Advanced topics in Markov-chain Monte Carlo

Lecture 1:
Transition matrices - from the balance conditions to mixing
Part 2/2: Transition matrices

Werner Krauth

ICFP -Master Course Ecole Normale Supérieure, Paris, France

18 January 2023

References

- D. A. Levin, Y. Peres, E. L. Wilmer, "**Markov Chains and Mixing Times**" (American Mathematical Society, 2008) Second edition: `http://pages.uoregon.edu/dlevin/MARKOV/mcmt2e.pdf`
- M. Weber, '**Eigenvalues of non-reversible Markov chains - A case study**" ZIB report (2017) `http://nbn-resolving.de/urn:nbn:de:0297-zib-62191`
- A. Sinclair, M. Jerrum, **Approximate Counting, Uniform Generation and Rapidly Mixing Markov Chains** Information and Computation 82, 93-133 (1989) (We only need Lemma 3.3, and its proof) `https://people.eecs.berkeley.edu/~sinclair/approx.pdf`
- F. Chen, L. Lovasz, I. Pak, **Lifting Markov Chains to Speed up Mixing.** Proceedings of the 17th Annual ACM Symposium on Theory of Computing, 275 (1999) `http://www.math.ucla.edu/~pak/papers/stoc2.pdf`

Transition matrix

- Space of samples: sample space Ω
- Markov chain: Sequence of random variables $(X_0, X_1, \ldots)$ where X_0 represents the initial distribution and X_{t+1} depends on X_t through the transition matrix.
- $P_{ij} \geq 0$: Conditional probability to move to j if at i.
- $\sum_{j \in \Omega} P_{ij} = 1 \; \forall i \in \Omega$ (stochasticity condition).
- Commonly: made up of two parts $P_{ij} = \mathcal{A}_{ij}\mathcal{P}_{ij}$
  $\mathcal{P} \Leftrightarrow$ filter and $\mathcal{A} \Leftrightarrow$ *a priori* probability
  Examples: Metropolis filter, heatbath filter.
- Commonly: $P_{ii} \Leftrightarrow$ rejection probability.
  Advanced MCMC algorithms often have no rejections.
- $V = \{(i, j) | P_{ij} > 0\}$: set of vertices of a graph $G = (\Omega, V)$.

Irreducibility

- P irreducible $\Leftrightarrow$ any i can be reached from any j in a finite number of steps.
- This is equivalent to $(P^t)_{ij} > 0 \; \forall i,j$ for some t, which may depend on i and j.
- P connects not only configurations (samples) i, but also probability distributions:

$$\pi_i^{\{t\}} = \sum_{j \in \Omega} \pi_j^{\{t-1\}} P_{ji} \quad \Rightarrow \quad \pi_i^{\{t\}} = \sum_{j \in \Omega} \pi_j^{\{0\}} (P^t)_{ji} \quad \forall i \in \Omega.$$

- $\pi^{\{0\}}$: Initial probability (user-supplied). If concentrated on a single initial configuration: $\pi^{\{0\}}$ is a (Kronecker) δ-function.
- P irreducible $\Rightarrow$ unique *stationary distribution* π with

$$\pi_i = \sum_{j \in \Omega} \pi_j P_{ji} \quad \forall i \in \Omega.$$

- No guarantee that $\pi^{\{t\}} \to \pi$ for $t \to \infty$, though!

Probability flows

- Balance condition (again):

$$\pi_i = \sum_{j \in \Omega} \pi_j P_{ji} \quad \forall i \in \Omega.$$

- "flow" from j to i $\Leftrightarrow$ stationary probability $\times$ probability to move:

$$\mathcal{F}_{ji} \equiv \pi_j P_{ji} \quad \Leftrightarrow \quad \overbrace{\sum_{k \in \Omega} \mathcal{F}_{ik}}^{\text{flows exiting } i} \quad = \quad \overbrace{\sum_{j \in \Omega} \mathcal{F}_{ji}}^{\text{flows entering } i} \qquad \forall i \in \Omega,$$

(NB: stochasticity condition used $\sum_{k \in \Omega} P_{ik} = 1$).

Ergodic theorem 1/2

- Irreducible $P \Leftrightarrow$ unique π, but not necessarily $\pi^{\{t\}} \to \pi$ for $t \to \infty$.
- Nevertheless, ergodicity follows from irreducibility alone.
- This is (essentially) the strong law of large numbers:
  For $\mathcal{O}$ a real function on Ω, π probability distribution on Ω, irreducible Markov chain with stationary distribution π:

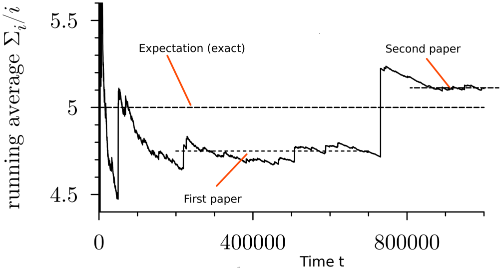
$$\langle \mathcal{O} \rangle := \sum_{i \in \Omega} \Omega_i \pi_i$$

then

$$P_{\pi^{\{0\}}} \left[\lim_{t \to \infty} \frac{1}{t} \sum_{i_t} \mathcal{O}(i_t) = \langle \mathcal{O} \rangle \right] = 1$$

- Ergodic theorem (for Markov chains):

$$P_{\pi\{0\}}\left[\lim_{t\to\infty}\frac{1}{t}\sum_{i_t}\mathcal{O}(i_t)=\langle\mathcal{O}\rangle\right]=1$$



Aperiodicity, convergence theorem

- Set of return times at configuration i: $\{t \geq 1 : (P^t)_{ii} > 0\}$
- Period: Greatest common divisor.
- $\{2, 4, 6, \ldots\} \Rightarrow$ period is 2
- $\{1000, 1001, 1002, \ldots\} \Rightarrow$ period is 1
- Period = 1: $\Leftrightarrow$ Markov chain is aperiodic
- For irreducible, aperiodic P: $P^t = (P^t)_{ij}$ is a positive matrix for some fixed t.
- For irreducible, aperiodic P: MCMC converges towards π from any starting distribution $\pi^{\{0\}}$.

Reversibility

- Reversible P satisfies the "detailed-balance" condition:

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \forall i, j \in \Omega.$$

- General P satisfies "global-balance" condition

$$\pi_i = \sum_{j \in \Omega} \pi_j P_{ji} \quad \forall i \in \Omega.$$

- DBC $\Rightarrow$ GBC (sum over j, use stochasticity: $\sum_{j \in \Omega} P_{ij} = 1$).
- DBC $\Leftrightarrow$ zero stationary net flow $\mathcal{F}_{ij} - \mathcal{F}_{ji}$ $\forall i, j \in \Omega$.
- Remember: GBC:

$$\overbrace{\sum_{k \in \Omega} \mathcal{F}_{ik}}^{\text{flows exiting } i} = \overbrace{\sum_{j \in \Omega} \mathcal{F}_{ji}}^{\text{flows entering } i} \qquad \forall i \in \Omega,$$

- DBC more restrictive, but far easier to check than GBC.

Spectrum of reversible transition matrix

- Reversible P:
$$\pi_i P_{ij} = \pi_j P_{ji} \quad \forall i, j \in \Omega.$$
- Reversible P: $A_{ij} = \pi_i^{1/2} P_{ij} \pi_j^{-1/2}$ is symmetric.
- Reversible P:
$$\sum_{j \in \Omega} \underbrace{\pi_i^{1/2} P_{ij} \pi_j^{-1/2}}_{A_{ij}} x_j = \lambda x_i \Leftrightarrow \sum_{j \in \Omega} P_{ij} \left[\pi_j^{-1/2} x_j \right] = \lambda \left[\pi_i^{-1/2} x_i \right].$$
- P and A have same eigenvalues.
- A symmetric: (Spectral theorem): All eigenvalues real, can expand on eigenvectors.
- Irreducible, aperiodic: Single eigenvalue with $\lambda = 1$, all others smaller in absolute value.

Classes for non-reversible transition matrix

Non-reversible P can be “unhappy” in different ways:

- P can be non-reversible, real eigenvalues, eigenvalues non-orthogonal.
- P can be non-reversible, real eigenvalues: Non-diagonalizable. (algebraic multiplicity $\neq$ geometric multiplicity).
- P can be non-reversible, pairs of complex eigenvalues.
- Most common case: Complex eigenvalues.
- For simple examples, see Weber (2017)

Total variation distance, mixing time

- Total variation distance:

$$||\pi^{\{t\}} - \pi||_{\text{TV}} = \max_{A\subset\Omega} |\pi^{\{t\}}(A) - \pi(A)| = \frac{1}{2}\sum_{i\in\Omega} |\pi_i^{\{t\}} - \pi_i|.$$

- (Above) first eq.: definition; second eq.: (tiny) theorem
- Distance:

$$d(t) = \max_{\pi^{\{0\}}} ||\pi^{\{t\}}(\pi^{\{0\}}) - \pi||_{\text{TV}}$$

- Mixing time:

$$t_{\text{mix}}(\epsilon) = \min\{t : d(t) \leq \epsilon\}$$

- Usually $\epsilon = 1/4$ is taken (arbitrary, must be smaller than $\frac{1}{2}$): $t_{\text{mix}} = t_{\text{mix}}(1/4)$

Diameter bounds, conductance

- Graph diameter L: minimum number of moves to travel between any $i, j \in \Omega$
- Diameter bound: or any $\epsilon < 1/2$, trivially satisfies

$$t_{\text{mix}} \geq L/2.$$

- Conductance (bottleneck ratio):

$$\Phi \equiv \min_{S \subset \Omega, \pi_S \leq \frac{1}{2}} \frac{\mathcal{F}_{S \to \overline{S}}}{\pi_S} = \min_{S \subset \Omega, \pi_S \leq \frac{1}{2}} \frac{\sum_{i \in S, j \notin S} \pi_i P_{ij}}{\pi_S}.$$

Total variation distance, mixing time (reminder)

- Total variation distance:

$$||\pi^{\{t\}} - \pi||_{\mathrm{TV}} = \max_{A\subset\Omega} |\pi^{\{t\}}(A) - \pi(A)| = \frac{1}{2}\sum_{i\in\Omega} |\pi_i^{\{t\}} - \pi_i|.$$

- (Above) first eq.: definition; second eq.: (tiny) theorem
- Distance:

$$d(t) = \max_{\pi^{\{0\}}} ||\pi^{\{t\}}(\pi^{\{0\}}) - \pi||_{\mathrm{TV}}$$

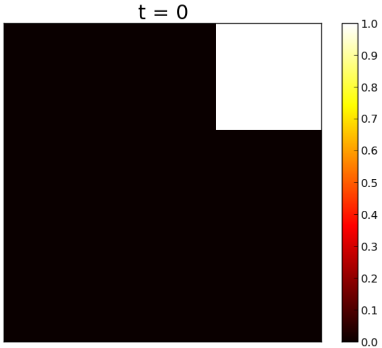
- Mixing time:

$$t_{\mathrm{mix}}(\epsilon) = \min\{t : d(t) \leq \epsilon\}$$

- Usually $\epsilon = 1/4$ is taken (arbitrary, must be smaller than $\frac{1}{2}$): $t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4)$

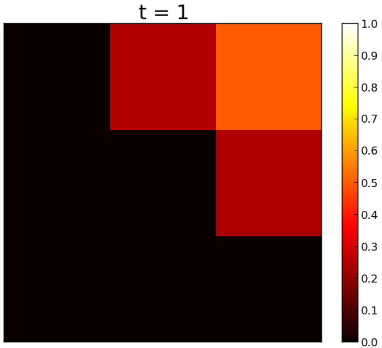
Mixing (reminder)

- Distribution $\pi^{t=0}$ (starting from upper right)



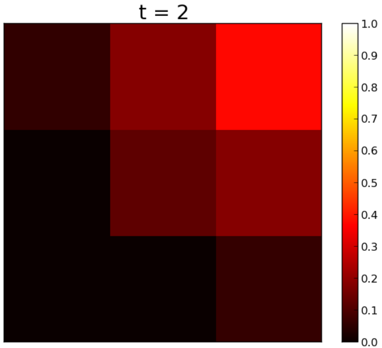
Mixing

- Distribution $\pi^{t=1}$ (starting from upper right)



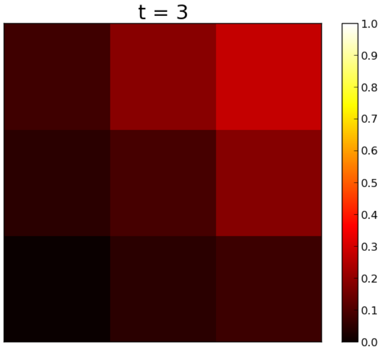
Mixing

- Distribution $\pi^{t=2}$ (starting from upper right)



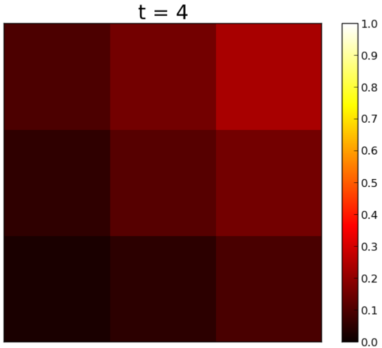
Mixing

- Distribution $\pi^{t=3}$ (starting from upper right)



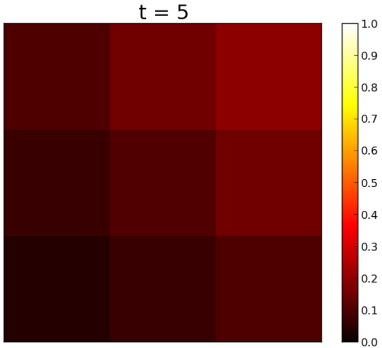
Mixing

- Distribution $\pi^{t=4}$ (starting from upper right)



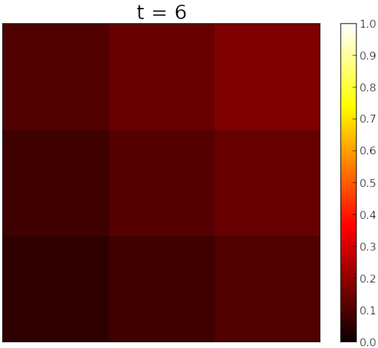
Mixing

- Distribution $\pi^{t=5}$ (starting from upper right)



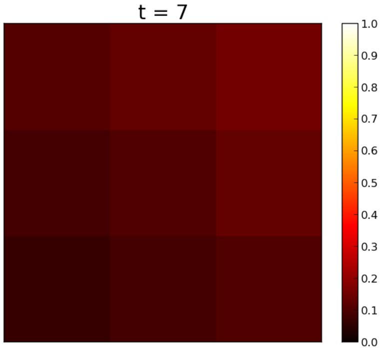
Mixing

- Distribution $\pi^{t=6}$ (starting from upper right)

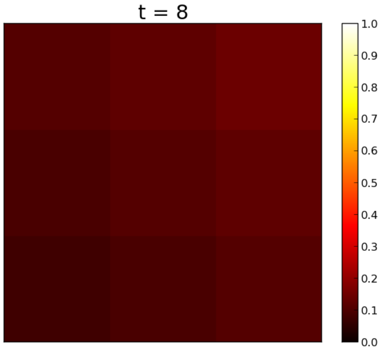


Mixing

- Distribution $\pi^{t=7}$ (starting from upper right)

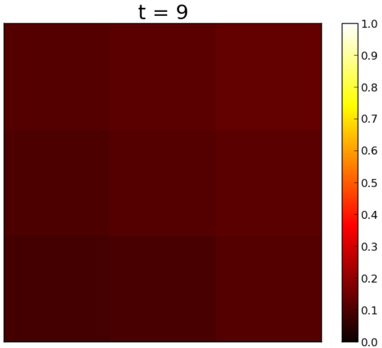


- Distribution $\pi^{t=8}$ (starting from upper right)



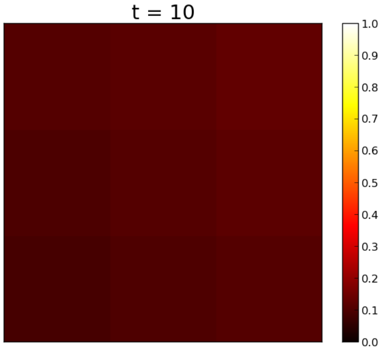
Mixing

- Distribution $\pi^{t=9}$ (starting from upper right)



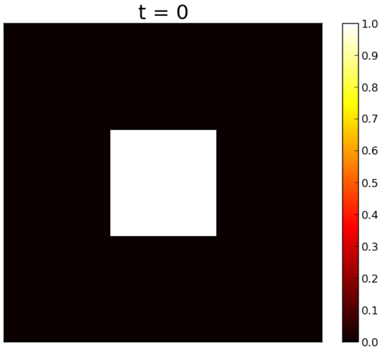
Mixing

- Distribution $\pi^{t=10}$ (starting from upper right)



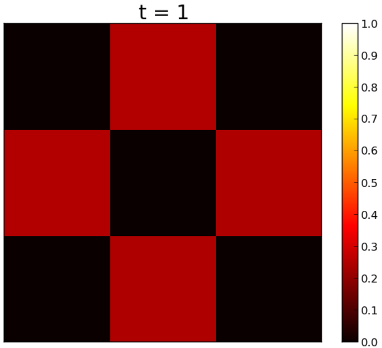
Mixing

- Distribution $\pi^{t=0}$ (starting from center)



Mixing

- Distribution $\pi^{t=1}$ (starting from center)



Diameter bounds, conductance

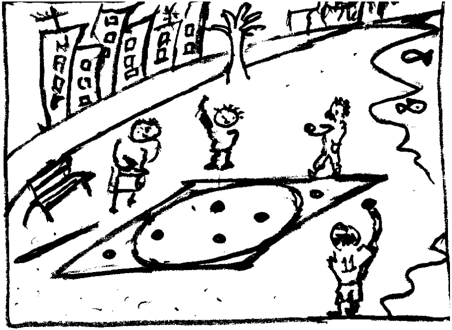
- Graph diameter L: minimum number of moves to travel between any $i, j \in \Omega$.
- NB: $L = 4$ for 3×3 pebble game.
- Diameter bound: for any $\epsilon < 1/2$, trivially satisfies

$$t_{\text{mix}} \geq L/2.$$

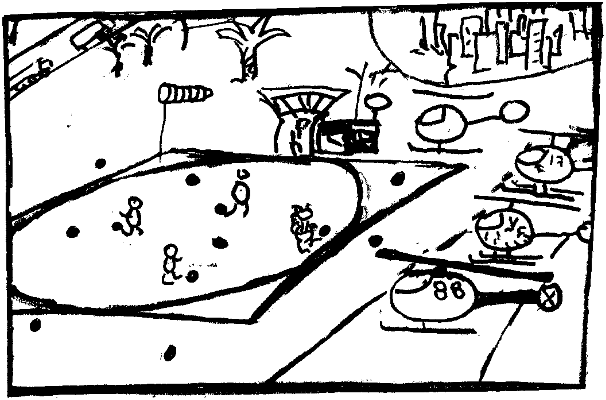
- Conductance (bottleneck ratio):

$$\Phi \equiv \min_{S \subset \Omega, \pi_S \leq \frac{1}{2}} \frac{\mathcal{F}_{S \to \overline{S}}}{\pi_S} = \min_{S \subset \Omega, \pi_S \leq \frac{1}{2}} \frac{\sum_{i \in S, j \notin S} \pi_i P_{ij}}{\pi_S}.$$

Direct Sampling

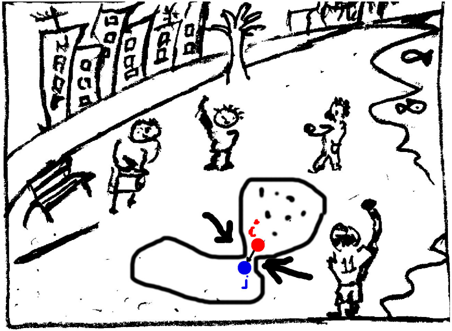


Markov-chain sampling



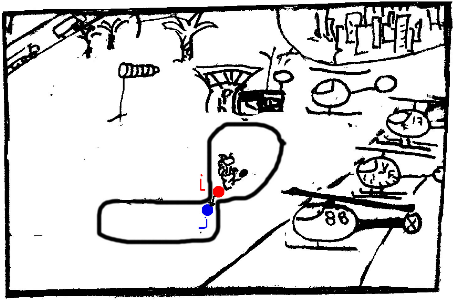
NB: ... slower than direct sampling

Direct sampling with bottleneck



NB: ... reaches a boundary site $i \in S$ with probability π_i/π_S

Direct sampling with bottleneck



NB: ... reaches a boundary site $i \in S$ less than with π_i/π_S

Conductance and correlations

Remember:

$$\Phi \equiv \min_{S\subset\Omega,\pi_S\leq\frac{1}{2}} \frac{\mathcal{F}_{S\to\overline{S}}}{\pi_S} = \min_{S\subset\Omega,\pi_S\leq\frac{1}{2}} \frac{\sum_{i\in S, j\notin S} \pi_i P_{ij}}{\pi_S}.$$

- Reversible Markov chains:

$$\frac{1}{\Phi} \leq \tau_{\text{corr}} \leq \frac{8}{\Phi^2}$$

  (second relation see Sinclair & Jerrum, Lemma (3.3) (p 15-17))
- Arbitrary Markov chain (see Chen et al):

$$\frac{1}{4\Phi} \leq \mathcal{A} \leq \frac{20}{\Phi^2},$$

  (set time: Expectation of $\max_S (t_S \times \pi_S)$ from equilibrium)

NB: One bottleneck, not many. Lower *and* upper bound.
NNB: $\mathcal{A}$ is not the mixing time as we have defined it (see Chen et al. (1999)).

Conductance and mixing

$$\Phi \equiv \min_{S\subset\Omega,\pi_S\leq\frac{1}{2}} \frac{\mathcal{F}_{S\to\overline{S}}}{\pi_S} = \min_{S\subset\Omega,\pi_S\leq\frac{1}{2}} \frac{\sum_{i\in S,j\notin S}\pi_i P_{ij}}{\pi_S}.$$

- Mixing-time bounds:

$$\frac{\text{const}}{\Phi} \leq t_{\text{mix}} \leq \frac{\text{const}'}{\Phi^2}\log\left(1/\pi_0\right)$$

  const and const′ depend on whether reversible or non-reversible. π_0: smallest weight (see Chen et al 1999).

NB: One bottleneck, not many. Lower *and* upper bound.
NNB: Conductance: more general than transition matrices