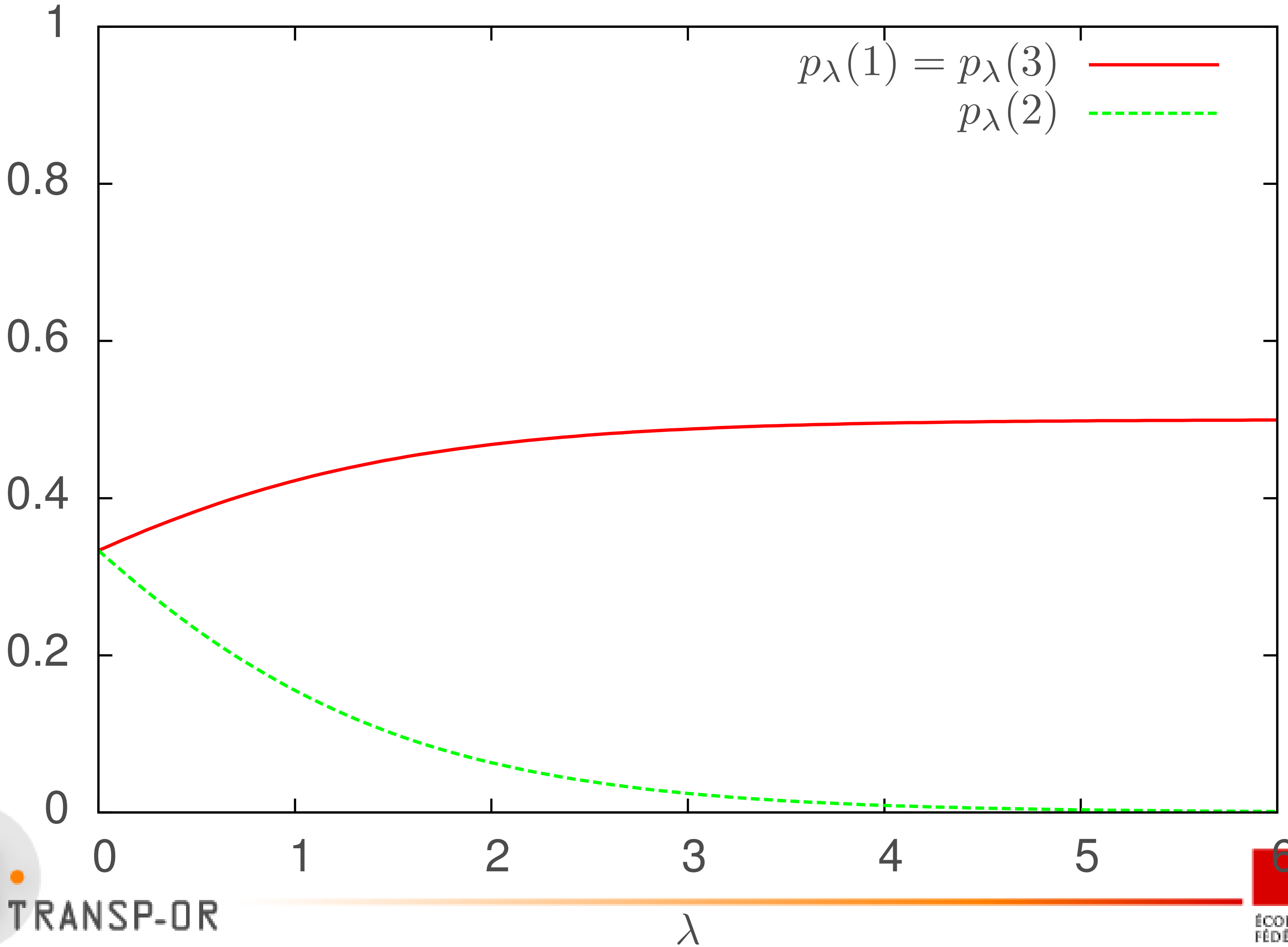
$$\mathcal{F} = \{1, 2, 3\} \; f(\mathcal{F}) = \{0, 1, 0\}$$

$$p_\lambda(1) = \frac{1}{2 + e^{-\lambda}}$$

$$p_\lambda(2) = \frac{e^{-\lambda}}{2 + e^{-\lambda}}$$

$$p_\lambda(3) = \frac{1}{2 + e^{-\lambda}}$$

Example



Simulated annealing

- If λ is large,
- we generate a Markov chain with stationary distribution $p_\lambda(x)$.
- The mass is concentrated on optimal solutions.
- As the normalizing constant is not needed, only $e^{\lambda(f^*-f(x))}$ is used.
- Construction of the Markov process through the concept of *neighborhood*.
- A *neighbor* y of x is obtained by simple modifications of x.
- The Markov process will proceed from neighbors to neighbors.
- The neighborhood structure must be designed such that the chain is irreducible, that is the whole space $\mathcal{F}$ must be covered.
- It must be designed also such that the size of the neighborhood is reasonably small.

Neighborhood

- Examples of neighborhoods:
  - x and y are neighbors if they differ only in one coordinate.
  - x and y are neighbors if two elements are interchanged.
- Denote $N(x)$ the set of neighbors of x.
- Define a Markov process where the next state is a randomly drawn neighbor.
- Transition probability:

$$Q_{xy} = \frac{1}{|N(x)|}$$

- Metropolis Hastings:

$$\alpha_{xy} = \min\left(\frac{p(y)Q_{yx}}{p(x)Q_{xy}}, 1\right) = \min\left(\frac{e^{-\lambda f(y)}|N(x)|}{e^{-\lambda f(x)}|N(y)|}, 1\right)$$

Notes

- The neighborhood structure can always be arranged so that each vector has the same number of neighbors. In this case,

$$\alpha_{xy} = \min\left(\frac{e^{-\lambda f(y)}}{e^{-\lambda f(x)}}, 1\right)$$

- If y is better than x, the next state is automatically accepted.
- Otherwise, it is accepted with a probability that depends on λ.
- If λ is high, the probability is small.
- When λ is small, it is easy to escape from local optima.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Notes

- In practice, it may be better to enumerate $\mathcal{F}$ (MH is asymptotic while $\mathcal{F}$ is finite).
- It is therefore usually used as a heuristic, where the value of λ is changed over time. For instance

$$\lambda_k = C \ln(1 + k), \; C > 0.$$

- The heuristic returns the best solution encountered during the process.