Probabilistic Graphical Models

Part 4: Approximate Inference

Manfred Jaeger

Aalborg University

Basic Sampling

Problem Structure

Input: Evidence $\boldsymbol{Y} = \boldsymbol{y}$, random variable X, value $x \in Val(X)$.
Output: Approximation q for $p := P(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$.

Absolute/Relative Error

For $p, q \in [0, 1]$: q is approximation for p with *absolute error* $\leq \epsilon$, if

$$\mid p - q \mid \leq \epsilon, \text{ i.e. } q \in [p - \epsilon, p + \epsilon].$$

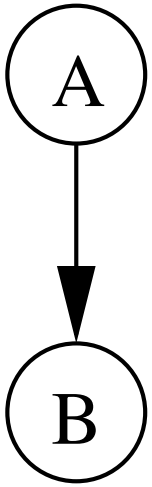
q is approximation for p with *relative error* $\leq \epsilon$, if

$$\mid 1 - q/p \mid \leq \epsilon, \text{ i.e. } q \in [p(1 - \epsilon), p(1 + \epsilon)].$$

- Relative error is not symmetric in p and q and not invariant under the transition $p \to (1 - p)$, $q \to (1 - q)$. Use with care!
- When q_1, q_2 are approximations for p_1, p_2 with absolute error $\leq \epsilon$, then no error bounds follow for q_1/q_2 as an approximation for p_1/p_2.
- When q_1, q_2 are approximations for p_1, p_2 with relative error $\leq \epsilon$, then q_1/q_2 approximates p_1/p_2 with relative error $\leq (2\epsilon)/(1 + \epsilon)$.

Observation: can use Bayesian network as random generator that produces full instantiations $\boldsymbol{X} = \boldsymbol{x}$ according to distribution $P(\boldsymbol{X})$.

Example:



	A	
	t	f
	.2	.8

	B	
A	t	f
t	.7	.3
f	.4	.6

- Generate random numbers r_1, r_2 uniformly from [0,1].
- Set $A = t$ if $r_1 \leq .2$ and $A = f$ else.
- Depending on the value of A and r_2 set B to t or f.

Generation of one random instantiation: linear in size of network.

Given a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ of complete instantiations generated (independently) by the sampling algorithm, approximate $P(\boldsymbol{Y} = \boldsymbol{y})$ as

$$q^* := \frac{1}{N} \left| \{ i \in 1, \ldots, N \mid \boldsymbol{Y} = \boldsymbol{y} \text{ in } \boldsymbol{x}_i \} \right|$$

Similarly, the sample provides an estimate for $P(X = x, \boldsymbol{Y} = \boldsymbol{y})$.

Let $\epsilon, \delta > 0$, $p = P(\boldsymbol{Y} = \boldsymbol{y})$. From Chebyshev's inequality:

$$N \geq \frac{1}{\epsilon^2 \delta} \rightarrow P_N(\mid q^* - p \mid \leq \epsilon) \geq 1 - \delta.$$

$$N \geq \frac{1-p}{p\epsilon^2 \delta} \rightarrow P_N(\mid 1 - q^*/p \mid \leq \epsilon) \geq 1 - \delta.$$

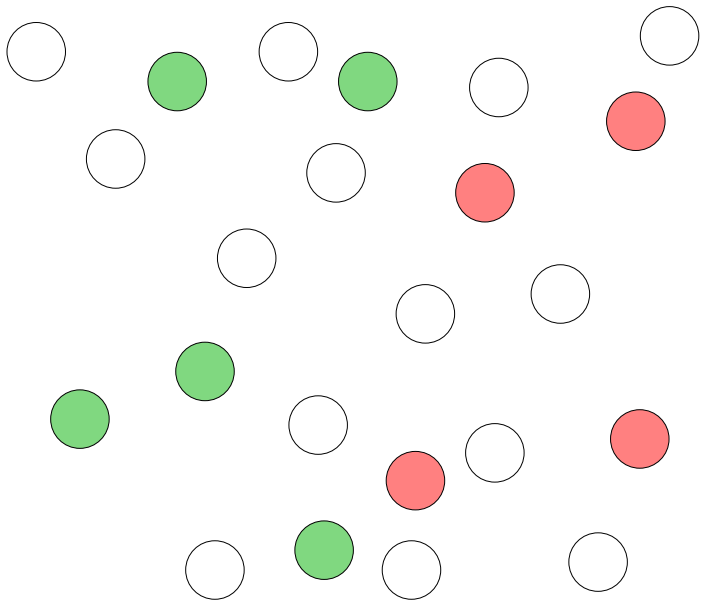
P_N: probability distribution on $Val(\boldsymbol{X})^N$ induced by independently sampling N instances.
Chebyshev's inequality: for any real-valued random variable X:

$$P(\mid X - E[X] \mid \geq \epsilon) \leq \frac{1}{\epsilon^2} Var[X]$$

- Sharper bounds can be obtained from Hoeffding and Chernoff inequalities

Problems for approximation with relative error bound:

1) N depends on p – but p is not known!
2) N increases as $1/p$.



Sample with

not $\boldsymbol{Y} = \boldsymbol{y}$

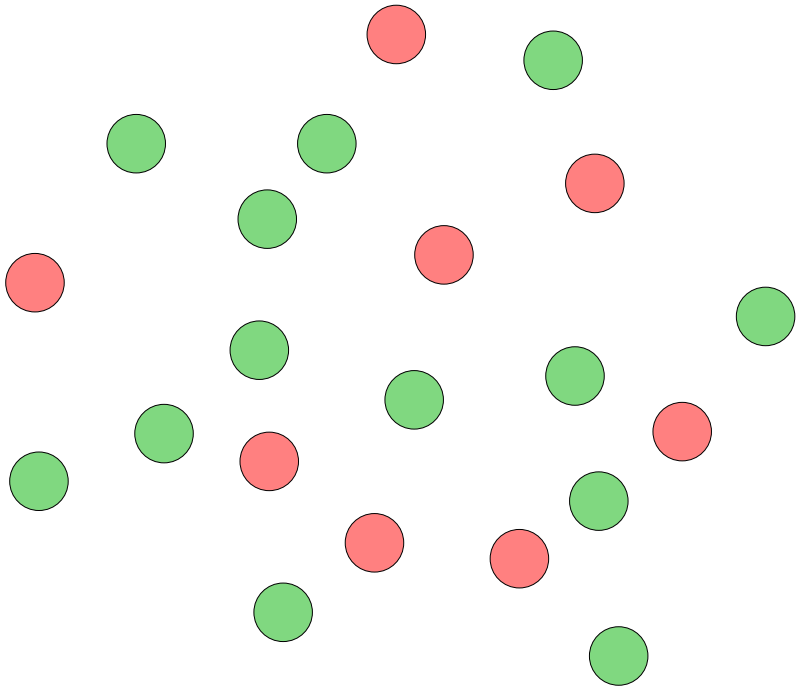
$\boldsymbol{Y} = \boldsymbol{y}, X \neq x$

$\boldsymbol{Y} = \boldsymbol{y}, X = x$

Approximation for $P(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$: $\dfrac{\#\,\bigcirc}{\#\,\bigcirc \cup \bigcirc}$

Problem of forward sampling: samples with $\boldsymbol{Y} \neq \boldsymbol{y}$ are useless!

Goal: find algorithm that samples according to $P(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y})$:



MCMC Sampling

Principle: obtain new sample from previous sample by randomly changing the value of only one selected variable.

```
Procedure Gibbs sampling
```
$\boldsymbol{x}_0 = (x_{0,1}, \ldots, x_{0,n}) :=$ `arbitrary instantiation of` $\boldsymbol{X}$ `with` $\boldsymbol{Y} = \boldsymbol{y}$
$t := 1$
`repeat forever`
  `choose` $X_k \in \boldsymbol{X} \setminus \boldsymbol{Y}$
  `generate randomly` $x_{t,k}$ `according to` $P(X_k \mid \boldsymbol{X} \setminus X_k = \boldsymbol{x}_{t-1} \setminus x_{t-1,k})$
  `set` $x_{t,j} := x_{t-1,j}$ `for` $j \neq k$.
  $t := t + 1$

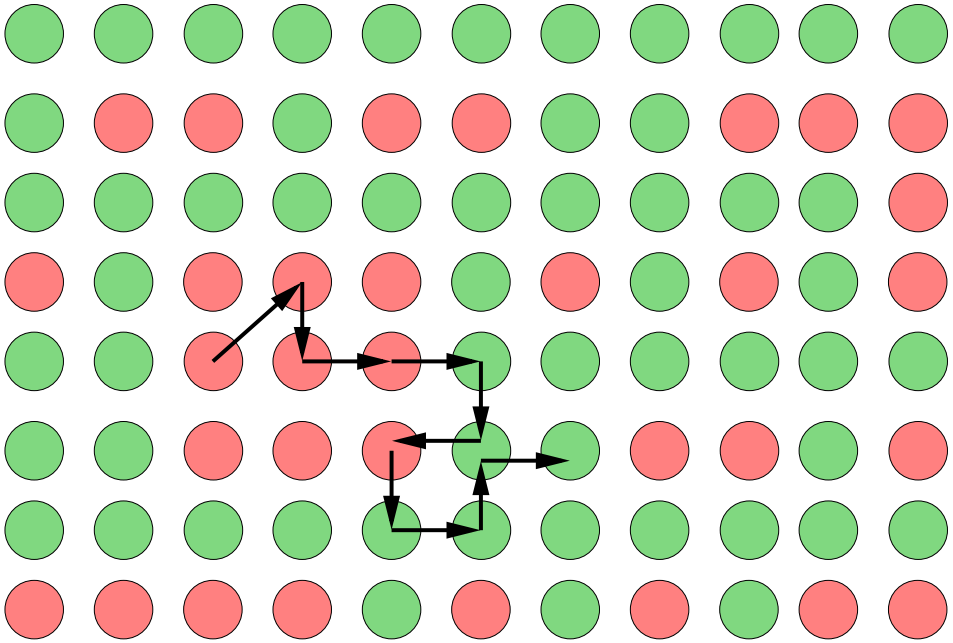
$$
\begin{aligned}
&P(X_k = x_k \mid \boldsymbol{X} \setminus X_k = \boldsymbol{x}_{t-1} \setminus x_{t-1,k}) \\
&\quad \approx P(X_k = x_k, \boldsymbol{X} \setminus X_k = \boldsymbol{x}_{t-1} \setminus x_{t-1,k}) \\
&\quad \approx P(X_k = x_k \mid \mathrm{Pa}(X_k) = (\boldsymbol{x}_{t-1} \restriction \mathrm{Pa}(X_k))) \cdot \\
&\qquad \prod_{i : X_k \in \mathrm{Pa}(X_i)} P(X_i = x_{t-1,i} \mid \mathrm{Pa}(X_i) \setminus X_k = (\boldsymbol{x}_{t-1} \restriction \mathrm{Pa}(X_i) \setminus X_k), X_k) \quad (*)
\end{aligned}
$$

where $\approx$ means: equals up to a constant that does not depend on X_k.

To sample $X_{t,k}$:

- evaluate $(*)$ for all $x_k \in Val(X_k)$
- normalize to obtain $P(X_k \mid \boldsymbol{X} \setminus X_k = \boldsymbol{x}_t \setminus x_{t,k})$
- sample value $x_{t,k}$ according to the resulting distribution

The process of Gibbs sampling can be understood as a *random walk* in the space of all instantiations with $\boldsymbol{Y} = \boldsymbol{y}$:



Reachable in one step: instantiations that differ from current one by value assignment to at most one variable (assume randomized choice of variable X_k).

Goal: want to sample instantiations $\boldsymbol{x} \in Val(\boldsymbol{X})$ from a distribution $\pi(\boldsymbol{X})$.
(e.g.: $\pi(\boldsymbol{X}) = P(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y})$)

```
Procedure Metropolis Hastings sampling
```
$\boldsymbol{x}_0 = (x_{0,1}, \ldots, x_{0,n}) :=$ `arbitrary instantiation of` $\boldsymbol{X}$ `with` $\pi(\boldsymbol{X} = \boldsymbol{x}_0) > 0$
$t := 1$
`repeat forever`
`generate` $\boldsymbol{x}_{cand}$ `according to a` proposal distribution $Q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1})$
`compute` acceptance probability $p_{acc} = min\{1, \frac{\pi(\boldsymbol{x}_{cand})Q(\boldsymbol{X}_t = \boldsymbol{x}_{t-1} \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{cand})}{\pi(\boldsymbol{x}_{t-1})Q(\boldsymbol{X}_t = \boldsymbol{x}_{cand} \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1})}\}$
`set` $\boldsymbol{x}_t = \begin{cases} \boldsymbol{x}_{cand} & \texttt{with probability } p_{acc} \\ \boldsymbol{x}_{t-1} & \texttt{with probability } 1 - p_{acc} \end{cases}$
$t := t + 1$

Independent sampling from π

Suppose we can sample from $\pi(\boldsymbol{X})$. Then M-H sampling with

$$Q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}) = \pi(\boldsymbol{X})$$

produces a series of independent samples from $\pi(\boldsymbol{X})$.

Gibbs sampling

Let $\pi(\boldsymbol{X}) = P(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y})$, and $Q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1})$ be defined by resampling a single, randomly selected variable X_k. Then

$$\frac{\pi(\boldsymbol{x}_{cand})}{\pi(\boldsymbol{x}_{t-1})} = \frac{Q(\boldsymbol{X}_t = \boldsymbol{x}_{cand} \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1})}{Q(\boldsymbol{X}_t = \boldsymbol{x}_{t-1} \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{cand})},$$

$p_{acc} = 1$, and M-H becomes Gibbs sampling.

- Need to design $Q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1})$ so that
  - can effectively sample from $Q()$
  - $$\frac{\pi(\boldsymbol{x}_{cand})}{\pi(\boldsymbol{x}_{t-1})} \approx \frac{Q(\boldsymbol{X}_t = \boldsymbol{x}_{cand} \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1})}{Q(\boldsymbol{X}_t = \boldsymbol{x}_{t-1} \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{cand})}$$

    so that acceptance probabilities are not too small
- Need to be able to evaluate ratios $\frac{\pi(\boldsymbol{x}_{cand})}{\pi(\boldsymbol{x}_{t-1})}$

Two conditions on the Markov Chain $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots$ defined by M-H sampling:

Irreducibility
For all $\boldsymbol{x}, \boldsymbol{x}' \in Val(\boldsymbol{X})$ with $\pi(\boldsymbol{x}) > 0$, $\pi(\boldsymbol{x}') > 0$: there exists $k > 0$ with

$$P(\boldsymbol{X}_k = \boldsymbol{x}' \mid \boldsymbol{X}_0 = \boldsymbol{x}) > 0$$

Aperiodicity
There exists $\boldsymbol{x}$ with $\pi(\boldsymbol{x}) > 0$, so that for all sufficiently large k:

$$P(\boldsymbol{X}_k = \boldsymbol{x} \mid \boldsymbol{X}_0 = \boldsymbol{x}) > 0$$

These are essentially conditions for the proposal distribution!

Convergence

If Irreducibility and Aperiodicity hold, then the sampling distribution of $\boldsymbol{X}_t$ converges to $\pi(\boldsymbol{X})$.

Estimate probability $\pi(A)$ $A \subseteq Val(\boldsymbol{X})$: (special case: $\pi(\boldsymbol{X}) = P(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y})$, $A = \{X = x\}$)

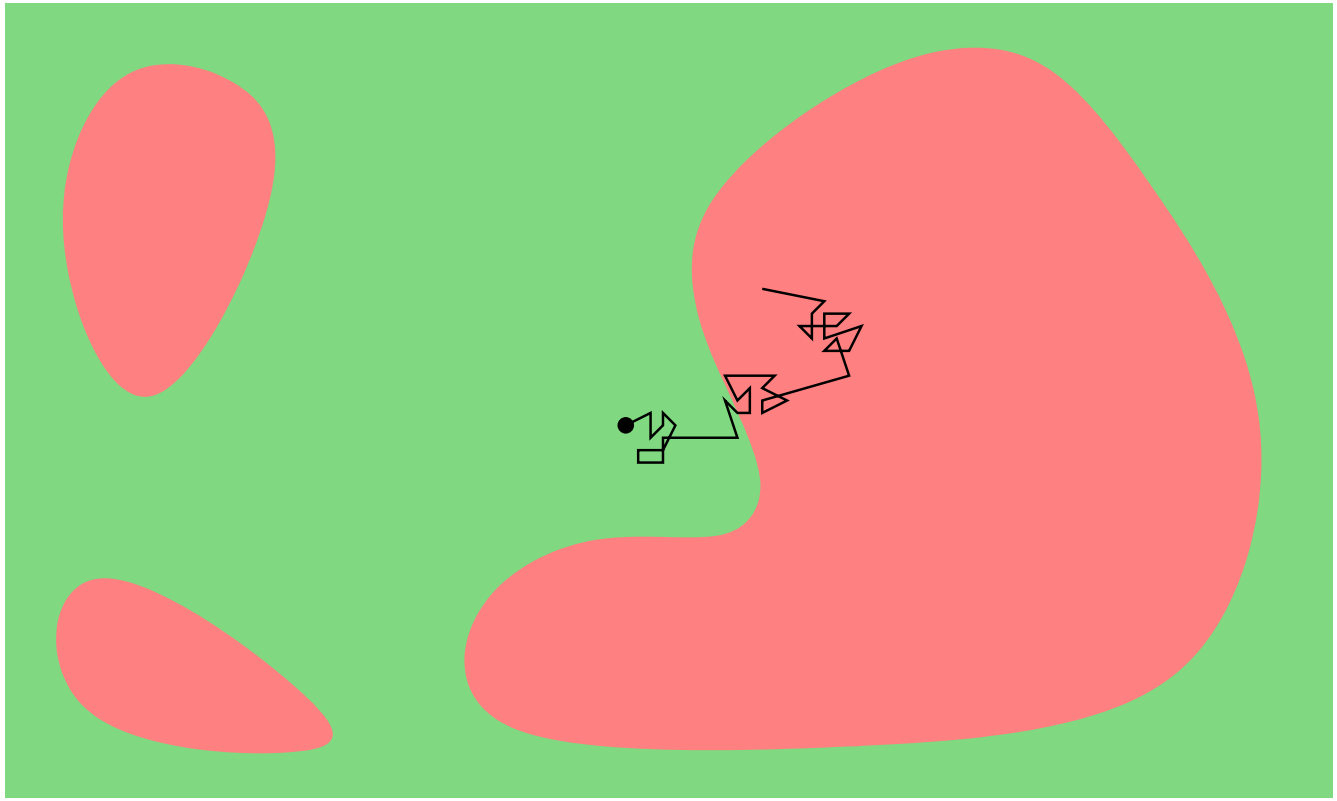
1. Find initial $\boldsymbol{x}$ with $\pi(\boldsymbol{x}) > 0$
2. **Burn in:** Perform k steps of M-H sampling
3. Sample values $\boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+2}, \ldots, \boldsymbol{x}_{k+N}$
4. Estimate:

$$\hat{\pi}(A) = \frac{1}{N} \left| \{ i \mid \boldsymbol{x}_{k+i} \in A \} \right|$$

Issues:

- How large needs k to be, so that $P(\boldsymbol{X}_k)$ is close to $\pi(\boldsymbol{X})$?
- The $\boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+2}, \ldots, \boldsymbol{x}_{k+N}$ are not independent. Accuracy of estimate $\hat{\pi}(A)$ very difficult to determine

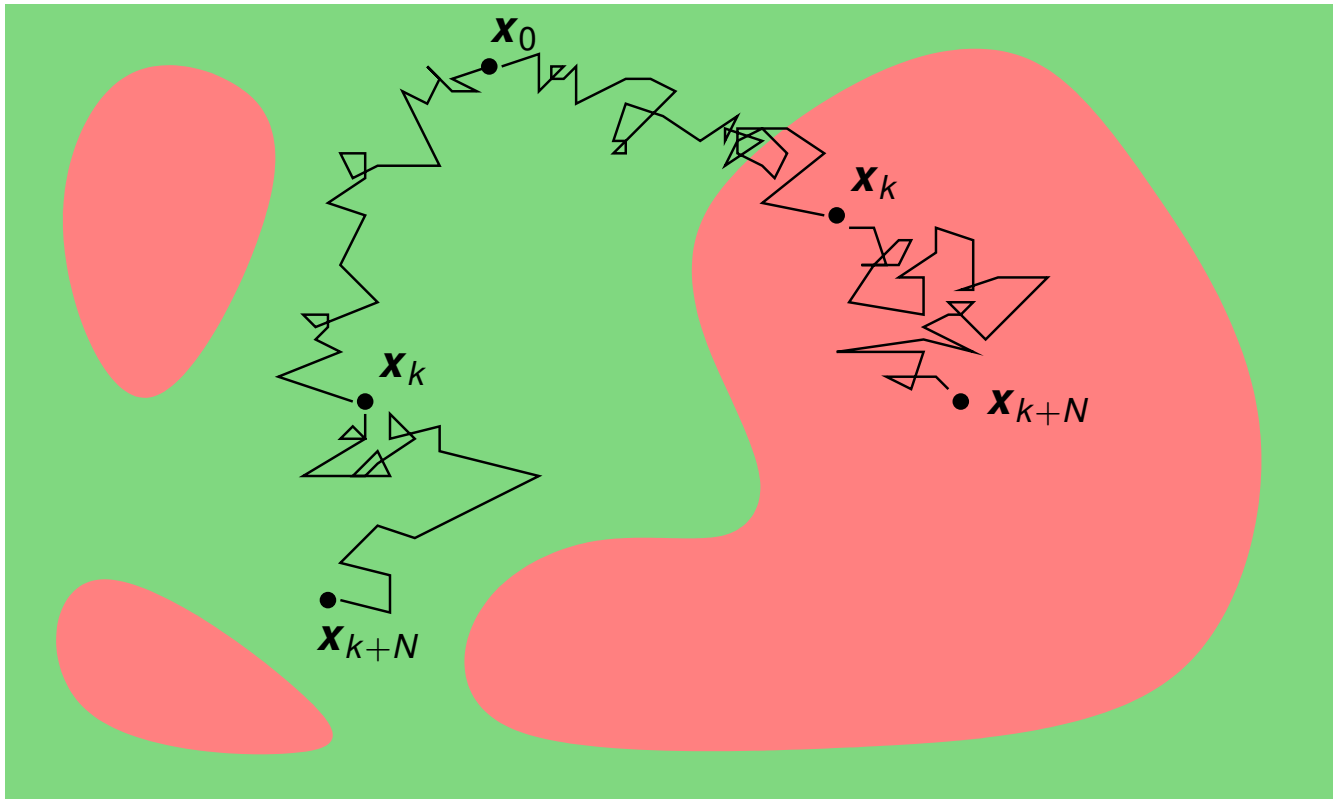
Markov chain $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ as random walk in $Val(\boldsymbol{X})$:



red: states $\boldsymbol{x}$ with $\boldsymbol{x} \in A$
green: states $\boldsymbol{x}$ with $\boldsymbol{x} \notin A$

$P(\boldsymbol{X}_{k+i})$ close to $\pi(\boldsymbol{X})$: probability that $\boldsymbol{x}_{k+i}$ is in the red region is close to $\pi(A)$.

This does not guarantee that $1/N\,|\{i \mid \boldsymbol{x}_{k+i} \in A\}|$ yields a good approximation for $\pi(A)$!



In practice, one tries to counteract these difficulties by running multiple Markov chains: