Markov-chain Monte Carlo: A modern primer

Lecture 1: Fundamentals
Part 1/2: MCMC—from balance to lifting

Werner Krauth
Département de physique, Ecole normale supérieure
Paris, France

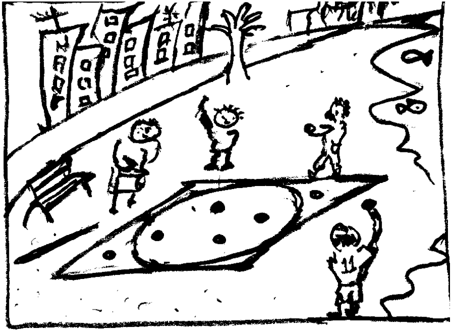
CANES Masterclass
Faculty of Natural & Mathematical Sciences, King's College London
London, Great Britain

28 February 2022

References

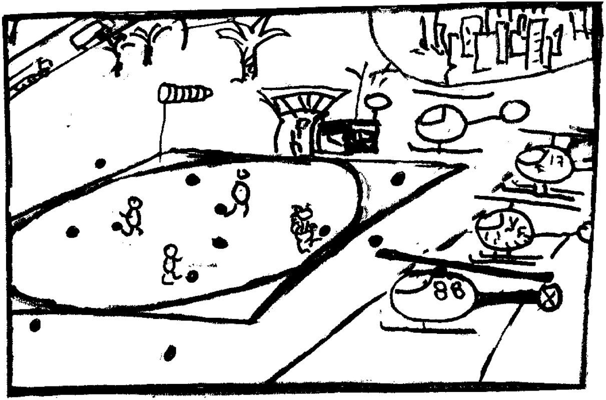
- D. A. Levin, Y. Peres, E. L. Wilmer, "**Markov Chains and Mixing Times**" (American Mathematical Society, 2008) Second edition: `http://pages.uoregon.edu/dlevin/MARKOV/mcmt2e.pdf`
- M. Weber, '**Eigenvalues of non-reversible Markov chains - A case study**" ZIB report (2017) `http://nbn-resolving.de/urn:nbn:de:0297-zib-62191`
- 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`

Direct sampling (1/1)



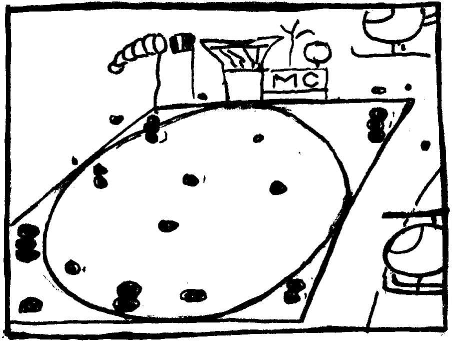
- Distribution π = uniform in square

Markov-chain sampling (1/2)



- Distribution $\pi =$ uniform in heliport square

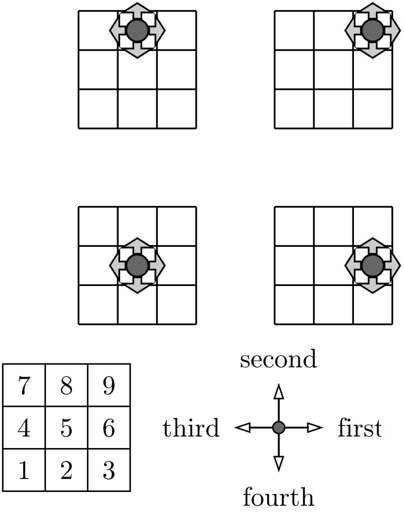
Markov-chain sampling (2/2)



- Metropolis et al. (1953).

Transition matrix (1/4)

- discretized version of heliport game



Transition matrix (2/4)

- Transition-matrix element P_{ij}: probability to move to j if at i:

$$P = \begin{bmatrix}
\boxed{\frac{1}{2}} & \frac{1}{4} & \cdot & \frac{1}{4} & \cdot & \cdot & \cdot & \cdot & \cdot \\
\frac{1}{4} & \boxed{\frac{1}{4}} & \frac{1}{4} & \cdot & \frac{1}{4} & \cdot & \cdot & \cdot & \cdot \\
\cdot & \frac{1}{4} & \boxed{\frac{1}{2}} & \cdot & \cdot & \frac{1}{4} & \cdot & \cdot & \cdot \\
\frac{1}{4} & \cdot & \cdot & \boxed{\frac{1}{4}} & \frac{1}{4} & \cdot & \frac{1}{4} & \cdot & \cdot \\
\cdot & \frac{1}{4} & \cdot & \frac{1}{4} & \boxed{0} & \frac{1}{4} & \cdot & \frac{1}{4} & \cdot \\
\cdot & \cdot & \frac{1}{4} & \cdot & \frac{1}{4} & \boxed{\frac{1}{4}} & \cdot & \cdot & \frac{1}{4} \\
\cdot & \cdot & \cdot & \frac{1}{4} & \cdot & \cdot & \boxed{\frac{1}{2}} & \frac{1}{4} & \\
\cdot & \cdot & \cdot & \cdot & \frac{1}{4} & \cdot & \frac{1}{4} & \boxed{\frac{1}{4}} & \frac{1}{4} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \frac{1}{4} & \cdot & \frac{1}{4} & \boxed{\frac{1}{2}}
\end{bmatrix}.$$

- All P are stochastic matrices: $\sum_j P_{ij} = 1$
- The pebble-game P is doubly stochastic.

Transition matrix (3/4)

- Initial distribution (NB: row vector)

$$\pi^{\{t=0\}} = \{0, \ldots, 0, 1\}.$$

- Distribution at time $t+1$ (short hand: $\pi^{\{t+1\}} = \pi^{\{t\}} P$)

$$\pi_i^{\{t+1\}} = \sum_{j=1}^{9} \pi_j^{\{t\}} P_{j \to i}$$

NB: P connects samples x_{t+1} to x_t, but also $\pi^{\{t+1\}}$ to $\pi^{\{t\}}$

- Left eigenvectors, eigenvalues

$$\{\pi_1^{\{t\}}, \ldots, \pi_9^{\{t\}}\} = \underbrace{\{\tfrac{1}{9}, \ldots, \tfrac{1}{9}\}}_{\substack{\text{first left eigenvector} \\ \text{eigenvalue } \lambda_1 = 1}} + \alpha_2 (0.75)^t \underbrace{\{-0.21, \ldots, 0.21\}}_{\substack{\text{second left eigenvector} \\ \text{eigenvalue } \lambda_2 = 0.75}} + \ldots$$

Transition matrix (4/4)

- Heliport square $\rightarrow$ sample space Ω.
- Players $\rightarrow$ 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 P.
- Four-arrow star$\rightarrow$ split matrix: $P_{ij} = \mathcal{A}_{ij}\mathcal{P}_{ij}$ for $i \neq j$
  $\mathcal{A} \Leftrightarrow$ *a priori* probability; $\mathcal{P} \Leftrightarrow$ filter
  Examples: Metropolis filter, heatbath filter.
- Pebble piles $\rightarrow$ $P_{ii} \Leftrightarrow$ (filter) rejection probability.
  NB: Modern MCMC algorithms often have no rejections.
- Eigensystem analysis $\not\rightarrow$ Not generally possible for "non-reversible" transition matrices.

Irreducibility

- P irreducible $\Leftrightarrow$ any i can be reached from any j.
- $\pi^{\{0\}}$: 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

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

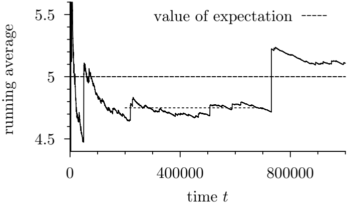
- No guarantee that $\pi^{\{t\}} \to \pi$ for $t \to \infty$, for any $\pi^{\{0\}}$

Ergodic theorem

- P irreducible $\Rightarrow \pi$ unique, but maybe $\pi^{\{t\}} \not\to \pi$ for $t \to \infty$.
- P irreducible $\Rightarrow$ Ergodic theorem ($\mathbb{E}(\mathcal{O}) := \sum_{i \in \Omega} \mathcal{O}_i \pi_i$):

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

(Strong law of large numbers for a single running average)



- Diagnostic tool if $\mathbb{E}(\mathcal{O})$ is known.

Probability flows

- Uniqueness of $\pi \Rightarrow$ balance condition on P:

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

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

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

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

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

Aperiodicity, convergence theorem

- Set of return times at configuration i: $\{t \geq 1 : (P^t)_{ii} > 0\}$
- $\{2, 4, 6, \ldots\} \Rightarrow$ 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: exponential convergence towards π from any starting distribution $\pi^{\{0\}}$.

Reversibility

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

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

- General P satisfies "global-balance" condition

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

- Detailed balance implies global balance.
- Global balance:

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

- 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:

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

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

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

- Mixing time:

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

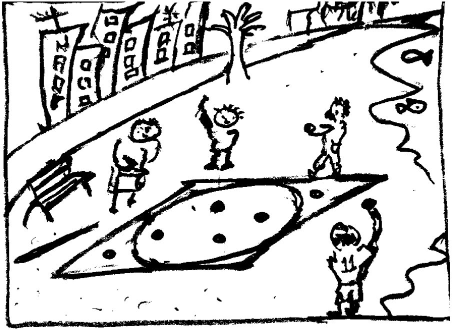
- Usually $\epsilon = 1/4$ is taken (arbitrary, must be smaller than $\frac{1}{2}$): $t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4)$

Diameter bound

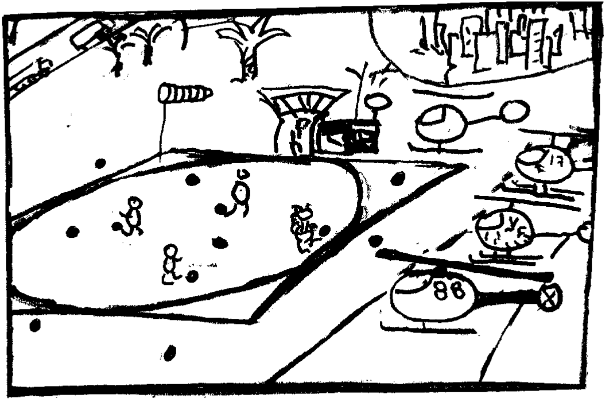
- Graph diameter L: minimum number of moves to travel between any $i, j \in \Omega$.
  NB: $L = 4$ for 3×3 pebble game.
- Diameter bound: for any $\epsilon < 1/2$, trivially satisfies

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

Conductance (bottleneck ratio) (1/6)

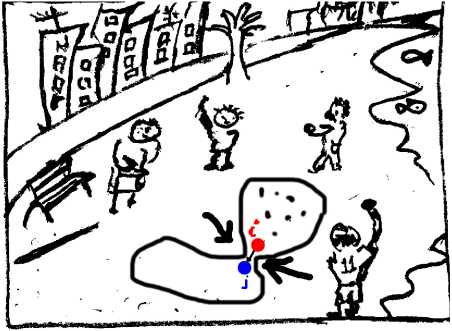


Conductance (bottleneck ratio) (2/6)



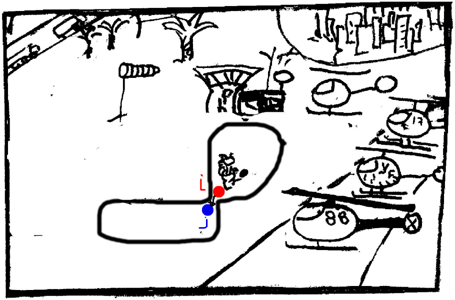
NB: ... this is less efficient than direct sampling

Conductance (bottleneck ratio) (3/6)



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

Conductance (bottleneck ratio) (4/6)



NB: ... reaches a boundary site $i \in \partial S$ with probability $\leq \pi_i / \pi_S$

Conductance (bottleneck ratio) (5/6)

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

- Reversible Markov chains:

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

("$\leq$": Sinclair & Jerrum (1986), Lemma (3.3))

- Arbitrary Markov chain (see Chen et al. (1999)):

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

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

NB: One bottleneck, not many. Lower *and* upper bound.

Conductance (bottleneck ratio) (6/6)

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

- 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).

Lifting (Chen et al (1999)) (1/2)

- Markov chain $\Pi \Leftrightarrow$ Lifted Markov chain $\widehat{\Pi}$
- $\Omega \ni v$ (sample space) $\Leftrightarrow \widehat{\Omega} \ni i$ (lifted sample space)
- P (transition matrix) $\Leftrightarrow \widehat{P}$ (lifted transition matrix)
- π_v (stationary probability) $\Leftrightarrow \hat{\pi}_i$
- **Condition 1**: sample space is copied ("lifted"), π preserved

$$\pi_v = \hat{\pi}\left[f^{-1}(v)\right] = \sum_{i \in f^{-1}(v)} \hat{\pi}_i,$$

- **Condition 2**: flows are preserved

$$\underbrace{\pi_v P_{vu}}_{\text{collapsed flow}} = \sum_{i \in f^{-1}(v), j \in f^{-1}(u)} \overbrace{\hat{\pi}_i \widehat{P}_{ij}}^{\text{lifted flow}} .$$

- Usually: $\widehat{\Omega} = \Omega \times \mathcal{L}$, with $\mathcal{L}$ a set of lifting variables σ

Lifting (Chen et al (1999)) (2/2)

- Required: Mapping from $\widehat{\Omega}$ (lifted sample space) to Ω that preserves stationary probability distribution.
- Required: Lifted transition matrix $\widehat{P}$ that preserves flow.
- Optional: $\widehat{\Omega} = \Omega \times \mathcal{L}$ (with $\mathcal{L}$: set of lifting variables).
- Optional:

$$\frac{\hat{\pi}(u,\sigma)}{\pi(u)} = \frac{\hat{\pi}(v,\sigma)}{\pi(v)} \quad \forall\, u, v \in \Omega; \forall\, \sigma \in \mathcal{L}. \qquad (1)$$

- There are many liftings $\widehat{P}$ for a given lifted sample space $\widehat{\Omega}$.
- Liftings are popular for transfering parts of the moves into the sample space.
- Lifting do not increase conductance.