CS395T: Continuous Algorithms, Part XIV
Spectral graph theory

Kevin Tian

1 Discrete Markov chains

In this lecture, we begin our exploration of algorithms for sampling from distributions. We specifically give tools for designing and analyzing *Markov chain Monte Carlo* algorithms, whose strategy for producing samples from a target stationary distribution $\boldsymbol{\pi}$ is to run a random walk which converges, in appropriate senses which we will make formal, to $\boldsymbol{\pi}$. We focus this lecture on the discrete Markov chain setting, where our goal is to produce a sample from a distribution $\boldsymbol{\pi}$ supported on $[d]$, which we view as a set of states that the random walk can take on.

A discrete Markov chain is specified by its *transition matrix* $\mathbf{P} \in \mathbb{R}^{d\times d}_{\geq 0}$; we view random walks and transition matrices interchangeably. We let $\mathbf{P}_{i:}$ denote the distribution of one step of the random walk starting at the state $i \in [d]$. This means $\mathbf{P}_{ij}$ is the probability that a random walk starting at $i \in [d]$ moves to $j \in [d]$, so $\mathbf{P}_{i:} \in \Delta^d$, the probability simplex in $\mathbb{R}^d$, for all $i \in [d]$. One pleasing consequence of viewing the random walk transitions as matrices is we can easily compute the evolution of distributions over $[d]$, when taking steps according to $\mathbf{P}$. To see this, let $\boldsymbol{\mu} \in \Delta^d$, and observe that if we randomly choose a starting state $i \in [d]$ proportional to $\boldsymbol{\mu}$, then the probability we end up at state $j \in [d]$ after taking one step of the random walk given by $\mathbf{P}$ is

$$\sum_{i\in[d]} \boldsymbol{\mu}_i \mathbf{P}_{ij} = \left[\mathbf{P}^\top \boldsymbol{\mu}\right]_j .$$

Iterating on this calculation shows that the transition matrix given by taking $k \in \mathbb{N}$ steps of the random walk in a row is $\mathbf{P}^k$. We additionally observe that $\mathbf{P}\mathbf{1}_d = \mathbf{1}_d$ since all rows of $\mathbf{P}$ are in Δ^d, so $\mathbf{1}_d$ is a right eigenvector of $\mathbf{P}$ with eigenvalue 1. This implies that there is a corresponding left eigenvector $\boldsymbol{\pi}^\top$, also of eigenvalue 1, which means $\boldsymbol{\pi}^\top \mathbf{P} = \boldsymbol{\pi}^\top$. In other words, $\boldsymbol{\pi}$ is a *stationary distribution* for $\mathbf{P}$, which means that choosing an initial state according to $\boldsymbol{\pi}$ and taking one step preserves the distribution $\boldsymbol{\pi}$. We have shown that a stationary distribution always exists.

A natural follow-up question is: when is the stationary distribution (i.e., left eigenvector with eigenvalue 1) unique? If our goal is to use $\mathbf{P}$ to induce a target distribution $\boldsymbol{\pi}$, then we should at least be sure our algorithm is well-posed, i.e., there are not two different target distributions $\boldsymbol{\pi}, \boldsymbol{\pi}'$ we could converge to. Towards this end, we aim to develop a better understanding of spectral properties of $\mathbf{P}$. We begin with a simple proof that the spectrum is bounded.

Lemma 1. *If $\mathbf{P} \in \mathbb{R}^{d\times d}_{\geq 0}$ is a transition matrix with eigenvalue-eigenvector pair $(\lambda, \mathbf{v})$, then $|\lambda| \leq 1$.*

Proof. By definition, $\|\mathbf{P}\mathbf{v}\|_\infty = |\lambda|\|\mathbf{v}\|_\infty$.[1] However, we also know that $\|\mathbf{P}\|_{\infty\to\infty} = 1$, since the $\infty \to \infty$ norm is the largest ℓ_1 norm of a row, and $\mathbf{P}$ is a transition matrix. Therefore,

$$|\lambda| \|\mathbf{v}\|_\infty = \|\mathbf{P}\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_\infty \implies |\lambda| \leq 1.$$

□

To understand when there is a unique eigenvalue with magnitude 1, we turn to the *Perron-Frobenius theorem* [Per07, Fro12], a famous result in matrix analysis which says that any square matrix which is entrywise positive has a unique largest eigenvalue, that eigenvalue is real, and the corresponding eigenvector has positive entries. The Perron-Frobenius theorem and Lemma 1 immediately imply

[1] We define the ℓ_∞ norm for complex vectors to be the largest magnitude of any entry.

that for any $\mathbf{P}$ which is entrywise positive, i.e., all transitions have positive probability, there is a unique stationary distribution of the corresponding random walk. More generally, we require the following definition to characterize uniqueness of stationary distributions.

Definition 1 (Irreducibility and aperiodicity)**.** *Let $\mathbf{P} \in \mathbb{R}_{\geq 0}^{d \times d}$ be a transition matrix.*

1. *We say $\mathbf{P}$ is* irreducible *if for all $(i, j) \in [d] \times [d]$, there exists $k \in \mathbb{N}$ such that $\mathbf{P}_{ij}^k \neq 0$.*[2]
2. *We say $\mathbf{P}$ is* aperiodic *if for all $i \in [d]$, $\{k \in \mathbb{N} \mid \mathbf{P}_{ii}^k \neq 0\}$ has greatest common divisor 1.*[3]

We now prove uniqueness of stationary distributions for irreducible and aperiodic transitions.

Proposition 1. *Let $\mathbf{P} \in \mathbb{R}_{\geq 0}^{d \times d}$ be an irreducible and aperiodic transition matrix. Then, $\mathbf{P}$ has a unique left eigenvector with eigenvalue 1 which is entrywise positive, and all other eigenvalues of $\mathbf{P}$ have magnitude < 1.*

Proof. We assert that for some $k \in \mathbb{N}$, $\mathbf{P}^k$ is entrywise positive. Note that any eigenvector-eigenvalue pair $(\lambda, \mathbf{v})$ for $\mathbf{P}$ induces an eigenvector-eigenvalue pair for $\mathbf{P}^k$. Moreover, $\mathbf{P}^k$ has a left eigenvector with eigenvalue 1, because $\mathbf{P}^k \mathbf{1}_d = \mathbf{1}_d$, and left and right eigenvectors come in pairs corresponding to eigenvalues. This eigenvalue is unique by Lemma 1 and the Perron-Frobenius theorem, which also guarantees that the corresponding eigenvector is entrywise positive.

It remains to prove our earlier assertion. We sketch a proof, deferring details to Proposition 1.7, [LPW09]. The key technical claim we require is that any $S \subseteq \mathbb{N}$ which is relatively prime and closed under addition (i.e., $s, t \in S \implies s + t \in S$) contains all but finitely many elements of $\mathbb{N}$.[4]

Assuming this is true, let $S_i := \{k \in \mathbb{N} \mid [\mathbf{P}_{ii}]^k \neq 0\}$ for all $i \in [d]$. Then, S_i is relatively prime by aperiodicity, and it is closed under addition, since $[\mathbf{P}^{s+t}]_{ii} \geq [\mathbf{P}^s]_{ii}[\mathbf{P}^t]_{ii}$.[5] By applying our key technical claim, it follows that $\bigcap_{i \in [d]} S_i$ also contains all but finitely many elements of $\mathbb{N}$. Next, irreducibility implies that for every $(i, j) \in [d] \times [d]$, there is some r_{ij} such that $[\mathbf{P}^{r_{ij}}]_{ij} > 0$, and $[\mathbf{P}^k]_{ij}$ is positive as long as $k = r_{ij} + k'$, for some k' such that $\mathbf{P}_{ii}^{k'}$ is positive. Therefore, there is some $k \in \mathbb{N}$ such that $\mathbf{P}^k$ is entrywise positive, proving our assertion. □

In order to continue our analysis of the spectra of transition matrices, it is helpful at this juncture to introduce an alternative viewpoint, where we equate transition matrices with graphs.

Definition 2 (Graph-induced matrices)**.** *Let $G = (V, E, \mathbf{w})$ be a weighted graph on vertices V and directed edges $E \subset V \times V$, where $\mathbf{w} \in \mathbb{R}_{\geq 0}^E$ gives the weight of each edge. We let $\mathbf{D}_G$ denote the* out-degree matrix *of G, which is the diagonal matrix satisfying $[\mathbf{D}_G]_{vv} = \sum_{u \in V} \mathbf{w}_{(v,u)}$. We let $\mathbf{A}_G$ denote the* adjacency matrix *of G, where $[\mathbf{A}_G]_{uv} = \mathbf{w}_{(u,v)}$ for all $(u, v) \in E$.*

A key observation is that any weighted directed graph $G = (V, E, \mathbf{w})$ in the sense of Definition 2 is naturally identified with a transition matrix on states $[d]$, where $d = |V|$ and we equate vertices $v \in V$ with states $i \in [d]$ in an arbitrary but consistent way. To see this equivalence, consider

$$\mathbf{P} := \mathbf{D}_G^{-1} \mathbf{A}_G. \tag{1}$$

Note that the i^{th} row[6] of $\mathbf{P}$ is the outgoing weights from the i^{th} vertex, normalized by the out-degree of the vertex. It is straightforward to check that this implies $\mathbf{P}$ is a transition matrix. This equivalence has a simple interpretation: one step of the random walk induced by G starting from a vertex chooses an outgoing edge proportional to its weight. To see that this equivalence goes both ways, every transition matrix can be simply modeled by a graph such that every out-degree is 1 (i.e., it has $\mathbf{I}_V$ as an out-degree matrix), and an adjacency matrix given by $\mathbf{P}^\top$.

This equivalence is particularly interpretable when G is an undirected graph, meaning $\mathbf{A}_G^\top = \mathbf{A}_G$ (for every edge (u, v), there is an edge (v, u) with equal weight). In this case (and assuming irreducibility and aperiodicity), we have the following straightforward characterization.

[2] In other words, regardless of the starting state, there is positive probability of visiting every other state.

[3] In other words, the iteration counts where it is possible to return to the starting state are relatively prime.

[4] In the case where there are two generating elements of S, this is sometimes called the Chicken McNugget theorem. For a general proof, see Lemma 1.30, [LPW09].

[5] The probability that a random walk cycles from i to i in $s + t$ steps is at least the probability that it cycles in the first s steps, times the probability it cycles in the last t steps.

[6] We will interchange states of a Markov chain and vertices of a corresponding graph in the rest of the lecture.

Lemma 2. *Let $G = (V, E, \mathbf{w})$ be a weighted graph, and (following Definition 2) suppose $\mathbf{A}_G^\top = \mathbf{A}_G$ and $\mathbf{D}_G = \mathbf{diag}(\mathbf{d}_G)$. Then defining $\mathbf{P}$ as in (1), and supposing $\mathbf{P}$ is irreducible and aperiodic with unique stationary distribution $\boldsymbol{\pi}$, we have*

$$\boldsymbol{\pi} = \frac{\mathbf{d}_G}{\|\mathbf{d}_G\|_1}.$$

Proof. This follows from a direct computation:

$$\boldsymbol{\pi}^\top \mathbf{P} = \left(\frac{\mathbf{d}_G}{\|\mathbf{d}_G\|_1}\right)^\top \mathbf{D}_G^{-1}\mathbf{A}_G = \frac{1}{\|\mathbf{d}_G\|_1}\mathbf{1}_V^\top \mathbf{A}_G = \frac{\mathbf{d}_G^\top}{\|\mathbf{d}_G\|_1} = \boldsymbol{\pi}^\top.$$

The third equality above used that $\mathbf{A}_G^\top \mathbf{1}_V$ gives the in-degrees of each vertex, but this is exactly the out-degrees $\mathbf{d}_G$ of every vertex as well because $\mathbf{A}_G^\top = \mathbf{A}_G$.[7] □

Lemma 2 shows that the stationary distribution of a random walk corresponding to an undirected graph is proportional to the vertex degrees. The notion of undirectedness in graphs is closely related to a concept in Markov chain theory known as reversibility, which we next define.

Definition 3 (Reversible Markov chain)**.** *We say that a Markov chain associated with a transition matrix $\mathbf{P} \in \mathbb{R}_{\geq 0}^{d\times d}$ is* reversible *(or, $\mathbf{P}$ is reversible) if it has stationary distribution $\boldsymbol{\pi}$, and*

$$\boldsymbol{\pi}_i \mathbf{P}_{ij} = \boldsymbol{\pi}_j \mathbf{P}_{ji} \text{ for all } (i,j) \in [d]\times[d]. \tag{2}$$

Note that if we prove (2) holds for a transition matrix $\mathbf{P}$ and any $\boldsymbol{\pi} \in \Delta^d$, then $\boldsymbol{\pi}$ is stationary, because the probability we end up at i after taking one step from a distribution initialized via $\boldsymbol{\pi}$ is

$$\sum_{j\in[d]} \boldsymbol{\pi}_j \mathbf{P}_{ji} = \sum_{j\in[d]} \boldsymbol{\pi}_i \mathbf{P}_{ij} = \boldsymbol{\pi}_i. \tag{3}$$

Just as every Markov chain can be identified with a directed graph via (1), every reversible Markov chain is naturally identified with an undirected graph. To see this, consider a graph G with out-degrees $\boldsymbol{\pi}$ and adjacency matrix $\mathbf{A}_G = \boldsymbol{\Pi}\mathbf{P}$, which induces $\mathbf{P}$ via (1). Then, (2) implies

$$[\mathbf{A}_G]_{ij} = \boldsymbol{\pi}_i \mathbf{P}_{ij} = \boldsymbol{\pi}_j \mathbf{P}_{ji} = [\mathbf{A}_G]_{ji},$$

i.e., G is undirected. Similarly, starting from an undirected graph G, the corresponding transition matrix (1) satisfies (2), which is verifiable by undoing the above sequence of derivations with the observation that $\boldsymbol{\pi}$ is proportional to the diagonal elements of $\mathbf{D}_G$ via Lemma 2.

We next observe that for an arbitrary (potentially non-reversible) transition matrix $\mathbf{P}$, and a target $\boldsymbol{\pi} \in \Delta^d$, there is a simple modification to $\mathbf{P}$ proposed by [MRR⁺53, Has70] called the Metropolis-Hastings rule, which forces $\mathbf{P}$ to both be reversible and have stationary distribution $\boldsymbol{\pi}$.

Lemma 3 (Metropolis-Hastings)**.** *Let $\mathbf{P} \in \mathbb{R}_{\geq 0}^{d\times d}$ be a transition matrix, and for $\boldsymbol{\pi} \in \Delta^d$, denote the modified transition matrix $\widetilde{\mathbf{P}} \in \mathbb{R}_{\geq 0}^{d\times d}$ by*[8]

$$\widetilde{\mathbf{P}}_{ij} = \mathbf{P}_{ij} \min\left(1, \frac{\boldsymbol{\pi}_j \mathbf{P}_{ji}}{\boldsymbol{\pi}_i \mathbf{P}_{ij}}\right) \text{ for all } (i,j) \in [d]\times[d] \text{ with } i \neq j, \tag{4}$$

and such that $\{\widetilde{\mathbf{P}}_{ii}\}_{i\in[d]}$ are chosen so $\widetilde{\mathbf{P}}$ is a transition matrix. Then $\boldsymbol{\pi}$ is stationary for $\widetilde{\mathbf{P}}$.

Proof. We verify (2) holds for $\widetilde{\mathbf{P}}$, which implies $\boldsymbol{\pi}$ is stationary by (3). To see (2), we compute

$$\boldsymbol{\pi}_i \widetilde{\mathbf{P}}_{ij} = \mathbf{P}_{ij} \min\left(\boldsymbol{\pi}_i, \frac{\boldsymbol{\pi}_j \mathbf{P}_{ji}}{\mathbf{P}_{ij}}\right) = \min\left(\boldsymbol{\pi}_i \mathbf{P}_{ij}, \boldsymbol{\pi}_j \mathbf{P}_{ji}\right) = \mathbf{P}_{ji} \min\left(\boldsymbol{\pi}_j, \frac{\boldsymbol{\pi}_i \mathbf{P}_{ij}}{\mathbf{P}_{ji}}\right) = \boldsymbol{\pi}_j \widetilde{\mathbf{P}}_{ji}.$$

□

In general, the Metropolis-Hastings correction (4) may change the support of the associated graph's adjacency matrix. However, under mild conditions, e.g., $\mathbf{P}_{ji} > 0$ iff $\mathbf{P}_{ij} > 0$, the support of the adjacency matrix is unchanged, so the correction (4) preserves irreducibility and aperiodicity.

[7] We remark that this proof generalizes to *Eulerian graphs*, i.e., graphs which have equal in-degrees and out-degrees for every vertex (but are not necessarily undirected).

[8] Alternatively, $\widetilde{\mathbf{P}}$ takes a step according to a proposal $\mathbf{P}$, and accepts the move with a probability in $[0, 1]$.

2 Mixing times

Beyond simple characterizations such as Lemma 2, why do we insist on reversibility as a desirable property for Markov chains? One reason is that the resulting matrix $\mathbf{P}$ is similar to a symmetric matrix, via the characterization (1): for any G identified with a reversible transition matrix $\mathbf{P}$,

$$\mathbf{D}_G^{\frac{1}{2}}\mathbf{P}\mathbf{D}_G^{-\frac{1}{2}} = \mathbf{D}_G^{-\frac{1}{2}}\mathbf{A}_G\mathbf{D}_G^{-\frac{1}{2}} = \left(\mathbf{D}_G^{-\frac{1}{2}}\mathbf{A}_G\mathbf{D}_G^{-\frac{1}{2}}\right)^\top, \tag{5}$$

where the last equation used that graphs corresponding to reversible Markov chains are undirected. This means that the spectral theorem applies to $\mathbf{P}$, which gives us powerful algebraic tools to analyze its convergence. Before stating a such a convergence result, we first give a brief digression on the total variation distance, a natural measure of the convergence of sampling algorithms.

Definition 4 (Total variation distance)**.** *For two distributions P, Q on the same continuous sample space Ω, we define their* total variation distance *(where we overload P, Q to denote the respective probability density functions) by*

$$D_{\mathrm{TV}}(P, Q) := \frac{1}{2}\int_{\omega\in\Omega} |P(\omega) - Q(\omega)|\,\mathrm{d}\omega.$$

When Ω is discrete, we analogously let

$$D_{\mathrm{TV}}(P, Q) = \frac{1}{2}\sum_{\omega\in\Omega} |P(\omega) - Q(\omega)|\,.$$

The total variation distance enjoys the following characterizations (see Section 4.1, [LPW09]).

Fact 1. *For two distributions P, Q, on the same sample space Ω, we have the following equivalent characterizations of $D_{\mathrm{TV}}(P, Q)$.*

1. $D_{\mathrm{TV}}(P, Q) = \sup_{A\subseteq\Omega} \Pr_{\omega\sim P}[\omega \in A] - \Pr_{\omega\sim Q}[\omega \in A]$.
2. $D_{\mathrm{TV}}(P, Q) = \inf_{\gamma\in\mathcal{C}(P,Q)} \Pr_{(\omega,\omega')\sim\gamma}[\omega \neq \omega']$, *where $\mathcal{C}(P, Q)$ is the set of all* couplings *of (P, Q), i.e., distributions γ on $\Omega\times\Omega$ such that for $(\omega, \omega') \sim \gamma$, the marginal distribution of ω is P, and the marginal distribution of ω' is Q.*

The coupling characterization in Fact 1 is particularly useful when composing sampling algorithms, because of the union bound. For example, suppose we first run a sampling algorithm $\mathcal{A}$ to approximately produce a sample from a distribution P up to total variation distance ϵ, and then given the output of $\mathcal{A}$, we run another algorithm $\mathcal{A}'$ which approximately samples from P' up to total variation distance ϵ', provided it was initialized with P. By using Fact 1, we can bound the total variation distance between $\mathcal{A}' \circ \mathcal{A}$ to P', by first coupling the output of $\mathcal{A}$ to P, and then pretending $\mathcal{A}'$ was initialized with P, which happens except with probability ϵ under some coupling. The overall failure probability of the best coupling is then upper bounded by $\epsilon + \epsilon'$.

We mention that the optimal coupling γ in the second part of Fact 1 is the one that, with probability $1 - D_{\mathrm{TV}}(P, Q)$, returns (ω, ω) for $\omega \in \Omega$ sampled with probability

$$\frac{\min\{P(\omega), Q(\omega)\}}{\int_{\omega'\in\Omega} \min\{P(\omega'), Q(\omega')\}\mathrm{d}\omega'} = \frac{\min\{P(\omega), Q(\omega)\}}{1 - D_{\mathrm{TV}}(P, Q)},$$

and otherwise returns (ω', ω'') for an arbitrary coupling of the densities

$$\begin{aligned}\frac{P(\omega') - \min\{P(\omega'), Q(\omega')\}}{\int_{\omega\in\Omega} (P(\omega) - \min\{P(\omega), Q(\omega)\})\,\mathrm{d}\omega} &= \frac{P(\omega') - \min\{P(\omega'), Q(\omega')\}}{D_{\mathrm{TV}}(P, Q)},\\ \frac{P(\omega'') - \min\{P(\omega''), Q(\omega'')\}}{\int_{\omega\in\Omega} (P(\omega) - \min\{P(\omega), Q(\omega)\})\,\mathrm{d}\omega} &= \frac{Q(\omega'') - \min\{P(\omega''), Q(\omega'')\}}{D_{\mathrm{TV}}(P, Q)}.\end{aligned}$$

The marginal density of the first coordinate is then $\min\{P, Q\} + P - \min\{P, Q\} = P$, and similarly the second marginal is distributed as Q. In the above calculations, we used that

$$\begin{aligned}\int_{\omega\in\Omega} (P(\omega) - \min\{P(\omega), Q(\omega)\})\,\mathrm{d}\omega &= 1 - \int_{\omega\in\Omega} \min\{P(\omega), Q(\omega)\}\mathrm{d}\omega\\ &= \int_{\omega\in\Omega} (Q(\omega) - \min\{P(\omega), Q(\omega)\})\,\mathrm{d}\omega,\end{aligned} \tag{6}$$

and therefore

$$D_{\mathrm{TV}}\left(P, Q\right) = \frac{1}{2}\int_{\omega\in\Omega} |P(\omega) - Q(\omega)| \mathrm{d}\omega = \frac{1}{2}\int_{\omega\in\Omega} \left(\max\{P(\omega), Q(\omega)\} - \min\{P(\omega), Q(\omega)\}\right) \mathrm{d}\omega$$
$$= \frac{1}{2}\left(\int_{\omega\in\Omega} \left(P(\omega) - \min\{P(\omega), Q(\omega)\}\right) \mathrm{d}\omega + \int_{\omega\in\Omega} \left(Q(\omega) - \min\{P(\omega), Q(\omega)\}\right) \mathrm{d}\omega\right).$$

Finally, (6) showed that the two quantities in the final expression above are equal, so they both are $D_{\mathrm{TV}}(P, Q)$. We are now ready to give our first convergence analysis on random walks, measured in the total variation distance. Our proof relies on the spectral theorem applied to $\mathbf{P}$, in a similar way as the power method for principal component analysis (Theorem 2, Part XI).

Theorem 1 (Spectral convergence of random walks)**.** *Let $\mathbf{P} \in \mathbb{R}^{d\times d}_{\ge 0}$ be a reversible, irreducible, and aperiodic transition matrix with stationary distribution $\boldsymbol{\pi}$, let $\mathbf{\Pi} := \mathbf{diag}\left(\boldsymbol{\pi}\right)$, and let $\epsilon \in (0, 1)$. Let $\{\lambda_i\}_{i\in[d]}$ be the eigenvalues of $\mathbf{\Pi}^{\frac{1}{2}}\mathbf{P}\mathbf{\Pi}^{-\frac{1}{2}}$ in nondecreasing order, with $1 = \lambda_1 > \lambda_2$. Then for any $\boldsymbol{\mu} \in \Delta^d$, we have*

$$D_{\mathrm{TV}}\left(\left(\mathbf{P}^k\right)^\top \boldsymbol{\mu}, \boldsymbol{\pi}\right) \le \epsilon, \text{ if } k \ge \log_{\frac{1}{\max(|\lambda_2|, |\lambda_d|)}}\left(\frac{1}{2\epsilon \min_{i\in[d]} \boldsymbol{\pi}_i}\right).$$

Proof. By Proposition 1, we have that $|\lambda_i| < 1$ for all $i \ne 1$, and all eigenvalues of $\widetilde{\mathbf{P}} := \mathbf{\Pi}^{\frac{1}{2}}\mathbf{P}\mathbf{\Pi}^{-\frac{1}{2}}$ are real because $\widetilde{\mathbf{P}}$ is symmetric via reversibility of $\mathbf{P}$. Next, observe that if we let $\{\mathbf{v}_i\}_{i\in[d]}$ be the eigenvectors of $\widetilde{\mathbf{P}}$ sorted the same way as $\{\lambda_i\}_{i\in[d]}$, so that $\mathbf{v}_1 = \sqrt{\boldsymbol{\pi}}$ where $\sqrt{\cdot}$ is entrywise,

$$\mathbf{P}^k = \mathbf{\Pi}^{-\frac{1}{2}}\widetilde{\mathbf{P}}^k\mathbf{\Pi}^{\frac{1}{2}} = \mathbf{\Pi}^{-\frac{1}{2}}\sum_{i\in[d]} \lambda_i \mathbf{v}_i\mathbf{v}_i^\top \mathbf{\Pi}^{\frac{1}{2}} = \mathbf{1}_d\boldsymbol{\pi}^\top + \sum_{i=2}^d \lambda_i^k \left(\mathbf{\Pi}^{-\frac{1}{2}}\mathbf{v}_i\right)\left(\mathbf{\Pi}^{\frac{1}{2}}\mathbf{v}_i\right)^\top.$$

Therefore, for all $(i, j) \in [d] \times [d]$, we have by our lower bound on k,

$$\left[\mathbf{P}^k\right]_{ij} - \boldsymbol{\pi}_j = \sqrt{\frac{\boldsymbol{\pi}_j}{\boldsymbol{\pi}_i}}\sum_{\ell=2}^d \lambda_\ell^k [\mathbf{v}_\ell]_i[\mathbf{v}_\ell]_j \le \left(\max_{2\le\ell\le d} |\lambda_\ell|^k\right)\frac{\boldsymbol{\pi}_j}{\sqrt{\boldsymbol{\pi}_i\boldsymbol{\pi}_j}} \le \frac{\max\left(|\lambda_2|, |\lambda_d|\right)^k}{\min_{\ell\in[d]}\boldsymbol{\pi}_\ell}\cdot\boldsymbol{\pi}_j \le 2\epsilon\boldsymbol{\pi}_j. \quad (7)$$

Finally, we compute the j^{th} coordinate of $(\mathbf{P}^k)^\top\boldsymbol{\mu} - \boldsymbol{\pi}$:

$$\left|\left[(\mathbf{P}^k)^\top\boldsymbol{\mu}\right]_j - \boldsymbol{\pi}_j\right| = \left|\left(\sum_{i\in[d]}\left[\mathbf{P}^k\right]_{ij}\boldsymbol{\mu}_i\right) - \boldsymbol{\pi}_j\right| \le 2\epsilon\boldsymbol{\pi}_j,$$

using (7). The conclusion follows by summing the above display over $j \in [d]$, since $\left\|\boldsymbol{\pi}\right\|_1 = 1$. □

The takeaway from Theorem 1 is that if all eigenvalues of $\mathbf{P}$ are contained in the range $[-1 + \rho, 1 - \rho]$ for some $\rho > 0$, then the random walk achieves ϵ total variation from $\boldsymbol{\pi}$ from an arbitrary starting distribution, after $\approx \frac{1}{\rho}\log(\frac{d}{\epsilon})$ steps, assuming the stationary distribution $\boldsymbol{\pi}$ places at least a polynomially-small amount of weight on every state. To make Theorem 1 more interpretable, a common strategy is to assume that $\mathbf{P}$ is *lazy*, i.e., there is a transition matrix $\mathbf{P}'$ such that

$$\mathbf{P} = \frac{1}{2}(\mathbf{I}_d + \mathbf{P}'). \quad (8)$$

If $\mathbf{P}'$ is irreducible and aperiodic, so is $\mathbf{P}$ because the support of the relevant graph has not changed. In this case, it is further straightforward to check via the characterizations (1), (5) that the eigenvalues of $\mathbf{P}$ are the average of the eigenvalues of $\mathbf{P}'$ (which are all in $(-1, 1]$ by Proposition 1) and the eigenvalues of $\mathbf{I}_d$, so they all lie in $(0, 1]$. In this case, the convergence rate in Theorem 1 can be restated as achieving ϵ total variation when

$$k \approx \frac{1}{1 - \lambda_2}\log\left(\frac{1}{\epsilon\min_{i\in[d]}\boldsymbol{\pi}_i}\right).$$

As discussed previously, this matches the expected behavior of the power method applied to $\mathbf{P}$ to compute a leading eigenvector (Theorem 2, Part XI). The quantity $1 - \lambda_2$ is referred to as a *spectral* *gap* in this context; the further λ_2 is separated from 1, the faster the walk mixes. The matrix (8) is called lazy because in each iteration, it chooses to skip the iteration with probability $\frac{1}{2}$.[9]

Finally, we mention that the linear convergence rate of Theorem 1 is a more generic phenomenon than the proof belies. That is, constant total variation guarantees can always be boosted to ϵ distance at a $\log \frac{1}{\epsilon}$ overhead, even without the reversibility assumption or any spectral characterization.[10] We now illustrate this phenomenon with the following formal definition.

Definition 5 (Mixing time)**.** *Let $\mathbf{P} \in \mathbb{R}^{d\times d}_{\geq 0}$ be an irreducible and aperiodic transition matrix with stationary distribution $\boldsymbol{\pi}$. For any $\epsilon \in (0, 1)$, we define $\tau_{\text{mix}}(\epsilon)$, the* ϵ-mixing time *of $\mathbf{P}$, to be the smallest integer $k \in \mathbb{N}$ such that*[11]

$$\max_{\boldsymbol{\mu}\in\Delta^d} D_{\text{TV}}\left(\left(\mathbf{P}^k\right)^\top \boldsymbol{\mu}, \boldsymbol{\pi}\right) \leq \epsilon. \tag{9}$$

If $\epsilon = \frac{1}{4}$, we call this the mixing time *of $\mathbf{P}$ for short.*

As an application of the coupling characterization of D_{TV} in Fact 1, note that if (9) holds for a value of k, then it also holds for any $k' \geq k$. To see why, there is a coupling $\gamma_k \in \Gamma((\mathbf{P}^k)^\top \boldsymbol{\mu}, \boldsymbol{\pi})$ that sets $i = i'$ except with probability ϵ, for $(i, i') \sim \gamma$. Now consider the coupling of $\Gamma((\mathbf{P}^{k'})^\top \boldsymbol{\mu}, \boldsymbol{\pi})$ which first draws $(i, i') \sim \gamma_k$, and then advances (i, i') in the same way according to $\mathbf{P}$ for $k' - k$ steps if $i = i'$, and otherwise arbitrarily applies $\mathbf{P}$ for $k' - k$ steps. The output distribution has marginals $((\mathbf{P}^{k'})^\top \boldsymbol{\mu}, \boldsymbol{\pi})$ since $\mathbf{P}^\top \boldsymbol{\pi} = \boldsymbol{\pi}$, so it is a valid coupling, and preserves $i = i'$ with at least the same probability as before, so the total variation remains bounded by ϵ.

We now quantitatively strengthen this argument. To do so, we introduce the notation

$$\Delta(k) := \max_{i\in[d]} D_{\text{TV}}\left(\left(\mathbf{P}^k\right)^\top \mathbf{e}_i, \boldsymbol{\pi}\right), \; \overline{\Delta}(k) := \max_{(i,j)\in[d]\times[d]} D_{\text{TV}}\left(\left(\mathbf{P}^k\right)^\top \mathbf{e}_i, \left(\mathbf{P}^k\right)^\top \mathbf{e}_j\right). \tag{10}$$

Lemma 4. *Let $\mathbf{P} \in \mathbb{R}^{d\times d}_{\geq 0}$ be an irreducible and aperiodic transition matrix with stationary distribution $\boldsymbol{\pi}$. Following notation (10), $\Delta(k) \leq \overline{\Delta}(k) \leq 2\Delta(k)$ for all $k \in \mathbb{N}$.*

Proof. The upper bound on $\overline{\Delta}(k)$ is immediate because the total variation distance satisfies the triangle inequality, which can be seen by applying the second characterization in Fact 1.[12]

For the lower bound, we use the first characterization in Fact 1. Let $A \subseteq [d]$, and note that

$$\sum_{a\in A}\sum_{j\in[d]} \left[\mathbf{P}^k\right]_{ja} \boldsymbol{\pi}_j = \sum_{a\in A} \boldsymbol{\pi}_a,$$

by stationarity of $\boldsymbol{\pi}$ for $\mathbf{P}^k$. Therefore,

$$\begin{aligned}\left|\sum_{a\in A}\left[\left(\mathbf{P}^k\right)^\top \mathbf{e}_i\right]_a - \boldsymbol{\pi}_a\right| &= \left|\sum_{a\in A}\sum_{j\in[d]} \boldsymbol{\pi}_j \left(\left[\left(\mathbf{P}^k\right)^\top \mathbf{e}_i\right]_a - \left[\left(\mathbf{P}^k\right)^\top \mathbf{e}_j\right]_a\right)\right| \\ &\leq \sum_{j\in[d]} \boldsymbol{\pi}_j D_{\text{TV}}\left(\left(\mathbf{P}^k\right)^\top \mathbf{e}_i, \left(\mathbf{P}^k\right)^\top \mathbf{e}_j\right) \leq \overline{\Delta}(k).\end{aligned}$$

Taking the maximum over all possible A proves that $\Delta(k) \leq \overline{\Delta}(k)$. □

We also observe that $\overline{\Delta}$ decays at a linear rate.

Lemma 5. *Let $\mathbf{P} \in \mathbb{R}^{d\times d}_{\geq 0}$ be an irreducible and aperiodic transition matrix with stationary distribution $\boldsymbol{\pi}$. Following notation (10), $\overline{\Delta}(s + t) \leq \overline{\Delta}(s)\overline{\Delta}(t)$ for all $s, t \in \mathbb{N}$.*

[9] This does not qualitatively change the algorithm, but facilitates simpler analyses.

[10] Of course, the spectral characterization helps us achieve constant D_{TV} in the first place.

[11] By convexity of the ℓ_1 norm, the maximum is achieved by μ deterministically choosing a state.

[12] Given three distributions P, Q, R, first draw a sample from Q, and then draw conditional samples from $P \mid Q$, $R \mid Q$ from their joint distributions prescribed by the optimal couplings. This gives a coupling of (P, Q, R), where the total probability that the samples from (P, R) disagree is $\leq D_{\text{TV}}(P, Q) + D_{\text{TV}}(Q, R)$ by a union bound.

Proof. Let $\{\omega_k\}_{k\geq 0}$ be a random sequence of states in $[d]$ evolving according to $\mathbf{P}$ from $\omega_0 = i$, and similarly let $\{\omega'_k\}_{k\geq 0}$ evolve according to $\mathbf{P}$ from $\omega'_0 = j$. Note that for any $A \subseteq [d]$ and $s, t \in \mathbb{N}$,

$$\Pr[\omega_{s+t} \in A] = \mathbb{E}_{\omega_s}\left[\Pr[\omega_{s+t} \in A \mid \omega_s]\right],$$

by the law of iterated expectations. Therefore, letting γ be a coupling of ω_s and ω'_s achieving $\Pr_{(\omega_s,\omega'_s)\sim\gamma}[\omega_s \neq \omega'_s] \leq \overline{\Delta}(s)$, we can bound

$$\begin{aligned}\Pr[\omega_{s+t} \in A] - \Pr[\omega'_{s+t} \in A] &= \mathbb{E}_{\omega_s}\left[\Pr[\omega_{s+t} \in A \mid \omega_s]\right] - \mathbb{E}_{\omega'_s}\left[\Pr[\omega'_{s+t} \in A \mid \omega'_s]\right] \\ &\leq \mathbb{E}_{(\omega_s,\omega'_s)\sim\gamma}\left[\left|\Pr[\omega_{s+t} \in A \mid \omega_s] - \Pr[\omega'_{s+t} \in A \mid \omega'_s]\right| \cdot \mathbf{1}_{\omega_s \neq \omega'_s}\right] \\ &\leq \mathbb{E}_{(\omega_s,\omega'_s)\sim\gamma}\left[\overline{\Delta}(t) \cdot \mathbf{1}_{\omega_s \neq \omega'_s}\right] \leq \overline{\Delta}(t) \cdot \overline{\Delta}(s).\end{aligned}$$

In the second inequality, we viewed ω_{s+t} and ω'_{s+t} as the results of t-step random walks initialized at ω_s and ω'_s respectively. The conclusion follows by taking the maximum over all possible A. □

We conclude with our generic boosting of mixing times to high accuracy.

Corollary 1. *Let $\mathbf{P} \in \mathbb{R}^{d\times d}_{\geq 0}$ be an irreducible and aperiodic transition matrix with stationary distribution $\boldsymbol{\pi}$. For all $\epsilon < \frac{1}{4}$,*

$$\tau_{\text{mix}}(\epsilon) \leq \tau_{\text{mix}}\left(\frac{1}{4}\right)\left\lceil \log_2\left(\frac{1}{\epsilon}\right)\right\rceil.$$

Proof. Let $k := \tau_{\text{mix}}(\frac{1}{4})$, so by definition, $\Delta(k) \leq \frac{1}{4}$. Lemma 4 then shows $\overline{\Delta}(k) \leq \frac{1}{2}$, so $\overline{\Delta}(k \cdot \lceil \log_2(\frac{1}{\epsilon})\rceil) \leq \epsilon$ by repeatedly using Lemma 5. The claim follows by using Lemma 4 once more. □

Corollary 1 shows that for discrete Markov chains, the qualitative challenge is establishing a bound on $\tau_{\text{mix}}(\epsilon)$ for constant ϵ, because this also implies a bound for all smaller ϵ at logarithmic overhead. Interestingly, although Corollary 1 continues to hold for continuous sample spaces, the definition of $\tau_{\text{mix}}(\frac{1}{4})$ is often too stringent, since an initialization at a single point mass does not induce a finite relative density to the target distribution, making bounding the convergence rate challenging.[13] Nonetheless, we will see an analog of Corollary 1 in the continuous sampling setting as well.

3 Cheeger's inequality

Recall from Theorem 1 (and the discussion immediately following it) that for a reversible, irreducible, and aperiodic transition matrix $\mathbf{P}$ with the lazy form (8), the goal of establishing a mixing time bound is equivalent to lower bounding the spectral gap, $1 - \lambda_2$, where λ_2 is the second-largest eigenvalue of $\mathbf{P}$. In this section, we develop a combinatorial strategy to lower bound $1 - \lambda_2$, and hence establish rapid mixing. It will help to adopt a graph-based point of view, and let $G = (V, E, \mathbf{w})$ be an undirected weighted graph such that $\mathbf{P} = \mathbf{D}_G^{-1}\mathbf{A}_G$; recall this is without loss of generality from the discussion after Definition 3. In this notation, we let

$$\mathbf{N}_G := \mathbf{I}_V - \mathbf{D}_G^{-\frac{1}{2}}\mathbf{A}_G\mathbf{D}_G^{-\frac{1}{2}}. \tag{11}$$

By using the characterization (1) and rearranging, we note that we also have

$$\mathbf{N}_G = \mathbf{D}_G^{\frac{1}{2}}\left(\mathbf{I}_V - \mathbf{P}\right)\mathbf{D}_G^{-\frac{1}{2}}, \tag{12}$$

so the second-smallest eigenvalue of $\mathbf{N}_G$ and the second-largest eigenvalue of $\mathbf{P}$ sum to 1. Therefore, to establish a spectral gap, we can equivalently lower bound the second-smallest eigenvalue of $\mathbf{N}_G$ (note that $\mathbf{N}_G$ has a kernel, because $\mathbf{P}$ has an eigenvalue of 1). We also remark that

$$\mathbf{N}_G = \mathbf{D}_G^{-\frac{1}{2}}\mathbf{L}_G\mathbf{D}_G^{-\frac{1}{2}},$$

where $\mathbf{L}_G$ is the *Laplacian matrix* introduced in Definition 5, Part II. For this reason, $\mathbf{N}_G$ is often called a normalized Laplacian. For the remainder of the section, we let

$$\lambda_G^\star := \lambda_{d-1}\left(\mathbf{N}_G\right) \tag{13}$$

[13] For example, the analog of $\min_{i\in[d]} \boldsymbol{\pi}_i$ in Theorem 1 is always 0.

be the second-smallest eigenvalue of a normalized Laplacian, which per our discussion is also the spectral gap of an associated random walk $\mathbf{P}$, i.e., $\lambda^\star_G = 1 - \lambda_2(\mathbf{P})$.

To bound $\lambda^\star_G$, we use the following combinatorial notion.

Definition 6 (Conductance)**.** *For an undirected graph $G = (V, E, \mathbf{w})$, let $\deg(v) := \sum_{(u,v)\in E} \mathbf{w}_{(u,v)}$ denote the degree of vertex $v \in V$. For a subset $S \subseteq V$, let $\deg(S) := \sum_{v\in S} \deg(v)$, and let*

$$\partial(S) := \sum_{(u,v)\in(S\times V\setminus S)\cap E} \mathbf{w}_{(u,v)}$$

denote the total weight crossing the boundary from S to $V \setminus S$. We define the conductance *of G by*

$$\Phi_G := \min_{\substack{S\subseteq V \\ S\notin\{\emptyset,V\}}} \Phi_G(S), \text{ where } \Phi_G(S) := \frac{\partial(S)}{\min\left(\deg(S), \deg(V \setminus S)\right)}. \tag{14}$$

To demystify the definition (14), if $\deg(S) \le \deg(V \setminus S)$, then a direct calculation shows $\Phi_G(S)$ is the probability that a random walk, initialized within S according to the stationary distribution, leaves S in one step. Cheeger's inequality states that for all undirected graphs G,[14]

$$\frac{\lambda^\star_G}{2} \le \Phi_G \le \sqrt{2\lambda^\star_G}. \tag{15}$$

Intuitively, it is not surprising that Φ_G governs the mixing time of a random walk on G, because if Φ_G is large, then no set S is a "bottleneck" in the sense that we are reasonably-likely to leave S in a single step.[15] We focus on proving the upper bound in (15), because this is the useful inequality to establish a lower bound on the spectral gap $\lambda^\star_G$.[16] We mention that the bound is tight asymptotically, as witnessed by an unweighted path graph on d vertices; Hoeffding's inequality shows that it takes about d^2 steps of a random walk to reach one end from the other, and one can check that $\Phi_G = \Theta(\frac{1}{d})$ for this graph, witnessed by including the first $\frac{d}{2}$ vertices in S.

Theorem 2 (Cheeger's inequality)**.** *Let $G = (V, E, \mathbf{w})$ be a connected undirected graph. Following the notation (13), (14), $\Phi_G \le \sqrt{2\lambda^\star_G}$.*

Proof. Let $\mathbf{D}_G = \mathbf{diag}\left(\mathbf{\Delta}\right)$, where $\mathbf{\Delta}_v := \deg(v)$ for all $v \in V$. Recall that $\mathbf{1}_V$ is the leading right eigenvector of $\mathbf{P}$, so following the characterization (12), the kernel of $\mathbf{N}_G$ is spanned by $\sqrt{\mathbf{\Delta}}$ where $\sqrt{\cdot}$ is applied entrywise. Hence, we can characterize $\lambda^\star_G$, the second-smallest eigenvalue of $\mathbf{N}_G$, by

$$\lambda^\star_G = \min_{\substack{\mathbf{x}\in\mathbb{R}^V \\ \mathbf{x}\perp\sqrt{\mathbf{\Delta}}}} \frac{\mathbf{x}^\top \mathbf{N}_G \mathbf{x}}{\|\mathbf{x}\|_2^2} = \min_{\substack{\mathbf{y}\in\mathbb{R}^V \\ \mathbf{y}\perp\mathbf{\Delta}}} \frac{\mathbf{y}^\top \mathbf{L}_G \mathbf{y}}{\mathbf{y}^\top \mathbf{D}_G \mathbf{y}}, \tag{16}$$

where $\mathbf{L}_G := \mathbf{D}_G - \mathbf{A}_G$ is the Laplacian matrix of G. The first equality above used the min-max eigenvalue theorem (Proposition 2, Part VI), and the second changed variables $\sqrt{\mathbf{\Delta}} \circ \mathbf{y} \leftarrow \mathbf{x}$, where $\circ$ denotes entrywise multiplication. We claim that for any nonzero vector $\mathbf{y} \perp \mathbf{\Delta}$, we can produce a random set $S_t \subseteq V$, from a family of sets S_t parameterized by t, such that

$$\mathbb{E}_t\left[\partial(S_t)\right] \le \sqrt{2\rho} \cdot \mathbb{E}_t\left[\min\left(\deg\left(S_t\right), \deg\left(V \setminus S_t\right)\right)\right], \text{ where } \rho := \frac{\mathbf{y}^\top \mathbf{L}_G \mathbf{y}}{\mathbf{y}^\top \mathbf{D}_G \mathbf{y}}. \tag{17}$$

Assuming (17), for some realization S_t, we must have $\partial(S_t) \le \sqrt{2\rho} \cdot \min(\deg(S_t), \deg(V \setminus S_t))$, else this would contradict (17) by averaging over t. The realization S_t satisfying this inequality then certifies $\Phi_G \le \sqrt{2\rho}$, and the claim $\Phi_G \le \sqrt{2\lambda^\star_G}$ follows by minimizing ρ over $\mathbf{y}$, using (16). For the remainder of the proof, we fix a nonzero $\mathbf{y} \in \mathbb{R}^V$ with $\mathbf{y} \perp \Delta$, and establish (17). We also assume

[14] Cheeger's inequality was proved originally in the context of certain expansion properties of manifolds [Che69], and the proof is actually somewhat simpler in this continuous setting. For an exposition on the relationship between the continuous and discrete proofs of Cheeger's inequality, we refer the reader to the blog post [Tre13].

[15] In contrast, for a classic example of a graph G that does have a very clear bottleneck, consider a "dumbbell graph" consisting of two complete graphs on $\frac{1}{2}|V|$ vertices each joined by a single edge. A random walk on a dumbbell graph is unlikely to cross between the two complete graphs, which is reflected in its poor conductance.

[16] Tragically, the lower bound in (15) is actually the easier bound to prove; see Chapter 21, [Spi19].

without loss of generality that $\mathbf{y}$ has coordinates indexed by $i \in [d]$ sorted in increasing order (for $d := |V|$), by relabeling vertices as necessary. We now construct S_t after two modification steps.

In the first step, we observe that for any constant shift $c \in \mathbb{R}$ and $\mathbf{z} := \mathbf{y} + c\mathbf{1}_V$, we have $\mathbf{y}^\top \mathbf{L}_G \mathbf{y} = \mathbf{z}^\top \mathbf{L}_G \mathbf{z}$, because $\mathbf{1}_V$ is in the kernel of $\mathbf{L}_G$.[17] Furthermore, we have for any $c \in \mathbb{R}$,

$$\mathbf{z}^\top \mathbf{D}_G \mathbf{z} = \sum_{i \in [d]} \boldsymbol{\Delta}_i \left(\mathbf{y}_i + c\right)^2 \geq \sum_{i \in [d]} \boldsymbol{\Delta}_i \mathbf{y}_i^2 = \mathbf{y}^\top \mathbf{D}_G \mathbf{y},$$

where the only inequality used that the left-hand side is a quadratic in c, whose first-order optimality condition $2\sum_{i \in [d]} \boldsymbol{\Delta}_i \mathbf{y}_i = 2\sum_{i\in[d]} \boldsymbol{\Delta}_i c$ implies $c = 0$ is the minimizer, since $\mathbf{y} \perp \boldsymbol{\Delta}$. Therefore, any shift $c \in \mathbb{R}$ will yield $\mathbf{z} = \mathbf{y} + c\mathbf{1}_V$ with

$$\frac{\mathbf{z}^\top \mathbf{L}_G \mathbf{z}}{\mathbf{z}^\top \mathbf{D}_G \mathbf{z}} \leq \rho, \tag{18}$$

because the numerator stayed the same and the denominator did not decrease.

In the second step, we choose c. Let $j \in [d]$ be the largest vertex index such that $\sum_{i \in [j]} \boldsymbol{\Delta}_i \leq \frac{1}{2}\deg(V)$. We pick c as a function of $\mathbf{y}$, so that $\mathbf{z}_{j+1} = 0$. We also normalize $\mathbf{z}$ so that $\mathbf{z}_1^2 + \mathbf{z}_d^2 = 1$, which is without loss of generality since scaling by a constant does not affect the ratio (18).

Now, we define the family S_t to be the threshold family, where for a given value of t,

$$S_t := \left\{i \in [d] \mid \mathbf{z}_i \leq t\right\}.$$

We choose t according to the density $2|t|$, over $t \in [\mathbf{z}_1, \mathbf{z}_d]$. Note that for any $[a, b] \subseteq [\mathbf{z}_1, \mathbf{z}_d]$,

$$\Pr[t \in [a, b]] = \int_a^b 2|t|\mathrm{d}t = \mathrm{sign}(b) \cdot b^2 - \mathrm{sign}(a) \cdot a^2, \tag{19}$$

so that $\Pr[t \in [\mathbf{z}_1, \mathbf{z}_d]] = \mathbf{z}_1^2 + \mathbf{z}_d^2 = 1$ by assumption, so this is a valid probability distribution. To finish our proof of (17), it is enough to bound $\mathbb{E}_t[\partial(S_t)]$ and $\mathbb{E}_t[\min(\deg(S_t), \deg(V \setminus S_t))]$.

We start with the latter. The way we chose $\mathbf{z}$ (via the shift c in the second step above) is so that if $t > 0$, $\deg(S_t) \leq \deg(V \setminus S_t)$, and this inequality is reversed if $t > 0$.[18] Thus,

$$\begin{aligned}\mathbb{E}_t\left[\min(\deg(S_t), \deg(V \setminus S_t))\right] &= \sum_{\substack{i \in [d] \\ \mathbf{z}_i < 0}} \boldsymbol{\Delta}_i \cdot \Pr\left[\mathbf{z}_i \leq t \leq 0\right] + \sum_{\substack{i \in [d] \\ \mathbf{z}_i \geq 0}} \boldsymbol{\Delta}_i \cdot \Pr\left[\mathbf{z}_i \geq t \geq 0\right] \\ &= \sum_{i \in [d]} \boldsymbol{\Delta}_i \cdot \mathbf{z}_i^2 = \mathbf{z}^\top \mathbf{D}_G \mathbf{z},\end{aligned} \tag{20}$$

where we used the formula (19) twice. Next, before bounding $\mathbb{E}_t[\partial(S_t)]$, we make a brief observation which follows by casework on signs: for any $a, b \in \mathbb{R}$,

$$\mathrm{sign}(a) \cdot a^2 - \mathrm{sign}(b) \cdot b^2 \leq |a - b| \left(|a| + |b|\right). \tag{21}$$

Therefore,

$$\begin{aligned}\mathbb{E}_t\left[\partial(S_t)\right] &= \sum_{\substack{(i,j) \in E \\ i<j}} \mathbf{w}_{(i,j)} \cdot \Pr\left[\mathbf{z}_i \leq t \leq \mathbf{z}_j\right] \\ &= \sum_{\substack{(i,j) \in E \\ i<j}} \mathbf{w}_{(i,j)} \cdot \left(\mathrm{sign}(\mathbf{z}_i) \cdot \mathbf{z}_i^2 - \mathrm{sign}(\mathbf{z}_j) \cdot \mathbf{z}_j^2\right) \\ &\leq \sum_{\substack{(i,j) \in E \\ i<j}} \mathbf{w}_{(i,j)} \cdot |\mathbf{z}_i - \mathbf{z}_j| \left(|\mathbf{z}_i| + |\mathbf{z}_j|\right) \\ &\leq \sqrt{\sum_{\substack{(i,j) \in E \\ i<j}} \mathbf{w}_{(i,j)} \cdot \left(\mathbf{z}_i - \mathbf{z}_j\right)^2} \cdot \sqrt{\sum_{\substack{(i,j) \in E \\ i<j}} \mathbf{w}_{(i,j)} \cdot \left(|\mathbf{z}_i| + |\mathbf{z}_j|\right)^2} \\ &\leq \sqrt{\mathbf{z}^\top \mathbf{L}_G \mathbf{z}} \cdot \sqrt{2\mathbf{z}^\top \mathbf{D}_G \mathbf{z}}.\end{aligned} \tag{22}$$

[17] This can also be directly seen from the Laplacian quadratic form computed in Lemma 10, Part II.

[18] The event $t = 0$ is a measure-zero event, so we ignore it in calculations.

The first inequality applied (21) edgewise, the second was the Cauchy-Schwarz inequality, and the last used the definition of $\mathbf{L}_G$ (see Lemma 10, Part II) and the scalar inequality $(a+b)^2 \le 2a^2+2b^2$. Putting together (18), (20), and (22), we finally have proved (17), which establishes our claim:

$$\mathbb{E}_t\left[\partial(S_t)\right] \le \sqrt{\mathbf{z}^\top \mathbf{L}_G \mathbf{z}} \cdot \sqrt{2\mathbf{z}^\top \mathbf{D}_G \mathbf{z}} \le \sqrt{2\rho} \cdot \mathbf{z}^\top \mathbf{D}_G \mathbf{z} = \sqrt{2\rho} \cdot \mathbb{E}_t\left[\min\left(\deg(S_t), \deg(V \setminus S_t)\right)\right].$$

□

By combining Theorems 1 and 2, we have shown the following.

Corollary 2. *Let $\mathbf{P} \in \mathbb{R}^{d\times d}_{\ge 0}$ be a reversible, irreducible, aperiodic and lazy transition matrix with stationary distribution $\boldsymbol{\pi}$, and let $G = (V, E, \mathbf{w})$ be an weighted undirected graph such that (1) holds. Then for any $\boldsymbol{\mu} \in \Delta^d$, we have*

$$D_{\mathrm{TV}}\left(\left(\mathbf{P}^k\right)^\top \boldsymbol{\mu}, \boldsymbol{\pi}\right) \le \epsilon, \text{ if } k \ge \frac{2}{\Phi_G^2} \cdot \log\left(\frac{1}{2\epsilon \min_{i\in[d]} \boldsymbol{\pi}_i}\right).$$

One interesting consequence of the proof of Theorem 1 is a practical strategy for partitioning a graph G into two well-connected and large components, known as a *sparse cut*. Specifically, Theorem 1 takes a modification of an eigenvector of an appropriate matrix (e.g., $\mathbf{N}_G$) and constructs a set S_t by thresholding the coordinates of y at a cutoff value t. The resulting cuts $(S_t, V \setminus S_t)$ in G are known as sweep cuts. Spectral partitioning in practice often uses variations of this strategy, which was explicitly suggested by [SM00] and analyzed in [KVV04]. More generally, multiway generalizations of Cheeger's inequality and resulting spectral partitioning methods, i.e., partitions which allow for > 2 pieces of a graph, were provided by [LGT12, KLL+13].

Corollary 2 shows that undirected graphs G which have good conductance behavior, i.e., large Φ_G, induce random walks which mix quickly. We call a graph G a *ϕ-expander* if $\Phi_G \ge \phi$, which means that every partition of V into $S, V \setminus S$ has at least a ϕ fraction of edges of the smaller side's degree crossing between S and $V \setminus S$. Intuitively, these are exactly the types of graphs which have no bottlenecks, as discussed before Theorem 1. For example, the Laplacians of expander graphs are well-conditioned as a result of Theorem 2, which implies that the leverage scores (Definition 2, Part VIII) of the graph edges are all bounded; this turns out to have important consequences for the fault tolerance of expander graphs, since no single edge is too important. Moreover, expanders have a number of additional extremely useful properties which have been exploited throughout theoretical computer science, such as coding theory [SS96] and pseudorandomness [Vad12].

The breakthrough work of [ST14] launched a revolution in modern graph algorithms by quantifying the realization that all graphs can in fact be decomposed into a small number of disjoint expanders, and a small fraction of crossing edges which can then be recursed upon. We provide a simple example of such an *expander decomposition* for the reader's interest.

Proposition 2. *Let $G = (V, E, \mathbf{1}_E)$ be an undirected, unweighted graph with $n := |V|$, and let $\phi \in (0, 1)$. There exists a partitioning of V into disjoint subsets $\{V_i\}_{i\in[k]}$ such that $\bigcup_{i\in[k]} V_i = V$, where every induced subgraph[19] $G[V_i]$ is a ϕ-expander, and*

$$|\{e = (u, v) \in E \mid u \in V_i, v \in V_j, i \ne j\}| \le 2\phi \log_2(n)|E|.$$

Proof. Consider the algorithm which maintains a list L of subsets of the vertices of G, initialized to a single element V, such that the elements of L always partition V into disjoint subsets. Moreover, any time there exists $U \in L$ such that $G[U]$ is not a ϕ-expander, i.e., there exists a partition U_1, U_2 such that $U_1 \cup U_2 = U$, $U_1 \cap U_2 = \emptyset$, and the number of edges between U_1 and U_2 in $G[U]$ is at most $\phi \cdot \min(\deg(U_1), \deg(U_2))$ (where degrees are measured with respect to edges in $G[U]$), we delete U from L and add U_1, U_2. This algorithm terminates after at most n steps, since there are only n vertices, and at termination every piece is a ϕ-expander by definition.

It remains to bound the total number of edges cut by this process. We do so by a charging argument, where every vertex in G is initialized with a potential of 0. Every time we partition U into two pieces U_1, U_2 with $\deg(U_1) \le \deg(U_2)$ (where again, degrees are measured with respect to

[19] For $U \subseteq V$, we let $G[U]$ denote the induced subgraph onto U, i.e., the subgraph of G which keeps all edges between vertices in U with the same weight as in G, and deletes all other edges and vertices.

$G[U]$), we add $\phi\deg(u)$ to the potential of each $u \in U_1$ on the smaller side, so the total potential of all vertices always upper bounds the number of cut edges by definition. At the end of the algorithm, each vertex can only have a maximum potential of $\phi\log_2(n)\deg(u)$, since it can only be on the smaller side of a cut at most $\log_2(n)$ times. Summing over all vertices yields the claim. □

For example, Proposition 2 shows that every undirected, unweighted graph can be partitioned into $\Omega(\frac{1}{\log n})$-expanders and a small constant fraction of edges which cross between partition pieces.

As stated, Proposition 2 is not known to be implementable in polynomial time, because the *sparsest cut* problem (finding the value of Φ_G or a set S witnessing it) is NP-hard, and we do not know any polynomial-time constant factor approximations. Nonetheless, by exploiting variants of flow-cut duality, a line of research has obtained efficient algorithms for computing polylogarithmic factor approximations to the sparsest cut in nearly-linear time, as well as cuts achieving various tradeoffs along the approximation factor-runtime tradeoff curve [LR99, ARV09, KRV09, OSVV08, She09]. By using local variants of these sparsest cut approximation algorithms with runtime proportional only to the size of the partition piece being isolated, we now have expander decomposition algorithms which run in nearly-linear time and achieve the guarantee in Proposition 2 up to a polylogarithmic factor, even for weighted graphs. The first such example can be found in [SW19], building upon a weaker variant of this primitive from [ST14].

More generally, variants of the expander decomposition strategy that partition a potentially poorly-behaved graph into pieces having significantly stronger local properties (e.g., good conductance), and then recurse on the remainder, have led to many breakthroughs in classical problems in recent years. We refer the reader to the excellent course [Sar21] for more on these techniques, as well as the thesis [Li21] for a unified perspective on modern applications of the decomposition toolbox.

Source material

Portions of this lecture are based on reference material in [AF02, LPW09, Spi19], as well as the author's own experience working in the field.

References

[AF02] David Aldous and James Allen Fill. *Reversible Markov Chains and Random Walks on Graphs.* 2002.

[ARV09] Sanjeev Arora, Satish Rao, and Umesh V. Vazirani. Expander flows, geometric embeddings and graph partitioning. *J. ACM*, 56(2):5:1–5:37, 2009.

[Che69] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In *Proceedings of the Princeton conference in honor of Professor S. Bochner*, pages 195–199, 1969.

[Fro12] Georg Frobenius. Ueber matrizen aus nicht negativen elementen. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, pages 456–477, 1912.

[Has70] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[KLL$^+$13] Tsz Chiu Kwok, Lap Chi Lau, Yin Tat Lee, Shayan Oveis Gharan, and Luca Trevisan. Improved cheeger's inequality: analysis of spectral partitioning algorithms through higher order spectral gap. In *Symposium on Theory of Computing Conference, STOC'13*, pages 11–20. ACM, 2013.

[KRV09] Rohit Khandekar, Satish Rao, and Umesh V. Vazirani. Graph partitioning using single commodity flows. *J. ACM*, 56(4):19:1–19:15, 2009.

[KVV04] Ravi Kannan, Santosh S. Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.

[LGT12] James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. Multi-way spectral partitioning and higher-order cheeger inequalities. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012*, pages 1117–1130. ACM, 2012.

[Li21] Jason Li. *Preconditioning and Locality in Algorithm Design.* PhD thesis, Carnegie Mellon University, 2021.

[LPW09] David Asher Levin, Yuval Peres, and Elizabeth Wilmer. *Markov Chains and Mixing Times.* American Mathematical Society, 2009.

[LR99] Frank Thomson Leighton and Satish Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *J. ACM*, 46(6):787–832, 1999.

[MRR$^+$53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

[OSVV08] Lorenzo Orecchia, Leonard J. Schulman, Umesh V. Vazirani, and Nisheeth K. Vishnoi. On partitioning graphs via single commodity flows. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, 2008*, pages 461–470. ACM, 2008.

[Per07] Oskar Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263, 1907.

[Sar21] Thatchaphol Saranurak. Expanders and fast graph algorithms. https://sites.google.com/site/thsaranurak/teaching/Expander, 2021. Accessed: 2024-02-29.

[She09] Jonah Sherman. Breaking the multicommodity flow barrier for o(vlog n)-approximations to sparsest cut. In *50th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2009*, pages 363–372. IEEE Computer Society, 2009.

[SM00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 22(8):888–905, 2000.

[Spi19] Daniel A. Spielman. *Spectral and Algebraic Graph Theory*. 2019.

[SS96] Michael Sipser and Daniel A. Spielman. Expander codes. *IEEE Trans. Inf. Theory*, 42(6):1710–1722, 1996.

[ST14] Daniel A. Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Anal. Appl.*, 35(3):835–885, 2014.

[SW19] Thatchaphol Saranurak and Di Wang. Expander decomposition and pruning: Faster, stronger, and simpler. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, pages 2616–2635. SIAM, 2019.

[Tre13] Luca Trevisan. The cheeger inequality in manifolds. https://lucatrevisan.wordpress.com/2013/03/20/the-cheeger-inequality-in-manifolds/, 2013. Accessed: 2024-02-28.

[Vad12] Salil P. Vadhan. Pseudorandomness. *Found. Trends Theor. Comput. Sci.*, 7(1-3):1–336, 2012.