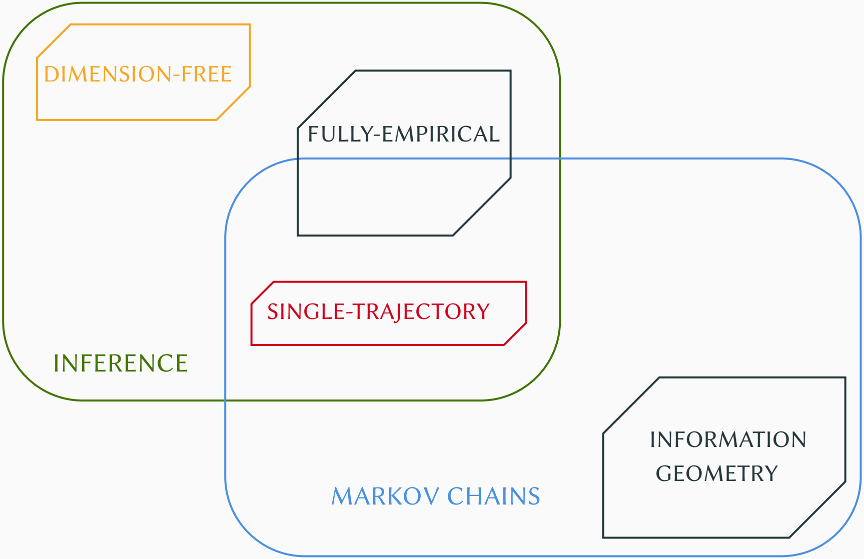
Inference in Markov Chain from a Single Finite Trajectory

Learning and Testing Markov Chains
Weekly Reading Group

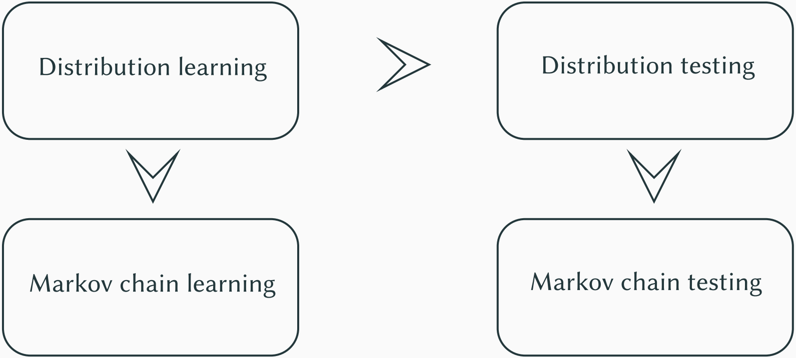
May 23, 2022

G. Wolfer – RIKEN AIP

Overview


DIMENSION-FREE
FULLY-EMPIRICAL
SINGLE-TRAJECTORY
INFERENCE
INFORMATION GEOMETRY
MARKOV CHAINS

Overview


Distribution learning
Distribution testing
Markov chain learning
Markov chain testing

Distribution learning

Distribution learning

(i) Unknown distribution over finite space $\mathcal{X}$: $\mu \in \mathcal{P}(\mathcal{X})$.

(ii) Access to a sample

$$X = (X_1, X_2, \dots, X_n) \sim \mu^{\otimes n}.$$

(iii) Total variation metric

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|.$$

(iv) Sample complexity for ε-precision, $(1-\delta)$-confidence

$$n_0(\varepsilon, \delta) \doteq \underset{n \in \mathbb{N}}{\arg\min} \left\{ \inf_{\hat{\mu}_n} \sup_{\mu} \mathbb{P}_{X \sim \mu^{\otimes n}} \left(\|\hat{\mu}_n - \mu\|_{\text{TV}} > \varepsilon \right) < \delta \right\},$$

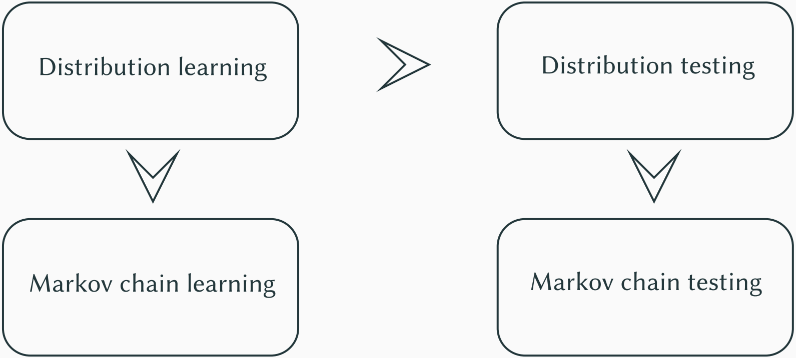
(v) Design lower and upper bounds for $n_0(\varepsilon, \delta)$,

$$L(\varepsilon, \delta) \leq n_0(\varepsilon, \delta) \leq U(\varepsilon, \delta).$$

(vi) Answer (folklore for $|\mathcal{X}| < \infty$) (Dimension free?)

$$n_0(\varepsilon, \delta) = \Theta\left(\frac{|\mathcal{X}| \vee \log 1/\delta}{\varepsilon^2} \right).$$

Overview


Distribution learning
Distribution testing
Markov chain learning
Markov chain testing

Background – Markov chains

(i) Discrete time, time homogeneous Markov chain,

$$X = (X_1, X_2, \ldots, X_m)$$

$$\forall x \in \mathcal{X}^m, \mathbb{P}\left(X = x\right) = \mu(x_1) \prod_{t=1}^{m-1} P(x_t, x_{t+1}).$$

(ii) Stationary distribution $\pi P = \pi$.

(iii) (Often) Irreducible, aperiodic.

(iv) Mixing time

$$t_{\text{mix}} \doteq \operatorname*{arg\,min}_{t \in \mathbb{N}} \max_{\mu} \left\| \mu P^t - \pi \right\|_{\text{TV}} \leq 1/4.$$

(v) (Sometimes) Reversible,

$$\pi(x)P(x, x') = \pi(x')P(x', x).$$

Sampling model – Single-trajectory

(i) Single-trajectory.

(ii) No restarts.

(iii) Arbitrary starting state.

(iv) Sample complexity: length of the trajectory.

Markov Chain Estimation

Estimation – Uniform metric

Uniform metric (suggested by John Lafferty)

$$\left\| P - P' \right\|_\infty \doteq \frac{1}{2} \max_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} \left| P(x, x') - P'(x, x') \right| .$$

$\Theta(|\mathcal{X}|^2)$-parameters:

$$m_0(\varepsilon) \overset{?}{\doteq} \Theta\left(\frac{|\mathcal{X}|^2}{\varepsilon^2} \right) .$$

Sample complexity (Wolfer and Kontorovich, 2019, 2021)

$$\Omega\left(\frac{|\mathcal{X}|}{\pi_{\min}\varepsilon^2} + \frac{|\mathcal{X}|}{\gamma_{\mathrm{ps}}} \right) \leq m_0(\varepsilon) \leq \tilde{\mathcal{O}}\left(\frac{|\mathcal{X}|}{\pi_{\min}\varepsilon^2} + \frac{1}{\pi_{\min}\gamma_{\mathrm{ps}}} \right),$$

with $\pi_{\min} \doteq \min_{x \in \mathcal{X}} \pi(x)$, $\gamma_{\mathrm{ps}} \doteq \max_{k \in \mathbb{N}} \gamma\left(((P^\star)^k P^k)/k \right)$. ▸ Details LBs

Estimation – Uniform metric

Extension to irreducible (Chan, Ding, and Li, 2021)

$$m_0(\varepsilon) = \tilde{\Theta}\left(\frac{|\mathcal{X}|}{\pi_{\min}\varepsilon^2} + t_{\mathrm{cov}}\right),$$

$$t_{\mathrm{cov}} \doteq \max_{x_1 \in \mathcal{X}} \mathbb{E}\left[\operatorname*{arg\,min}_{n \in \mathbb{N}} \left\{ \min_{x \in \mathcal{X}} \left\{ \sum_{t=1}^{n} \mathbf{1}[X_t = x] \right\} > 0 \right\} \middle| X_1 = x_1 \right].$$

(*i*) More delicate characterization of sample complexity with t_{cov}.

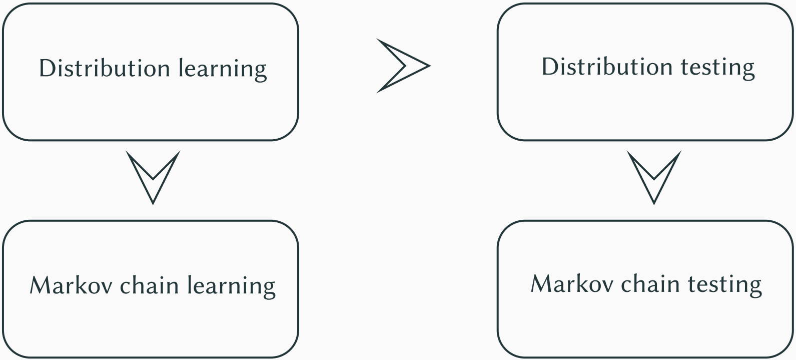
(*ii*) More difficult to compute than γ_{ps}, and no estimator available.

Extension to irreducible (Fried and Wolfer, 2021)

$$m_0(\varepsilon) = \tilde{\Theta}\left(\frac{|\mathcal{X}|}{\pi_{\min}\varepsilon^2} + \frac{1}{\pi_{\min}\gamma_{\mathrm{ps}}((P+I)/2)}\right),$$

Easy to simulate, easy to compute, possible to estimate.

Overview


Distribution learning
Distribution testing
Markov chain learning
Markov chain testing

Distribution testing

Identity Testing – Classical Results

Problem statement

Reference distribution $\mu_0 \in \mathcal{P}(\mathcal{X})$.

Access to iid sample $X_1, X_2, \ldots, X_n \sim \mu$ from unknown $\mu \in \mathcal{P}(\mathcal{X})$.

Distinguish between $H_0\colon \mu = \mu_0$ and $H_1\colon \|\mu - \mu_0\|_{\text{TV}} > \varepsilon$.

$$n_0(\varepsilon) = \min_{n \in \mathbb{N}} \left\{ \min_{\phi\colon \mathcal{X}^n \mapsto \{0,1\}} \left(\mathbb{P}_{\mu_0}(\phi = 1) + \max_{\mu \in H_1} \mathbb{P}_{\mu}(\phi = 0) \right) < \delta \right\}.$$

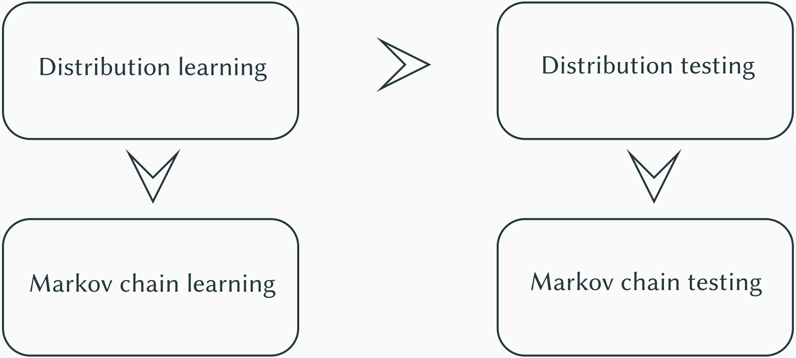
Uniformity testing: $\mu_0 = \text{Uniform}(\mathcal{X})$ **(Paninski, 2008)**

$$n_0 = \tilde{\Theta}\left(\frac{\sqrt{|\mathcal{X}|}}{\varepsilon^2} \right).$$

Instance optimal testing (Valiant and Valiant, 2017)

$$n_0 = \tilde{\Theta}\left(\frac{\|\mu_0\|_{2/3}}{\varepsilon^2} \right).$$

Overview


Distribution learning
Distribution testing
Markov chain learning
Markov chain testing

Markov Chain Identity Testing

Identity Testing Problem – Problem Statement

(i) Consider a reference kernel P_0.

(ii) Fix a metric (we will consider two) and a proximity parameter ε.

(iii) Sample a single trajectory from an unknown P.

(iv) Algorithm must distinguish between $P = P_0$ or $|P - P_0| > \varepsilon$.

Ergodic Reference – Under the Uniform Metric

Uniform metric

$$\left|P - P'\right| = \frac{1}{2}\left\|P - P'\right\|_{\infty}.$$

Sample complexity (Wolfer and Kontorovich, 2020)

$$\Omega\left(\frac{\sqrt{|\mathcal{X}|}}{\pi_0^\star \varepsilon^2} + \frac{|\mathcal{X}|}{\gamma_{\mathrm{ps}_0}}\right) \leq m_0 \leq \tilde{\mathcal{O}}\left(\frac{\sqrt{|\mathcal{X}|}}{\pi_0^\star \varepsilon^2} + \frac{1}{\pi_0^\star \gamma_{\mathrm{ps}_0}}\right).$$

Observe: (i) only depends on reference; (ii) unknown chain need not be ergodic; (iii) quadratic reduction; (iv) nearly matching bounds; (v) no dependence in initial state.

Extensions

(i) Instance specific bounds (Wolfer and Kontorovich, 2020, Th. 4.2) See.

(ii) Can extend to irreducible reference chains (Fried and Wolfer, 2021).

(iii) Can obtain rates in terms of cover times (Chan et al., 2021).

Symmetric Chains – Under the Kazakos Divergence

Divergence / Contrast function (Kazakos, 1978)

$$\left| P - P' \right| = 1 - \rho \left(P^{\circ 1/2} \circ P'^{\circ 1/2} \right)$$

(i) Not a proper metric.

(ii) Vanishes for chains with identical connected components.

(iii) Well-adapted for stochastic processes.

$$\lim_{n \to \infty} \frac{1}{n} D_{1/2}(Q^n \| Q'^n) = -2 \log(1 - \left| P - P' \right|).$$

(iv) Metric domination (Wolfer and Kontorovich, 2020, Lemma 8.1)

$$\left\| P - P' \right\|_\infty \geq 2 \left| P - P' \right|.$$

State-of-the-Art – Kazakos Divergence

	Conditions on $\bar{P}$	Conditions on P	Upper bound	Lower bound
[1]	$\bar{P} \in \mathcal{W}_{\text{sym}}$ $\bar{\pi} \propto 1$ $\bar{\pi}_{\text{min}} = 1/\lvert\mathcal{X}\rvert$	$P \in \mathcal{W}_{\text{sym}}$ $\pi \propto 1$ $\pi_{\text{min}} = 1/\lvert\mathcal{X}\rvert$ $\pi = \bar{\pi}$	$\tilde{\mathcal{O}}\left(\lvert\mathcal{X}\rvert/\varepsilon + Hit\right)$	$\Omega\left(\lvert\mathcal{X}\rvert/\varepsilon\right)$
[2]	$\bar{P} \in \mathcal{W}_{\text{sym}}$ $\bar{\pi} \propto 1$ $\bar{\pi}_{\text{min}} = 1/\lvert\mathcal{X}\rvert$	$P \in \mathcal{W}_{\text{sym}}$ $\pi \propto 1$ $\pi_{\text{min}} = 1/\lvert\mathcal{X}\rvert$ $\pi = \bar{\pi}$	$\tilde{\mathcal{O}}\left(\lvert\mathcal{X}\rvert/\varepsilon^4\right)$	–
[3]	$\bar{P} \in \mathcal{W}_{\text{rev}}$	$P \in \mathcal{W}_{\text{rev}}$ $\lVert \pi/\bar{\pi} - 1 \rVert_\infty < \varepsilon$	$\tilde{\mathcal{O}}\left(1/(\bar{\pi}_{\text{min}}\varepsilon^4)\right)$	–

Table 1: [1] Daskalakis et al. (2018); [2] Cherapanamjeri and Bartlett (2019); [3] Fried and Wolfer (2022)

References

Siu On Chan, Qinghua Ding, and Sing Hei Li. Learning and testing irreducible Markov chains via the k-cover time. In *Algorithmic Learning Theory*, pages 458–480. PMLR, 2021.

Y. Cherapanamjeri and P. L. Bartlett. Testing symmetric Markov chains without hitting. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 758–785. PMLR, 2019.

Doron Cohen, Aryeh Kontorovich, and Geoffrey Wolfer. Learning discrete distributions with infinite support. In *Advances in Neural Information Processing Systems*, volume 33, pages 3942–3951, 2020.

C. Daskalakis, N. Dikkala, and N. Gravin. Testing symmetric Markov chains from a single trajectory. In *Conference On Learning Theory*, pages 385–409. PMLR, 2018.

Sela Fried and Geoffrey Wolfer. On the α-lazy version of Markov chains in estimation and testing problems, 2021.

Sela Fried and Geoffrey Wolfer. Identity testing of reversible Markov chains. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 798–817. PMLR, 2022.

D. Kazakos. The Bhattacharyya distance and detection between Markov chains. *IEEE Transactions on Information Theory*, 24(6):747–754, 1978.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

Geoffrey Wolfer and Aryeh Kontorovich. Minimax learning of ergodic Markov chains. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 904–930. PMLR, 2019.

Geoffrey Wolfer and Aryeh Kontorovich. Minimax testing of identity to a reference ergodic Markov chain. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 191–201. PMLR, 26–28 Aug 2020.

Geoffrey Wolfer and Aryeh Kontorovich. Statistical estimation of ergodic Markov chain kernel over discrete state space. *Bernoulli*, 27(1):532–553, 02 2021. doi: 10.3150/20-BEJ1248.

Thank you for listening!

Dimension-free distribution learning ▸ Back

...but what if $\mathcal{X} \cong \mathbb{N}$?

$$n_0(\varepsilon, \delta) = \Theta\left(\frac{\|\mu\|_{1/2} \vee \log 1/\delta}{\varepsilon^2}\right).$$

...but what if $\|\mu\|_{1/2} = \infty$? Cut the tail!

$$n_0(\varepsilon, \delta) = \Theta\left(\frac{\left\|\mu_{\Theta(\varepsilon\delta)}\right\|_{1/2} \vee \log 1/\delta}{\varepsilon^2}\right).$$

...but what if no upper bound on half-norm? Do adaptively!
With probability at least $1 - \delta$,

$$\|\hat{\mu}_n - \mu\|_{\mathrm{TV}} \leq \underbrace{\frac{\|\hat{\mu}_n\|_{1/2}}{\sqrt{n}}}_{\text{converges}} + 3\sqrt{\frac{\log 2/\delta}{2n}}$$

(Cohen, Kontorovich, and Wolfer, 2020).

Instance specific upper bound

$$m_0 \leq \tilde{\mathcal{O}}\left(\frac{\Gamma(P_0)}{\varepsilon^2} + \frac{1}{\pi_0^\star \gamma_{\mathrm{ps}_0}}\right),$$

with

$$\Gamma(P) \doteq \max_{x \in \mathcal{X}} \left\{ \frac{\left\| e_x P \right\|_{2/3}}{\pi(x)} \right\}.$$

Example 1 (Simple random walk on Δ-regular graph)

$$m_0 \leq \tilde{\mathcal{O}}\left(|\mathcal{X}|\left(\frac{\sqrt{\Delta}}{\varepsilon^2} + \frac{1}{\gamma_{\mathrm{ps}_0}}\right)\right),$$

Proximity lower bound $\tilde{\Omega}\left(|\mathcal{X}| / (\varepsilon^2 \pi_{\min})\right)$ Back

Strategy: Make one state both *difficult to reach* and *hard to learn.*

$$\mathcal{G}_p = \left\{ P_\eta = \begin{pmatrix} p_1 & \dots & p_d & p_\star \\ \vdots & \vdots & \vdots & \vdots \\ p_1 & \dots & p_d & p_\star \\ \eta_1 & \dots & \eta_d & p_\star \end{pmatrix} : \eta = (\eta_1, \dots, \eta_d, p_\star) \in \Delta_{d+1} \right\}$$

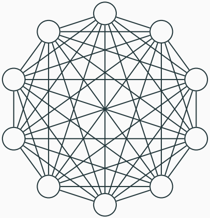
$\# \{\text{visits to special state}\} \sim \text{Binomial}(m, p_\star)$

$\pi_\star = p_\star$

$D\left(X_1^m \sim P_\eta \middle|\middle| X_1^m \sim P_{\eta'}\right) \leq p_\star m D\left(\eta || \eta'\right)$

Construct ε-packing w.r.t. $\|\cdot\|_\infty$ (with $\approx 2^{\Theta(|\mathcal{X}|)}$ elements, separated by $> \Theta(|\mathcal{X}|)$ in Hamming distance)

Mixing lower bound $\Omega\left(|\mathcal{X}| \log |\mathcal{X}| / \gamma_{\text{ps}}\right)$ ▸ Back



What about this construction ?

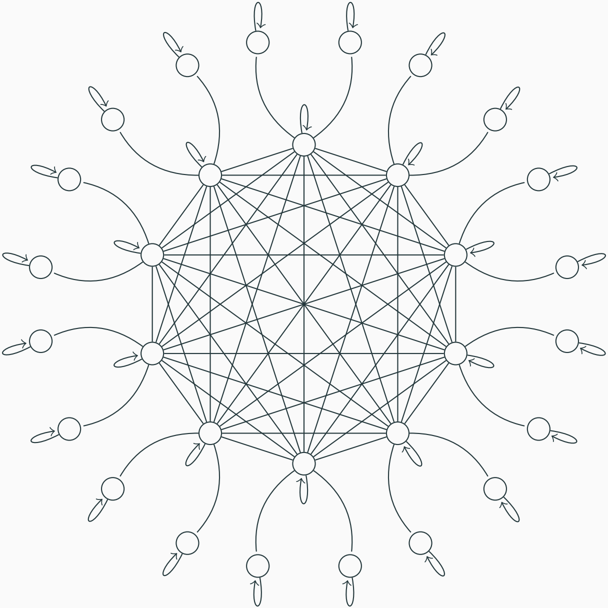
$P(x, x') \approx \mathbf{1}[x = x'](1 - \eta) + \mathbf{1}[x \neq x']\left(\frac{\eta}{|\mathcal{X}|}\right)$

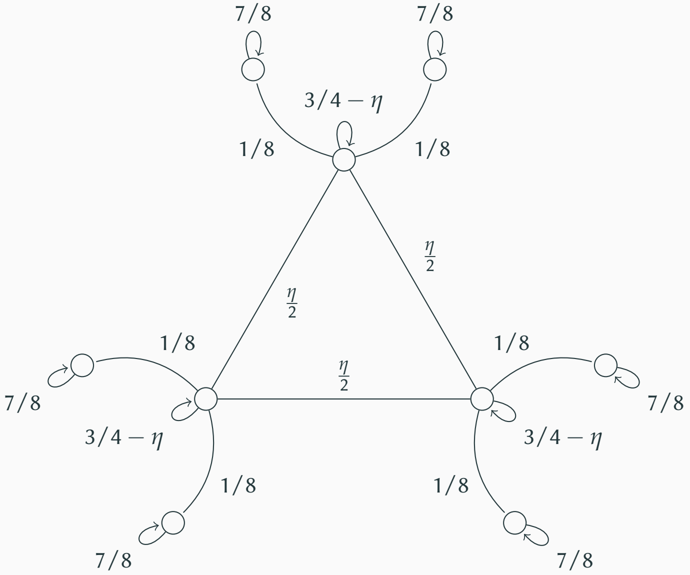
$\gamma_{\text{ps}} \approx \eta^{-1}$

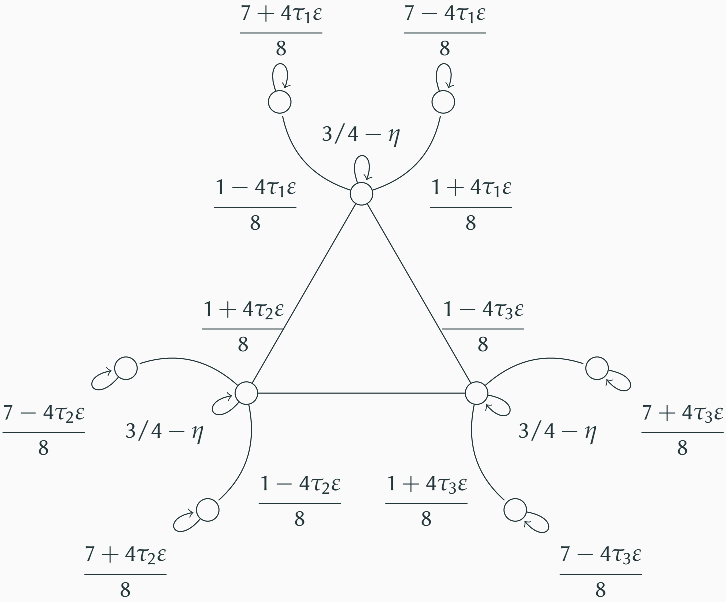
...but we need an ε-packing w.r.t $\|\cdot\|_\infty$

...$\eta, \gamma_{\text{ps}}, \varepsilon$ all coupled

...only yields a lower bound of $\Omega(|\mathcal{X}| / \varepsilon)$




7/8
7/8
3/4 − η
1/8
1/8
η/2
η/2
1/8
1/8
η/2
7/8
7/8
3/4 − η
3/4 − η
1/8
1/8
7/8
7/8


$\frac{7+4\tau_1\varepsilon}{8}$
$\frac{7-4\tau_1\varepsilon}{8}$
$3/4-\eta$
$\frac{1-4\tau_1\varepsilon}{8}$
$\frac{1+4\tau_1\varepsilon}{8}$
$\frac{1+4\tau_2\varepsilon}{8}$
$\frac{1-4\tau_3\varepsilon}{8}$
$\frac{7-4\tau_2\varepsilon}{8}$
$3/4-\eta$
$3/4-\eta$
$\frac{7+4\tau_3\varepsilon}{8}$
$\frac{1-4\tau_2\varepsilon}{8}$
$\frac{1+4\tau_3\varepsilon}{8}$
$\frac{7+4\tau_2\varepsilon}{8}$
$\frac{7-4\tau_3\varepsilon}{8}$

Control of the mixing time

$$\gamma_{\text{ps}} \approx \frac{1}{\eta} \tag{1}$$

Note: ε and η are uncoupled

Lower bound on cover time
T the time to cover all the nodes in the *central clique*

$$m \lessapprox \frac{|\mathcal{X}| \log |\mathcal{X}|}{\eta} \implies p(T > m) \geq \frac{1}{20} \tag{2}$$

Fail to cover $\implies$ have to toss a coin.

Back