Electron. J. Probab. **0** (2024), article no. 0, 1–29.
ISSN: 1083-6489 https://doi.org/10.1214/YY-TN

Entropy Contractions in Markov Chains: Half-Step, Full-Step and Continuous-Time

Pietro Caputo* Zongchen Chen† Yuzhou Gu‡ Yury Polyanskiy§

Abstract

This paper considers the speed of convergence (mixing) of a finite Markov kernel P with respect to the Kullback-Leibler divergence (entropy). Given a Markov kernel one defines either a discrete-time Markov chain (with the n-step transition kernel given by the matrix power P^n) or a continuous-time Markov process (with the time-t transition kernel given by $e^{t(P-\mathrm{Id})}$). The contraction of entropy for $n = 1$ or $t = 0+$ are characterized by the famous functional inequalities, the *strong data processing inequality (SDPI)* and the *modified log-Sobolev inequality (MLSI)*, respectively. When $P = KK^*$ is written as the product of a kernel and its adjoint, one could also consider the "half-step" contraction, which is the SDPI for K, while the "full-step" contraction refers to the SDPI for P. The work [17] claimed that these contraction coefficients (half-step, full-step, and continuous-time) are generally comparable, that is their ratio is bounded from above and below by two absolute constants. We disprove this and related conjectures by working out a number of different counterexamples. In particular, we construct (a) a continuous-time Markov process that contracts arbitrarily faster than its discrete-time counterpart; and (b) a kernel P such that P^{m+1} contracts arbitrarily better than P^m. Hence, our main conclusion is that the four standard inequalities comparing five common notions of entropy and variance contraction are generally not improvable (even in the subclass of factorizable Markov chains).

In the process of analyzing the counterexamples, we survey and sharpen the tools for bounding the contraction coefficients and characterize properties of extremizers of the respective functional inequalities. For example, we show that while the MLSI extremizers always have full support (unless the MLSI constant equals twice the spectral gap), the SDPI extremizers can have partial support. As our examples range from Bernoulli-Laplace model, random walks on graphs, to birth-death chains, the paper is also intended as a survey on computing MLSI, SDPI and other constants for these types of commonly occurring Markov chains.

Keywords: entropy contraction; logarithmic Sobolev Inequalities; strong data processing inequalities.

*Università Roma Tre, Italy. E-mail: pietro.caputo@uniroma3.it
†Georgia Institute of Technology, United States of America. E-mail: chenzongchen@gatech.edu
‡New York University, United States of America. E-mail: yuzhougu@nyu.edu
§Massachusetts Institute of Technology, United States of America. E-mail: yp@mit.edu

MSC2020 subject classifications: 60J10; 60J27.
Submitted to EJP on October 1, 2024, final version accepted on .
Supersedes arXiv:2409.07689.

1 Introduction

Markov chains are widely used in almost all areas of science and engineering, in the form of MCMC averaging in numerical analysis, approximation algorithms in computer science, generative modeling in artificial intelligence, and so on. Perhaps the most important problem in the study of Markov chains is to understand their equilibration properties. One example of such property is the mixing time, characterizing the time it takes for the marginal distribution to come close to the stationary one, as measured by some statistical distance: the total variation (most commonly) or Wasserstein distance, χ^2 or Kullback-Leibler (KL) divergence.

In this paper, we focus on Markov chains on a finite state space. There are two common ways to define or implement Markov processes. In a discrete-time Markov chain, the state is updated at every discrete time $t \in \mathbb{Z}_{\geq 1}$; meanwhile, in a continuous-time Markov chain, the update times are distributed as a Poisson point process on the positive real axis $\mathbb{R}_{\geq 0}$. For example, if the single-step kernel for the former is given by a row-stochastic matrix P then the corresponding continuous-time chain has kernel $T_t = e^{t(P-\mathrm{Id})}$ at any time $t \geq 0$. Due to concentration of the number of updates in the time interval $[0, t]$, both versions are known to have almost equal mixing times in *total variation* (e.g., [24, Theorem 20.3]). Thus, for the study of total-variation mixing time probabilists are free to switch between the discrete and continuous times as convenient.

One common approach (e.g., [18]) to show rapid mixing of Markov processes is to prove that they "make progress" step-wise in terms of some f-divergence, most commonly the χ^2 or the KL divergence. For discrete-time Markov chains, this means that the f-divergence to the target distribution decreases by a certain factor in every step, which can be described by the contraction coefficient of the associated Markov kernel. For continuous-time chains, this corresponds to that the derivative of f-divergence with respect to time is suitably bounded away from zero from above, and hence the f-divergence to the target distribution decreases at a certain speed. For the KL divergence, the respective contraction inequalities are known as the *strong data processing inequality (SDPI)* and the *modified log-Sobolev inequality (MLSI)*, respectively for discrete-time and continuous-time. See Eqs. (1.6) and (1.14) below for formal definitions.

For a large family of Markov chains, including the Glauber dynamics on product spaces and random walks on high-dimensional simplices (e.g., [23]), the transition kernel P consists of two stages, for which $P = KK^*$ is written as the product of a kernel and its adjoint. The first stage (K) is often described as the downward, forward, or noising move, and the second stage (K^*) as the upward, backward, or denoising move. Recent success on analyzing such two-stage processes (e.g., [3, 16, 15]) considers the decay of f-divergence in a single stage K or K^*, which for many specific problems turns out to be easier to handle. For such chains it is then natural to define *a half-step* contraction coefficient (corresponding to application of K only) and ask how it compares with full-step, multi-step and continuous-time ones. We note that a much earlier work [20] also studied such factorizable Markov kernels $P = KK^*$ and called P a *multiplicative reversiblization* of the kernel K. The main purpose of [20] is to use P to investigate the convergence of a nonreversible Markov chain K.

Given the equivalence between continuous-time and discrete-time mixing times, one naturally expects similar equivalence between the contraction coefficients. This turns out to be true for the χ^2 contraction ([32, Remark 4.2]). Thus, it was not surprising

when the work [17] claimed to show such equivalence for the KL divergence contraction as well (that is showing that the ratio of the MLSI and SDPI constants is universally bounded). Specifically, they claimed an equivalence (up to universal multiplicative factors) between the MLSI and the *half-step SDPI*, which implies in particular that the MLSI and the *full-step SDPI* (i.e., continuous-time and discrete-time entropy contraction) are equivalent. The present paper originated from us discovering a gap in their proof (see Section 3.6) and realizing that their claim cannot hold true because the ratio between the MLSI constant and the half-step SPDI constant can be arbitrarily large for the random transposition model (Example 3.15). However, the question of whether the MLSI and the full-step SDPI are equivalent remained open.

We answer this question by presenting an example (Example 3.11) separating the MLSI and the full-step SDPI. That is, there exist cases where the contraction rate of the discrete-time chain can be much slower than that of the continuous-time one. The example is adapted from Münch's counterexample to the Peres-Tetali conjecture [29], although there seems to be no direct relationship between the properties used here and op. cit.

The separation between continuous-time and discrete-time entropy contraction leads us to consider a related question of comparing the entropy contraction of the m-step kernel P^m versus the $(m+1)$-step one P^{m+1}. If a separation is possible, then it would explain how a continuous-time chain (which averages over many P^m's) could have better convergence properties at finite time than the corresponding discrete-time chain (see discussions at the end of Section 3.2). It turns out that a counterexample is again possible (see Example 3.9 below).

To avoid trivial counterexamples, we restrict our attention to *factorizable* kernels, which are kernels that can be written in the form $P = KK^*$. As discussed above, this is a natural class of Markov kernels to study. All our examples are factorizable.

In all, the main purpose of this paper is to provide a self-contained and thorough introduction to all common notions of relative entropy decay for finite-state Markov chains, for both continuous-time and discrete-time versions. We summarize known comparisons among these notions (including the recent half-step contraction) and we give examples demonstrating that in all cases where comparisons are not available there exist counterexamples, in the sense that the ratio can be arbitrarily large. Along the way, we correct several misstatements that appeared in previous works, and show how to get sharp upper and lower bounds for these coefficients and study the extremizers in the respective functional inequalities.

Organization. This paper's content is succinctly summarized in three tables: Table 1 gives definitions of five common contraction coefficients, Table 2 lists all known comparison inequalities along with (new) counterexamples for the missing comparisons, and Table 3 summarizes the list of counterexamples. The rest of the introduction defines all notions rigorously and recalls the standard comparison chain Eq. (1.28):

$$\rho \le \alpha \le \delta \le \rho_0 \le 2\lambda\,.$$

Then, after introducing factorizable Markov kernels ($P = KK^*$) in Section 2, Section 3 shows that every inequality above can have arbitrarily high ratio, even when restricted to factorizable kernels. Section 4 concludes with some results on the properties of extremizers of functional inequalities.

Notation. Throughout the paper, let $\mathcal{X}$ be a finite set, and $P : \mathcal{X} \to \mathcal{X}$ be a Markov kernel[1] with an invariant distribution $\pi > 0$ fully supported on $\mathcal{X}$. Assume that (π, P) is

[1] A Markov kernel $K : \mathcal{X} \to \mathcal{Y}$ is a collection $(K(x,\cdot))_{x\in\mathcal{X}}$ where each $K(x,\cdot)$ is a distribution on $\mathcal{Y}$. Our notation for Markov kernel is commonly used in information theory (e.g., [31]).

reversible, i.e., $\pi(x)P(x,y) = \pi(y)P(y,x)$ for all $x, y \in \mathcal{X}$. For a Markov kernel $K : \mathcal{X} \to \mathcal{Y}$ and a distribution π on $\mathcal{X}$, we define the reverse channel $K^*_\pi : \mathcal{Y} \to \mathcal{X}$ as

$$K^*_\pi(y,x) = \begin{cases} \frac{\pi(x)K(x,y)}{(\pi K)(y)}, & \text{if } (\pi K)(y) > 0, \\ \pi(x), & \text{otherwise.} \end{cases} \tag{1.1}$$

Note that π is an invariant distribution of KK^*_π and (π, KK^*_π) is reversible. Furthermore, for any $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, we have

$$\pi[fKg] = (\pi K)[gK^*_\pi f]. \tag{1.2}$$

For two sequences $\{f_n\}_n$, $\{g_n\}_n$ of non-negative real numbers, we say $f_n = O(g_n)$ if there exists $C > 0$ such that $f_n \le Cg_n$ for all large enough n; we say $f_n = \Omega(g_n)$ if there exists $c > 0$ such that $f_n \ge cg_n$ for all large enough n; we say $f_n = \Theta(g_n)$ if $f_n = O(g_n)$ and $f_n = \Omega(g_n)$.

1.1 Continuous-time contraction notions

For $f, g : \mathcal{X} \to \mathbb{R}$, define the Dirichlet form as

$$\mathcal{E}_{\pi,P}(f,g) = -\pi[fLg] \tag{1.3}$$

where $L = P - I$ is the Markov generator.

Let $\mathcal{P}(\mathcal{X})$ denote the space of distributions on $\mathcal{X}$. Recall that we assume $\pi \in \mathcal{P}(\mathcal{X})$ is fully supported. For any $\nu \in \mathcal{P}(\mathcal{X})$, the Radon–Nikodym derivative $f = \frac{d\nu}{d\pi}$ is thus always well-defined.

The *Poincaré constant* (also called the *spectral gap*) $\lambda = \lambda(\pi, P)$ is the largest number such that

$$\lambda \operatorname{Var}_\pi(f) \le \mathcal{E}_{\pi,P}(f,f), \qquad \forall f : \mathcal{X} \to \mathbb{R}, \tag{1.4}$$

where $\operatorname{Var}_\pi(f) := \pi[f^2] - (\pi[f])^2$ is the variance of f under π. When $f = \frac{d\nu}{d\pi}$ for some $\nu \in \mathcal{P}(\mathcal{X})$, we have $\operatorname{Var}_\pi(f) = \chi^2(\nu\|\pi)$ where $\chi^2(\cdot\|\cdot)$ stands for the χ^2-divergence. Eq. (1.4) is called the *Poincaré inequality*.

The *log-Sobolev constant* (LSC) $\rho = \rho(\pi, P)$ is the largest number such that

$$\rho \operatorname{Ent}_\pi(f) \le \mathcal{E}_{\pi,P}(\sqrt{f}, \sqrt{f}), \qquad \forall f : \mathcal{X} \to \mathbb{R}_{\ge 0}, \tag{1.5}$$

where $\operatorname{Ent}_\pi(f) := \pi[f \log f] - \pi[f] \log(\pi[f])$ is the entropy of f with the convention that $0 \log 0 = 0$. When $f = \frac{d\nu}{d\pi}$ for some $\nu \in \mathcal{P}(\mathcal{X})$, we have $\operatorname{Ent}_\pi(f) = D(\nu\|\pi)$ where $D(\cdot\|\cdot)$ stands for the Kullback-Leibler (KL) divergence. Eq. (1.5) is called the *log-Sobolev inequality* (LSI).

The *modified log-Sobolev constant* (MLSC) $\rho_0 = \rho_0(\pi, P)$ is the largest number such that

$$\rho_0 \operatorname{Ent}_\pi(f) \le \mathcal{E}_{\pi,P}(f, \log f), \qquad \forall f : \mathcal{X} \to \mathbb{R}_{\ge 0}. \tag{1.6}$$

Eq. (1.6) is called the *modified log-Sobolev inequality* (MLSI).

For reversible (π, P), we always have ([18, 8])

$$4\rho \le \rho_0 \le 2\lambda. \tag{1.7}$$

The constants ρ, ρ_0, and λ represent the contraction ability of the continuous-time Markov chain (also known as the Markov semigroup) $(T_t)_{t\ge 0}$, where $T_t = \exp(-tL)$.[2]

[2] In this work we focus on continuous-time Markov chains of this type and do not consider more general continuous-time Markov chains.

These constants are properties of the Markov generator L (which can be arbitrarily scaled), rather than the Markov kernel P.

Let ν be a distribution on $\mathcal{X}$. Define $\nu_t = \nu T_t$ to be the distribution of the Markov chain at time $t \geq 0$ when initialized at ν, and define $f_t = \frac{d\nu_t}{d\pi}$ to be the relative density. A direct computation yields

$$\frac{d}{dt} \operatorname{Var}_\pi(f_t) = -2\mathcal{E}_{\pi,P}(f_t, f_t), \tag{1.8}$$

$$\frac{d}{dt} \operatorname{Ent}_\pi(f_t) = -\mathcal{E}_{\pi,P}(f_t, \log f_t). \tag{1.9}$$

Therefore the Poincaré inequality and the modified log-Sobolev inequality can be equivalently stated as

$$\frac{d}{dt} \operatorname{Var}_\pi(f_t) \leq -2\lambda \operatorname{Var}_\pi(f_t), \tag{1.10}$$

$$\frac{d}{dt} \operatorname{Ent}_\pi(f_t) \leq -\rho_0 \operatorname{Ent}_\pi(f_t). \tag{1.11}$$

Eqs. (1.10) and (1.11) can also be understood as alternative definitions for λ and ρ_0, from which one immediately obtains

$$\operatorname{Var}_\pi(f_t) \leq \exp(-2\lambda t) \operatorname{Var}_\pi(f_0), \tag{1.12}$$

$$\operatorname{Ent}_\pi(f_t) \leq \exp(-\rho_0 t) \operatorname{Ent}_\pi(f_0). \tag{1.13}$$

We remark that the log-Sobolev inequality is equivalent to hypercontractivity by [18]. Furthermore, [7] shows that the log-Sobolev inequality is equivalent (up to a constant factor) to the Poincaré inequality on an Orlicz space. In this work we consider only the Poincaré inequality on the L^2 space.

We refer the reader to [18, 8] for more discussions on (modified) log-Sobolev constants.

1.2 Discrete-time contraction notions

Another class of contraction notions comes from contraction coefficients for f-divergences. For two distributions ν and π on the same state space satisfying $\nu \ll \pi$, the f-divergence between ν and π is defined as $D_f(\nu\|\pi) = \int f\left(\frac{d\nu}{d\pi}\right) d\pi$, where $f : [0, +\infty) \to (-\infty, +\infty]$ is a convex function satisfying $f(x) < \infty$ for all $x > 0$, $f(1) = 0$, and $f(0) = \lim_{t\to 0^+} f(t)$. We refer the reader to [31, Chapter 7] for an introduction to f-divergences. Let $K : \mathcal{X} \to \mathcal{Y}$ be a Markov kernel and π be a distribution on $\mathcal{X}$. For any f-divergence, we define the (input-restricted[3]) *f-contraction coefficient*

$$\eta_f(\pi, K) := \sup_{\substack{\nu \in \mathcal{P}(\mathcal{X}) \\ 0 < D_f(\nu\|\pi) < \infty}} \frac{D_f(\nu K\|\pi K)}{D_f(\nu\|\pi)}. \tag{1.14}$$

In other words, we have

$$D_f(\nu K\|\pi K) \leq \eta_f(\pi, K) D_f(\nu\|\pi), \tag{1.15}$$

known as the *strong data processing inequality* (SDPI). By the data processing inequality (DPI), we always have $0 \leq \eta_f(\pi, K) \leq 1$, and a smaller value means a stronger contraction ability for f-divergence. The most commonly used f-contraction coefficients include

[3] One could also consider the input-unrestricted contraction coefficient $\eta_f(K) := \sup_{\pi \in \mathcal{P}(\mathcal{X})} \eta_f(\pi, K)$. In this paper we focus on the input-restricted version.

the total variation (TV) contraction coefficient $\eta_{\mathrm{TV}}(\pi, K)$, the Kullback-Leibler (KL) contraction coefficient $\eta_{\mathrm{KL}}(\pi, K)$, and the χ^2-contraction coefficient $\eta_{\chi^2}(\pi, K)$. It is known ([1, 30]) that

$$\eta_{\chi^2}(\pi, K) \le \eta_{\mathrm{KL}}(\pi, K) \le \eta_{\mathrm{TV}}(\pi, K). \tag{1.16}$$

The TV contraction coefficient $\eta_{\mathrm{TV}}(\pi, K)$ is also known as the Dobrushin's coefficient ([19]), and satisfies

$$\eta_{\mathrm{TV}}(\pi, K) = \max_{x,x' \in \mathcal{X}} \mathrm{TV}(K(x,\cdot), K(x',\cdot)). \tag{1.17}$$

Via Eq. (1.16), Eq. (1.17) provides an easy upper bound for $\eta_{\chi^2}(\pi, K)$ and $\eta_{\mathrm{KL}}(\pi, K)$.

We refer the reader to [32, 31] for more discussions on contraction coefficients.

Let $P = KK_\pi^*$. We define the half-step entropy contraction coefficient

$$\alpha(\pi, K) = 1 - \eta_{\mathrm{KL}}(\pi, K) = \inf_{\substack{\nu \in \mathcal{P}(\mathcal{X}) \\ 0 < D(\nu\|\pi) < \infty}} \frac{D(\nu\|\pi) - D(\nu K\|\pi K)}{D(\nu\|\pi)} \tag{1.18}$$

and the full-step entropy contraction coefficient

$$\delta(\pi, P) = 1 - \eta_{\mathrm{KL}}(\pi, P) = \inf_{\substack{\nu \in \mathcal{P}(\mathcal{X}) \\ 0 < D(\nu\|\pi) < \infty}} \frac{D(\nu\|\pi) - D(\nu P\|\pi)}{D(\nu\|\pi)}. \tag{1.19}$$

By rearranging, we have inequalities

$$\mathrm{Ent}_{\pi K}(K_\pi^* f) - \mathrm{Ent}_\pi(f) \le -\alpha \,\mathrm{Ent}_\pi(f), \tag{1.20}$$

$$\mathrm{Ent}_\pi(Pf) - \mathrm{Ent}_\pi(f) \le -\delta \,\mathrm{Ent}_\pi(f), \tag{1.21}$$

for all $f : \mathcal{X} \to \mathbb{R}_{\ge 0}$. Eqs. (1.20) and (1.21) can be seen as definitions of α and δ, and can be compared with Eqs. (1.12) and (1.13). In the next section, we will discuss relationships between the discrete-time contraction notions α, δ and the continuous-time contraction notions ρ, ρ_0, λ.

1.3 Continuous-time versus discrete-time contraction

As we discussed above, the log-Sobolev constant $\rho(\pi, P)$, modified log-Sobolev constant $\rho_0(\pi, P)$, and the Poincaré constant $\lambda(\pi, P)$ represent the contraction ability of the continuous Markov chain, while the contraction coefficients represent the contraction ability of the discrete-time Markov chain. These constants allow one to derive mixing time bounds for associated Markov chains in the continuous-time and discrete-time settings respectively, see e.g. [10] and the references therein for more details. In this paper, we give a full comparison between the discrete-time contraction notions and the continuous-time contraction notions.

The χ^2-contraction coefficient has a close relationship with the Poincaré constant λ. By [32, Theorem 4.3], we have

$$1 - \eta_{\chi^2}(\pi, K) = \lambda(\pi, KK_\pi^*). \tag{1.22}$$

For $P = KK_\pi^*$, we have $1 - \lambda(\pi, P^2) = (1 - \lambda(\pi, P))^2$ (see Lemma 2.6). So

$$1 - \sqrt{\eta_{\chi^2}(\pi, P)} = \lambda(\pi, P). \tag{1.23}$$

In particular,

$$\frac{1}{2}\left(1 - \eta_{\chi^2}(\pi, P)\right) \le \lambda(\pi, P) \le 1 - \eta_{\chi^2}(\pi, P). \tag{1.24}$$

Symbol	Name	Inequality (Definition)
$\rho(\pi, P)$	Log-Sobolev Constant	$\rho \operatorname{Ent}_\pi(f) \le \mathcal{E}_{\pi,P}(\sqrt{f}, \sqrt{f})$
$\alpha(\pi, K)$	Half-Step Entropy Contraction	$\alpha \operatorname{Ent}_\pi(f) \le \operatorname{Ent}_\pi(f) - \operatorname{Ent}_{\pi K}(K_\pi^* f)$
$\delta(\pi, P)$	Full-Step Entropy Contraction	$\delta \operatorname{Ent}_\pi(f) \le \operatorname{Ent}_\pi(f) - \operatorname{Ent}_\pi(Pf)$
$\rho_0(\pi, P)$	Modified Log-Sobolev Constant	$\rho_0 \operatorname{Ent}_\pi(f) \le \mathcal{E}_{\pi,P}(f, \log f)$
$\lambda(\pi, P)$	Poincaré Constant	$\lambda \operatorname{Var}_\pi(f) \le \mathcal{E}_{\pi,P}(f, f)$

Table 1: Contraction notions discussed in this paper. We assume that $P = KK_\pi^*$. See Table 2 for relationships between the constants.

This shows that the rates of χ^2-divergence contraction for continuous-time and discrete-time Markov chains are within a factor of two of each other.

If we consider entropy (KL divergence) rather than χ^2-divergence, then previous works have shown a one-sided inequality. [6] shows that

$$\delta(\pi, P) \le \rho_0(\pi, P). \tag{1.25}$$

[27] proves that

$$\rho(\pi, P) \le \alpha(\pi, K). \tag{1.26}$$

By the data processing inequality, we have

$$\alpha(\pi, K) \le \delta(\pi, P). \tag{1.27}$$

Summarizing the above, we have a chain of inequalities

$$\rho \le \alpha \le \delta \le \rho_0 \le 2\lambda. \tag{1.28}$$

In other words, we have the following implications:

$$\begin{aligned} &\text{Log-Sobolev Inequality} \\ &\quad \Longrightarrow \text{Half-Step Entropy Contraction} \Longrightarrow \text{Full-Step Entropy Contraction} \\ &\qquad \Longrightarrow \text{Modified Log-Sobolev Inequality} \Longrightarrow \text{Poincaré Inequality} \end{aligned} \tag{1.29}$$

In Section 3, we show that the ratio between any two adjacent constants in Eq. (1.28) can be arbitrarily large. In particular, unlike the χ^2-divergence, the discrete-time entropy contraction $\delta(\pi, P)$ is not equivalent to the continuous-time entropy contraction $\rho_0(\pi, P)$.

Table 1 summarizes the main constants discussed in this paper, and Table 2 summarizes relationships between these contraction notions. While these contraction notions are in general non-equivalent (up to a constant factor), we will see in Section 3.5 that under extra conditions, some of them could become equivalent for certain chains. Notably, [34] shows that for a reversible pair (π, P), $\rho_0(\pi, P)$ and $\rho(\pi, P)$ are within a factor of $O(\log(1/p_{\min}))$ of each other, where $p_{\min} = \min_{x,y \in \mathcal{X}: P(x,y)>0} P(x, y)$.

2 Factorizable kernels

Definition 2.1 (Factorizable pairs)**.** *Let $\mathcal{X}$ be a finite set and $P : \mathcal{X} \to \mathcal{X}$ be a Markov kernel with invariant distribution π. We say (π, P) is factorizable if $P = KK_\pi^*$ for some finite Markov kernel $K : \mathcal{X} \to \mathcal{Y}$.*

Lemma 2.2 (Factorizable implies reversible)**.** *A factorizable pair (π, P) is reversible.*

Relationship	Explanation	Reference
$\rho(\pi, P) \le \alpha(\pi, K)$	Log-Sobolev Inequality $\Rightarrow$ Half-Step Entropy Contraction	[27, Prop. 6]
$\rho(\pi, P) \not\gtrsim \alpha(\pi, K)$	Half-Step Entropy Contraction $\not\Rightarrow$ Log-Sobolev Inequality	Section 3.1
$\alpha(\pi, K) \le \delta(\pi, P)$	Half-Step Entropy Contraction $\Rightarrow$ Full-Step Entropy Contraction	Data Processing Inequality
$\alpha(\pi, K) \not\gtrsim \delta(\pi, P)$	Full-Step Entropy Contraction $\not\Rightarrow$ Half-Step Entropy Contraction	Section 3.2
$\delta\left(\pi, P^m\right) \not\gtrsim \delta\left(\pi, P^{m+1}\right)$	$(m+1)$-Step Entropy Contraction $\not\Rightarrow$ m-Step Entropy Contraction	Section 3.2
$\delta(\pi, P) \le \rho_0(\pi, P)$	Full-Step Entropy Contraction $\Rightarrow$ Modified Log-Sobolev Inequality	[6, Lemma 2.7]
$\delta(\pi, P) \not\gtrsim \rho_0(\pi, P)$	Modified Log-Sobolev Inequality $\not\Rightarrow$ Full-Step Entropy Contraction	Section 3.3
$\rho_0(\pi, P) \le 2\lambda(\pi, P)$	Modified Log-Sobolev Inequality $\Rightarrow$ Poincaré Inequality	[8, Prop. 3.6]
$\rho_0(\pi, P) \not\gtrsim \lambda(\pi, P)$	Poincaré Inequality $\not\Rightarrow$ Modified Log-Sobolev Inequality	Section 3.4

Table 2: Relationships between contraction notions. The setting is the same as Table 1. $a \not\gtrsim b$ means $\frac{b}{a}$ can be arbitrarily large. Inequality A $\Rightarrow$ Inequality B means Inequality A with constant a implies Inequality B with constant Ca for some absolute constant $C > 0$.

Proof. Suppose $P = KK_\pi^*$ for a finite Markov kernel $K : \mathcal{X} \to \mathcal{Y}$. For $x, y \in \mathcal{X}$, we have

$$\begin{aligned} \pi(x)P(x,y) &= \pi(x) \sum_{z \in \mathcal{Y}} K(x,z) K_\pi^*(z,y) \\ &= \pi(x)\pi(y) \sum_{\substack{z \in \mathcal{Y} \\ (\pi K)(z) > 0}} \frac{1}{(\pi K)(z)} K(x,z)K(y,z) = \pi(y)P(y,x). \end{aligned} \tag{2.1}$$

So (π, P) is reversible. □

Factorizability is a reasonable assumption for several reasons. Recall that ρ, ρ_0, and λ measure the contraction ability of the Markov generator L, while δ measures the contraction ability of the Markov kernel P. The Markov generator and the Markov kernel are related by $L = P - I$. If we do not make extra assumptions on P, then it could happen that for two Markov kernels P_1 and P_2, P_1 contracts better than P_2, but the corresponding Markov generators L_1 and L_2 satisfy that L_2 contracts better than than L_1. Consider the example where $\mathcal{X} = [2]$, $\pi = \mathrm{Unif}(\mathcal{X})$, $P_c = \begin{pmatrix} 1-c & c \\ c & 1-c \end{pmatrix}$ for $c \in [0,1]$. Then the contraction ability as Markov kernels is increasing for $c \in \left[0, \frac{1}{2}\right]$ and decreasing for $c \in \left[\frac{1}{2}, 1\right]$. However, the contraction ability of the corresponding Markov generators $L_c = P_c - I$ is increasing on the whole interval $[0,1]$. To avoid such undesirable behavior, we need to impose extra assumptions like factorizability. Furthermore, Eqs. (1.26) and (1.27) imply that

$$\rho(\pi, P) \le \delta(\pi, P) \tag{2.2}$$

for factorizable P. While the statement of Eq. (2.2) does not involve factorizability, it does not hold if P is not factorizable. Consider the same example with $c = 1$, that is,

$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Then $\delta(\pi, P) = 0$ but $\rho(\pi, P) = 1$ (e.g., [18, Example 3.1]).

2.1 Important classes of factorizable kernels

Another reason that the factorizability assumption is reasonable is that many natural Markov chains considered in the literature are factorizable. In this section we discuss two important classes of factorizable kernels: lazy kernels and the Glauber dynamics.

Lemma 2.3 (Lazy implies factorizable)**.** *Let (π, P) be a reversible pair. If $P(x,x) \geq \frac{1}{2}$ for all $x \in \mathcal{X}$, then there exists a finite set $\mathcal{Y}$ and a Markov kernel $K : \mathcal{X} \to \mathcal{Y}$ such that $P = KK^*_\pi$.*

Proof. Let $\mathcal{Y} = \binom{\mathcal{X}}{1} \cup \binom{\mathcal{X}}{2}$.[4] Let $K : \mathcal{X} \to \mathcal{Y}$ be the map

$$K(x,e) = \begin{cases} 2P(x,x) - 1, & \text{if } e = \{x\}, \\ 2P(x,y), & \text{if } e = \{x,y\}, \\ 0, & \text{if } x \notin e. \end{cases} \tag{2.3}$$

Then $K^*_\pi(e, \cdot) = \mathrm{Unif}(e)$. For $y \neq x \in \mathcal{X}$, we have

$$(KK^*_\pi)(x,y) = K(x, \{x,y\}) K^*_\pi(\{x,y\}, y) = P(x,y). \tag{2.4}$$

Therefore $P = KK^*_\pi$. □

Lazy chains are quite common. In particular, many of our examples are lazy random walks on regular graphs.

Definition 2.4 (Lazy random walk Markov chain)**.** *Let $G = (V, E)$ be an undirected d-regular simple graph. We associate with it a canonical Markov chain called the (lazy) random walk. Let $\mathcal{X} = V$, $\mathcal{Y} = E$, $\pi = \mathrm{Unif}(\mathcal{X})$, and $K : \mathcal{X} \to \mathcal{Y}$ be $K(x,e) = \frac{1}{d}\mathbb{1}\{x \in e\}$ for $x \in \mathcal{X}$, $e \in \mathcal{Y}$. Then $K^*_\pi(e, \cdot) = \mathrm{Unif}(e)$ and $P = KK^*_\pi$ satisfies*

$$P(x,y) = \begin{cases} \frac{1}{2}, & \text{if } x = y, \\ \frac{1}{2d}, & \text{if } (xy) \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

The Glauber dynamics is an important class of Markov chains that have shown huge practical and theoretical success in sampling spin systems. Let us consider a general α-weighted block dynamics, defined as follows. Let π be a probability measure on $\Omega = [q]^n$ (e.g., the Ising model on a graph with n vertices) with full support. For any $\sigma, \eta \in \Omega$ and $A \subseteq [n]$, consider the conditional probability of η given the configuration σ on A:

$$\pi(\eta | \sigma_A) = \frac{\pi(\eta)\mathbb{1}\{\sigma|_A = \eta|_A\}}{\sum_{\eta' \in \Omega} \pi(\eta')\mathbb{1}\{\sigma|_A = \eta'|_A\}}. \tag{2.6}$$

Let $\mathcal{S} = 2^{[n]}$ be the set of all subsets of $[n]$. For any probability measure $\alpha = (\alpha_A)_{A \in \mathcal{S}}$ on $\mathcal{S}$, the α-weighted block dynamics is the Markov chain on Ω with transition matrix

$$P(\sigma, \sigma') = \sum_{A \in \mathcal{S}} \alpha_A \pi(\sigma' | \sigma_{A^c}) \tag{2.7}$$

where $A^c = [n] \backslash A$. The case $\alpha_A = \frac{1}{n}\mathbb{1}\{|A| = 1\}$ is known as the Glauber dynamics. The pair (π, P) is reversible. A factorization of the α-weighted block dynamics can be defined as follows. Let $\mathcal{X} = \Omega$, $\mathcal{Y} = \mathcal{S} \times \Omega$, and $K : \mathcal{X} \to \mathcal{Y}$ be defined as

$$K(\sigma, (A, \eta)) = \alpha_A \pi(\eta | \sigma_{A^c}). \tag{2.8}$$

[4] For a set $\mathcal{X}$ and an integer $k \geq 0$, $\binom{\mathcal{X}}{k}$ denotes the set of subsets of $\mathcal{X}$ of size k.

One can compute that $(\pi K)(A,\eta) = \alpha_A \pi(\eta)$ and that $K_\pi^*((A,\eta),\sigma) = \pi(\sigma|\eta_{A^c})$ for all $(A,\eta) \in \mathcal{S}\times\Omega$, $\sigma \in \Omega$. Therefore, for all $\sigma,\sigma' \in \Omega$,

$$\sum_{(A,\eta)\in\mathcal{S}\times\Omega} K(\sigma,(A,\eta))K_\pi^*((A,\eta),\sigma') = \sum_{A\in\mathcal{S}} \alpha_A \sum_{\eta\in\Omega} \pi(\eta|\sigma_{A^c})\pi(\sigma'|\eta_{A^c}) = P(\sigma,\sigma'). \tag{2.9}$$

Thus $P = KK_\pi^*$ is a factorization of the α-weighted block dynamics. Half-step entropy contractions for such Markov chains have been recently investigated in the context of spin systems under the name of block factorizations of entropy (e.g., [5]).

2.2 Non-uniqueness of factorization

In general, a factorizable pair (π,P) can be factorized in more than one different ways, and the associated half-step contraction rates may differ considerably. This is illustrated in the following example.

Example 2.5 (Complete graph)**.** Let $n \geq 3$ be an integer. Let $\mathcal{X} = [n]$, $\pi = \mathrm{Unif}(\mathcal{X})$. Let $P : \mathcal{X} \to \mathcal{X}$ be the lazy random walk on the complete graph. That is,

$$P(x,y) = \begin{cases} \frac{1}{2}, & \text{if } x = y, \\ \frac{1}{2(n-1)}, & \text{if } x \neq y. \end{cases} \tag{2.10}$$

Given an integer $2 \leq \ell \leq n$, let $\mathcal{Y}$ be the set of all subsets $A \subseteq [n]$ with either $|A| = 1$ or $|A| = \ell$, and define $K(\ell) : \mathcal{X} \to \mathcal{Y}$ as

$$K(\ell)(x,y) = \begin{cases} \frac{\ell-2}{2(\ell-1)}, & \text{if } y = \{x\}, \\ \frac{\ell}{2(\ell-1)\binom{n-1}{\ell-1}}, & \text{if } |y| = \ell, x \in y, \\ 0, & \text{otherwise.} \end{cases} \tag{2.11}$$

One can check that $K(\ell)_\pi^*(y,\cdot) = \mathrm{Unif}(y)$ and that $K(\ell)K(\ell)_\pi^* = P$. We also note that when $\ell = 2$ this reduces to the construction in the proof of Lemma 2.3.

[9, Theorem 1.1] computes the half-step entropy contraction coefficient $\alpha(\pi,K(\ell))$ for every ℓ, showing that

$$\alpha(\pi,K(\ell)) = \frac{\ell\log\ell}{2(\ell-1)\log n}, \tag{2.12}$$

achieved at and only at point distributions. From Eq. (2.12) we see that $\alpha(\pi,K(\ell))$ increases with ℓ from the minimum value $\alpha(\pi,K(2)) = \frac{\log 2}{\log n}$ to the maximum value $\alpha(\pi,K(n)) = \frac{n}{2(n-1)}$. The former is of the same magnitude of the log-Sobolev constant $\rho(\pi,P) = \frac{n-2}{2(n-1)\log(n-1)}$ ([18, Corollary A.5]) and the latter matches asymptotically with the full-step contraction rate $\delta(\pi,P) = \frac{1}{2} \pm o(1)$ ([22, Eq. (131)]).

2.3 Characterizations of factorizable kernels

An interesting question is to characterize the set of factorizable kernels for a fixed π. Lemma 2.3 provides a sufficient condition and it is certainly not necessary. For example, any pair (π,P) satisfying $P(x,\cdot) = \pi$ for all $x \in \mathcal{X}$ is factorizable (see Example 3.1). On the other hand, a necessary condition for factorizability is positive semidefiniteness.

Lemma 2.6 (Factorizable implies positive semidefinite)**.** *Let (π,P) be a reversible pair. If P is factorizable, then the matrix $\mathrm{Diag}(\pi)P$ is positive semidefinite (PSD), where $\mathrm{Diag}(\pi)$ denotes the $\mathcal{X}\times\mathcal{X}$ diagonal matrix with diagonal π.*

Proof. Let $A = \mathrm{Diag}(\pi)P$. Let $K : \mathcal{X} \to \mathcal{Y}$ be a Markov kernel such that $P = KK_\pi^*$. WLOG assume that πK has full support. Then

$$A_{x,y} = \pi(x)\sum_{z\in\mathcal{Y}} K(x,z)K_\pi^*(z,y) = \pi(x)\pi(y)\sum_{z\in\mathcal{Y}} \frac{1}{(\pi K)(z)}K(x,z)K(y,z). \tag{2.13}$$

So $A = MM^\top$ where $M = \mathrm{Diag}(\pi)K\,\mathrm{Diag}(\pi K)^{-1/2}$. This implies A is PSD. □

In particular, while a factorizable kernel is not necessarily lazy, the proof shows that P must have strictly positive diagonal entries.

For a distribution π on $\mathcal{X}$, let $\mathcal{F}_\pi$ denote the set of Markov kernels $P : \mathcal{X} \to \mathcal{X}$ such that (π, P) is factorizable.

Lemma 2.7 ($\mathcal{F}_\pi$ is convex)**.** *For any distribution π, the set $\mathcal{F}_\pi$ is convex.*

Proof. Let $P_0, P_1 \in \mathcal{F}_\pi$ and $K_0 : \mathcal{X} \to \mathcal{Y}_0$, $K_1 : \mathcal{X} \to \mathcal{Y}_1$ be the corresponding factors. Let $t \in [0,1]$. We prove that $P_t := (1-t)P_0 + tP_1$ is in $\mathcal{F}_\pi$. Let $K_t : \mathcal{X} \to \mathcal{Y}_0 \sqcup \mathcal{Y}_1$ be defined as

$$K_t(x,y) = \begin{cases} (1-t)K_0(x,y), & \text{if } y \in \mathcal{Y}_0, \\ tK_1(x,y), & \text{if } y \in \mathcal{Y}_1. \end{cases} \tag{2.14}$$

Then we can verify that K_t is a Markov kernel, and

$$K_t(K_t)^*_\pi = (1-t)K_0(K_0)^*_\pi + tK_1(K_1)^*_\pi = P_t. \tag{2.15}$$

□

For a distribution π on $\mathcal{X}$, let $\mathcal{P}_\pi$ denote the set of Markov kernels $P : \mathcal{X} \to \mathcal{X}$ such that $\mathrm{Diag}(\pi)P$ is PSD. By Lemmas 2.6 and 2.7, $\mathcal{F}_\pi$ is a convex subset of $\mathcal{P}_\pi$. When the state space is binary, the two sets are equal, but this is not true in general.

Lemma 2.8. *If $|\mathcal{X}| = 2$, then for any distribution π on $\mathcal{X}$, we have $\mathcal{F}_\pi = \mathcal{P}_\pi$.*

Proof. Direct calculation shows that $\mathcal{P}_\pi$ is the convex hull of $\{J\,\mathrm{Diag}(\pi), I\}$, where J is the all-ones matrix. Both extreme points are in $\mathcal{F}_\pi$. □

Lemma 2.9. *For $n \geq 5$, $\mathcal{X} = [n]$, and $\pi = \mathrm{Unif}(\mathcal{X})$, the set $\mathcal{F}_\pi$ is strictly smaller than $\mathcal{P}_\pi$.*

Proof. An $n \times n$ matrix A is called completely positive if it can be written as $A = MM^\top$ for some (not necessarily square) matrix M with non-negative entries. Clearly, all completely positive matrices are PSD. It is known ([26, 4]) that for $n \leq 4$, a PSD matrix is completely positive if and only if all its entries are non-negative, while for $n \geq 5$, there exist PSD matrices with strictly positive entries that are not completely positive.

Fix $n \geq 5$, $\mathcal{X} = [n]$, and $\pi = \mathrm{Unif}(\mathcal{X})$. Then $\mathcal{P}_\pi$ is exactly the set of doubly stochastic PSD matrices. Let A be an $n \times n$ PSD matrix with strictly positive entries that is not completely positive. By Sinkhorn's theorem ([25, 36]), there is a diagonal matrix D with strictly positive entries such that DAD is doubly stochastic. Let $P = DAD$. Clearly P is in $\mathcal{P}_\pi$. We claim that P is not in $\mathcal{F}_\pi$. Suppose for the sake of contradiction that $P = KK^*_\pi$ for some K. Then

$$A = D^{-1}PD^{-1} = \frac{1}{n}D^{-1}K\,\mathrm{Diag}(\pi K)^{-1}K^\top D^{-1} = MM^\top \tag{2.16}$$

where $M = \frac{1}{\sqrt{n}}D^{-1}K\,\mathrm{Diag}(\pi K)^{-1/2}$ is non-negative. This contradicts with the assumption that A is not completely positive. □

It remains an interesting open problem to characterize $\mathcal{F}_\pi$, even for uniform π.

3 Comparison between constants

In this section we compare constants in Table 1, showing that there is a separation between any two of them. Table 3 summarizes examples in this section.

Example	Description	Separation
Example 3.1	One-step chain	Log-Sobolev Constant $\rho(\pi, P)$ vs Half-Step Entropy Contraction $\alpha(\pi, K)$
Example 3.2	1-to-k chain	Half-Step Entropy Contraction $\alpha(\pi, K)$ vs Full-Step Entropy Contraction $\delta(\pi, P)$
Example 3.4	Bernoulli-Laplace model	Half-Step Entropy Contraction $\alpha(\pi, K)$ vs Full-Step Entropy Contraction $\delta(\pi, P)$
Example 3.7	Three-state chain	One-Step Entropy Contraction $\delta(\pi, P)$ vs Two-Step Entropy Contraction $\delta\left(\pi, P^2\right)$
Example 3.9	Birth-death chain	m-Step Entropy Contraction $\delta\left(\pi, P^m\right)$ vs $(m+1)$-Step Entropy Contraction $\delta\left(\pi, P^{m+1}\right)$
Example 3.11	Three-state chain	Full-Step Entropy Contraction $\delta(\pi, P)$ vs Modified Log-Sobolev Constant $\rho_0(\pi, P)$
Example 3.14	Expander graph	Modified Log-Sobolev Constant $\rho_0(\pi, P)$ vs Poincaré Constant $\lambda(\pi, P)$
Example 3.15	Random transposition model	Half-Step Entropy Contraction $\alpha(\pi, K)$ vs Modified Log-Sobolev Constant $\rho_0(\pi, P)$

Table 3: Examples in Section 3 and the separations they witness.

3.1 Log-Sobolev constant ρ vs half-step entropy contraction α

By [27], we always have $\rho(\pi, P) \le \alpha(\pi, K)$ for $P = KK_\pi^*$. The following example shows that their ratio can be arbitrarily large.

Example 3.1 (A one-step Markov chain)**.** Let $\mathcal{X}$ be a finite set, π be a distribution on $\mathcal{X}$ with full support. Let $\mathcal{Y} = \{*\}$ and $K : \mathcal{X} \to \mathcal{Y}$ be the unique Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$. Then $P = KK_\pi^*$ satisfies $P(x, y) = \pi(y)$ for all $x, y \in \mathcal{X}$. By [18, Theorem A.1],

$$\rho(\pi, P) = \frac{1 - 2\pi_*}{\log(1/\pi_* - 1)} \tag{3.1}$$

where $\pi_* = \min_{x \in \mathcal{X}} \pi(x)$. On the other hand, $\alpha(\pi, K) = 1 - \eta_{\mathrm{KL}}(\pi, K) = 1$. As $\pi_* \to 0$, we have $\frac{\alpha(\pi,K)}{\rho(\pi,P)} \to \infty$.

3.2 Half-step entropy contraction α vs full-step entropy contraction δ

By the data processing inequality, we always have $\alpha(\pi, K) \le \delta(\pi, P)$ for $P = KK_\pi^*$. Example 2.5 with $\ell = 2$ already shows that the ratio can be arbitrarily large. Below we present a few different examples.

Example 3.2 (A 1-to-k Markov chain)**.** Let $\mathcal{X} = [n]$, $\mathcal{Y} = [n]^k$, $\pi = \mathrm{Unif}(\mathcal{X})$, and $K(x, y) = \frac{1}{kn^{k-1}} \sum_{j \in [k]} \mathbb{1}\{y_j = x\}$. In other words, on input $x \in \mathcal{X}$, this chain generates a uniform length-k output string and then randomly overwrite one of the k positions with x. Then $K_\pi^* : \mathcal{Y} \to \mathcal{X}$ satisfies $K_\pi^*(y, x) = \frac{1}{k} \sum_{j \in [k]} \mathbb{1}\{y_j = x\}$. Let $P = KK_\pi^*$. The motivation for this chain comes from analysis of the Glauber dynamics on $\mathrm{Unif}\left(\mathcal{X}^k\right)$. The Markov kernel K_π^* is the k-to-1 walk, and its entropy contraction is called entropic independence in [2]. Prop. 3.3 shows that for constant $k \ge 2$, as $n \to \infty$, we have $\frac{\delta(\pi,P)}{\alpha(\pi,K)} \to \infty$.

Proposition 3.3 (A 1-to-k Markov chain)**.** *Let π, K, P be as in Example 3.2. Then we have $\alpha(\pi, K) = O\left(\frac{1}{\log n}\right)$ and $\delta(\pi, P) \ge 1 - \frac{1}{k}$. In particular, for fixed $k \ge 2$, $\frac{\delta(\pi,P)}{\alpha(\pi,K)} = \Omega(\log n)$.*

Proof. **Upper bound on $\alpha(\pi, K)$.** Let ν be the point distribution at $1 \in \mathcal{X}$. Then $D(\nu \| \pi) = \log n$. Consider the distribution νK. For $y \in \mathcal{Y}$, if y contains i copies of 1, then

$(\nu K)(y) = \frac{i}{kn^{k-1}}$. So

$$D(\nu K\|\pi K) = \sum_{1\le i\le k}\binom{k}{i}(n-1)^{k-i}\cdot\frac{i}{kn^{k-1}}\log\frac{ni}{k} \tag{3.2}$$
$$= \log n - \sum_{1\le i\le k}\binom{k-1}{i-1}\frac{(n-1)^{k-i}}{n^{k-1}}\log\frac{k}{i}.$$

For constant $k\ge 2$, as $n\to\infty$, we have $D(\nu K\|\pi K) = \log n - \Theta(1)$. Therefore

$$\alpha(\pi, K) \le 1 - \frac{D(\nu K\|\pi K)}{D(\nu\|\pi)} = \Theta\left(\frac{1}{\log n}\right). \tag{3.3}$$

Lower bound on $\delta(\pi, P)$. Let $M : \mathcal{Y}\to[k]\times[n]$ be the Markov kernel defined as $M(y,\cdot) = \mathrm{Unif}(\{(i, y_i) : i\in[k]\})$. By the data processing inequality,

$$1-\delta(\pi,P) = \eta_{\mathrm{KL}}(\pi,P) \le \eta_{\mathrm{KL}}(\pi K, K^*_\pi) \le \eta_{\mathrm{KL}}(\pi K, M). \tag{3.4}$$

Let ν be any distribution on $\mathcal{Y}$, and ν_i ($i\in[k]$) be the i-th marginal of ν. Then

$$D(\nu\|\pi K) = D\left(\nu\|\pi^{\otimes k}\right) = D(\nu\|\nu_1\otimes\cdots\otimes\nu_k) + \sum_{i\in[k]} D(\nu_i\|\pi) \ge \sum_{i\in[k]} D(\nu_i\|\pi). \tag{3.5}$$

On the other hand,

$$D(\nu M\|\pi K M) = D(\nu M\|\,\mathrm{Unif}([k])\otimes\pi) = \frac{1}{k}\sum_{i\in[k]} D(\nu_i\|\pi). \tag{3.6}$$

Therefore

$$\eta_{\mathrm{KL}}(\pi K, M) \le \frac{1}{k}. \tag{3.7}$$

So

$$\delta(\pi,P) \ge 1 - \frac{1}{k}. \tag{3.8}$$

□

Example 3.4 (Bernoulli-Laplace model)**.** Let n be a positive integer and $1\le k\le n-1$. We define a graph $G=(V,E)$. Let $V = \binom{[n]}{k}$ (i.e., size-k subsets of of [n]). Equivalently, V is the set of length-n bit strings with Hamming weight k. There is an edge (x,y) for $x,y\in V$ if and only if $\|x-y\|_1 = 2$ (considered as elements in $\{0,1\}^n$). The Bernoulli-Laplace model is the lazy random walk on G (Definition 2.4). That is, $\mathcal{X}=V$, $\pi=\mathrm{Unif}(\mathcal{X})$, $\mathcal{Y}=E$, $K:\mathcal{X}\to\mathcal{Y}$ is defined as $K(x,e) = \frac{1}{k(n-k)}\mathbb{1}\{x\in e\}$. For $(ij)\in\binom{[n]}{2}$, define map $\sigma_{ij}:\mathcal{X}\to\mathcal{X}$ by swapping the i-th coordinate and the j-th coordinate (under the bit string interpretation). Then

$$P(x,\cdot) = \frac{1}{2}\mathbb{1}_x + \frac{1}{2k(n-k)}\sum_{\substack{i\in x\\ j\in[n]\setminus x}}\mathbb{1}_{\sigma_{ij}(x)}. \tag{3.9}$$

Prop. 3.5 shows that for constant $k\ge 1$, as $n\to\infty$, we have $\frac{\delta(\pi,P)}{\alpha(\pi,K)}\to\infty$.

Proposition 3.5 (Bernoulli-Laplace model)**.** *Let π, K, P be as in Example 3.4 with $1\le k\le n-1$. Then $\alpha(\pi,K) = O\left(\frac{1}{\log\binom{n}{k}}\right)$ and $\delta(\pi,P)\ge\frac{n}{2k(n-k)}$. In particular, for constant $k\ge 1$, we have $\frac{\delta(\pi,P)}{\alpha(\pi,K)} = \Omega(\log n)$.*

Proof. **Upper bound on $\alpha(\pi, K)$.** Let ν be the point distribution on any $x \in \mathcal{X}$. Then $D(\nu\|\pi) = \log|\mathcal{X}| = \log\binom{n}{k}$,

$$\nu K = \frac{1}{k(n-k)} \sum_{\substack{i\in x \\ j\in[n]\setminus x}} \mathbb{1}_{\{x,\sigma_{ij}(x)\}}. \tag{3.10}$$

Because $\pi K = \mathrm{Unif}(\mathcal{Y})$,

$$D(\nu K\|\pi K) = \log|\mathcal{Y}| - \log(k(n-k)) = \log\binom{n}{k} - \log 2. \tag{3.11}$$

Therefore

$$\alpha(\pi, K) \le 1 - \frac{D(\nu K\|\pi K)}{D(\nu\|\pi)} = \frac{\log 2}{\log\binom{n}{k}}. \tag{3.12}$$

Lower bound on $\delta(\pi, P)$. We make use of the following useful result from [12].

Lemma 3.6 ([12, Theorem 1])**.** *Let d be a metric on $\mathcal{X}$. Let W_p denote the Wasserstein p-distance on $\mathcal{P}(\mathcal{X})$ If*

$$W_\infty(P(x,\cdot), P(y,\cdot)) \le d(x,y), \qquad \forall x, y \in \mathcal{X}, \tag{3.13}$$
$$W_1(P(x,\cdot), P(y,\cdot)) \le (1-\kappa)d(x,y), \qquad \forall x, y \in \mathcal{X}, \tag{3.14}$$

then

$$\delta(\pi, P) \ge \kappa. \tag{3.15}$$

For the Bernoulli-Laplace model, we let d be the graph distance on $V = \binom{[n]}{k}$. To prove Eqs. (3.13) and (3.14), it suffices to prove the result for adjacent x and y. By symmetry, WLOG assume that $x = \{1, 3, 4, \ldots, k+1\}$, $y = \{2, 3, \ldots, k+1\}$. We define a coupling between $P(x,\cdot)$ and $P(y,\cdot)$ such that Eqs. (3.13) and (3.14) are both satisfied.

(1) For $3 \le i \le k+1$, $k+2 \le j \le n$, couple $\sigma_{ij}(x)$ with $\sigma_{ij}(y)$. This happens with probability $\frac{(k-1)(n-k-1)}{2k(n-k)}$ and incurs distance 1.

(2) For $k+2 \le j \le n$, couple $\sigma_{1j}(x)$ with $\sigma_{2j}(y)$. This happens with probability $\frac{n-k-1}{2k(n-k)}$ and incurs distance 0.

(3) For $3 \le i \le k+1$, couple $\sigma_{2i}(x)$ with $\sigma_{1i}(y)$. This happens with probabilities $\frac{k-1}{2k(n-k)}$ and incurs distance 0.

(4) Couple $\sigma_{12}(x)$ with y, and x with $\sigma_{12}(y)$ each with weight $\frac{1}{2k(n-k)}$. This happens with probability $\frac{1}{k(n-k)}$ and incurs distance 0.

(5) At this point, all remaining mass in $P(x,\cdot)$ (resp. $P(y,\cdot)$) is at x (resp. y). Couple them directly. This happens with probability $\frac{1}{2} - \frac{1}{2k(n-k)}$ and incurs distance 1.

To summarize, the coupling has distance at most one and expected distance $1 - \frac{n}{2k(n-k)}$. By Lemma 3.6, we have

$$\delta(\pi, P) \ge \frac{n}{2k(n-k)}. \tag{3.16}$$

□

We remark that for the Bernoulli-Laplace model, the exact value of $\alpha(\pi, K)$ has been determined in [9, Theorem 1.12], where it is shown that Eq. (3.12) is tight. For $\delta(\pi, P)$, by considering a point distribution at any $x \in \mathcal{X}$, we have

$$\delta(\pi, P) \le \frac{\log(2n(n-k))}{2\log\binom{n}{k}}. \tag{3.17}$$

Therefore, for n, k satisfying $\log k = (1 - \Omega(1)) \log n$, we have $\delta(\pi, P) = \Theta\left(\frac{1}{k}\right)$.

Examples 3.2 and 3.4 show that there is a separation between $\alpha(\pi, K)$ and $\delta(\pi, P = KK_\pi^*)$. In these examples, K_π^* can be quite different from K. One natural question is whether there is a separation between $\delta(\pi, P)$ and $\delta\left(\pi, P^2\right)$. That is, can running the same Markov kernel twice result in much better contraction than running only once? The following example shows that indeed such a separation exists. This example is adapted from [29]'s counterexample to the Peres-Tetali conjecture. We note, however, that the properties of this chain we use here and in Example 3.11 are different from those used op. cit.

Example 3.7 (A three-state Markov chain)**.** Let M be a positive real number. Let $\mathcal{X} = [3]$, $\pi = \left(\frac{M}{M+2}, \frac{1}{M+2}, \frac{1}{M+2}\right)$. Let $P : \mathcal{X} \to \mathcal{X}$ be defined as

$$P = \begin{pmatrix} 1 - \frac{1}{4M} & \frac{1}{4M} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}. \tag{3.18}$$

It is easy to see that π is the invariant distribution of P and (π, P) is reversible. Note that by Lemma 2.3, P is factorizable. By Prop. 3.8, $\frac{\delta\left(\pi, P^2\right)}{\delta(\pi, P)} \to \infty$ as $M \to \infty$.

Proposition 3.8 (A three-state Markov chain)**.** *Let π, P be as in Example 3.7. Then $\delta(\pi, P) = O\left(\frac{1}{\log M}\right)$ and $\delta\left(\pi, P^2\right) = \Omega(1)$. In particular, $\frac{\delta\left(\pi, P^2\right)}{\delta(\pi, P)} = \Omega(\log M)$.*

Proof. **Upper bound on $\delta(\pi, P)$.** Let ν be the point distribution at $3 \in \mathcal{X}$. Then $D(\nu\|\pi) = \log(M+2)$,

$$D(\nu P\|\pi) = \frac{1}{4}\log\frac{M+2}{4} + \frac{3}{4}\log\frac{3(M+2)}{4} = \log(M+2) - h(1/4) \tag{3.19}$$

where $h(x) : [0,1] \to [0, \log 2]$ is the binary entropy function

$$h(x) = -x\log x - (1-x)\log(1-x). \tag{3.20}$$

Therefore

$$\delta(\pi, P) \le 1 - \frac{D(\nu P\|\pi)}{D(\nu\|\pi)} = O\left(\frac{1}{\log M}\right). \tag{3.21}$$

Lower bound on $\delta\left(\pi, P^2\right)$. We prove that for large M, we have $P^2(x, 1) = \Omega(1)$ for all $x \in \mathcal{X}$. In fact, $P^2(1,1) \ge P(1,1)^2 = \Omega(1)$, $P^2(2,1) \ge P(2,1)P(1,1) = \Omega(1)$, and $P^2(3,1) \ge P(3,2)P(2,1) = \Omega(1)$. So $\mathrm{TV}(P^2(x,\cdot), P^2(x',\cdot)) = 1 - \Omega(1)$ for all $x, x' \in \mathcal{X}$. By Eq. (1.17), the Dobrushin's coefficient satisfies $\eta_{\mathrm{TV}}(\pi, P^2) = 1 - \Omega(1)$. By Eq. (1.16), we have

$$\delta\left(\pi, P^2\right) \ge 1 - \eta_{\mathrm{TV}}(\pi, P^2) = \Omega(1). \tag{3.22}$$

□

We generalize Example 3.7 as follows, showing that for any positive integer m, there exists a Markov kernel such that running $(m+1)$ steps results in entropy contraction much better than running m steps. We note that there is a relatively simple characterization of the LSI for birth-death chains ([28, 13, 14]), but for MLSI or SDPI no such characterizations are known, except for partial progress in [33, 11].

Example 3.9 (A birth-death Markov chain)**.** We fix a positive integer m and let M be a large positive real number. Let $\mathcal{X} = [m+2]$, $\pi(x) = \frac{1}{M+m+1} + \mathbb{1}\{x = 1\}\frac{M-1}{M+m+1}$. Let $P : \mathcal{X} \to \mathcal{X}$ be a birth-death Markov chain, where $P(x,y) = 0$ for $|x-y| \geq 2$, $P(x, x-1) = \frac{1}{4}$ for $2 \leq x \leq m+2$, $P(x, x+1) = \frac{1}{4}$ for $2 \leq x \leq m+1$, $P(1,2) = \frac{1}{4M}$, and $P(x,x) = 1 - P(x, x-1) - P(x, x+1)$. It is easy to verify that (π, P) is a reversible pair and $P(x,x) \geq \frac{1}{2}$ for all $x \in [m+2]$. By Lemma 2.3, (π, P) is factorizable. Prop. 3.10 shows that as $M \to \infty$, we have $\frac{\delta\left(\pi, P^{m+1}\right)}{\delta(\pi, P^m)} \to \infty$.

Proposition 3.10 (A birth-death Markov chain)**.** *Let π, P be as in Example 3.9. Then $\delta(\pi, P^m) = O\left(\frac{1}{\log M}\right)$ and $\delta\left(\pi, P^{m+1}\right) = \Omega(1)$. In particular, $\frac{\delta\left(\pi, P^{m+1}\right)}{\delta(\pi, P^m)} = \Omega(\log M)$.*

Proof. **Upper bound on $\delta(\pi, P^m)$.** Let ν be the point distribution at $m+2 \in \mathcal{X}$. Then $D(\nu\|\pi) = \log(M+m+1)$. Note that $(\nu P^m)(1) = 0$, $(\nu P^m)(x) = c_x$ for some $c_x = \Theta_m(1)$ for $2 \leq x \leq m+2$, where Θ_m hides a constant factor depending on m. Furthermore, $(c_2, \ldots, c_{m+2})$ is a distribution on $\{2, \ldots, m+2\}$. Then

$$D(\nu P\|\pi) = \log(M+m+1) - H(c_2, \ldots, c_{m+2}) \tag{3.23}$$

where H is the entropy function

$$H(c_2, \ldots, c_{m+2}) = -\sum_{2 \leq i \leq m+2} c_i \log c_i. \tag{3.24}$$

Because $c_i = \Theta_m(1)$ for all $2 \leq i \leq m+2$, we have $H(c_2, \ldots, c_{m+2}) = \Theta_m(1)$. Therefore

$$\delta(\pi, P^m) \leq 1 - \frac{D(\nu P\|\pi)}{D(\nu\|\pi)} = \Theta_m\left(\frac{1}{\log M}\right). \tag{3.25}$$

Lower bound on $\delta\left(\pi, P^{m+1}\right)$. Note that for any $x \in [m+2]$, we have

$$P^{m+1}(x,1) \geq P(x, x-1) \cdots P(2,1) \cdot P(1,1)^{m+1-x} = \Omega_m(1) \tag{3.26}$$

where Ω_m hides a constant factor depending on m. By Eqs. (1.17) and (1.16),

$$\delta\left(\pi, P^{m+1}\right) \geq 1 - \eta_{\mathrm{TV}}(\pi, P^{m+1}) = \Omega_m(1). \tag{3.27}$$

□

If a Markov kernel (π, P) separates $\delta(\pi, P^m)$ and $\delta\left(\pi, P^{m+1}\right)$, then it also separates $\delta(\pi, T_t)$ (where $T_t = e^{t(P-\mathrm{Id})}$) and $\delta(\pi, P^m)$ for finite t. In other words, the continuous-time chain contracts entropy at finite time better than the discrete-time counterpart. To see this, note that for any function f on $\mathcal{X}$, we have

$$\begin{aligned}
\mathrm{Ent}_\pi(T_t f) &\leq \mathbb{E}_{n \sim \mathrm{Pois}(t)} \mathrm{Ent}_\pi\left(P^n f\right) \\
&\leq \mathbb{E}_{n \sim \mathrm{Pois}(t)} \left[\mathbb{1}\{n \leq m\} \mathrm{Ent}_\pi(f) + \mathbb{1}\{n \geq m+1\} \mathrm{Ent}_\pi\left(P^{m+1} f\right)\right] \\
&\leq \mathbb{P}[\mathrm{Pois}(t) \leq m] \mathrm{Ent}_\pi(f) + \mathbb{P}[\mathrm{Pois}(t) \geq m+1]\left(1 - \delta\left(\pi, P^{m+1}\right)\right) \mathrm{Ent}_\pi(f) \\
&\leq \left(1 - \mathbb{P}[\mathrm{Pois}(t) \geq m+1]\delta\left(\pi, P^{m+1}\right)\right) \mathrm{Ent}_\pi(f),
\end{aligned} \tag{3.28}$$

where the first step is by convexity of Ent_π, and the second step is by the data processing inequality. Therefore

$$\delta(\pi, T_t) \geq \mathbb{P}[\mathrm{Pois}(t) \geq m+1]\delta\left(\pi, P^{m+1}\right), \tag{3.29}$$

which is separated from $\delta(\pi, P^m)$ for finite t, assuming that $\delta\left(\pi, P^{m+1}\right)$ is separated from $\delta(\pi, P^m)$.

3.3 Full-step entropy contraction δ vs modified log-Sobolev constant ρ_0

By [6, Lemma 2.7], we always have $\delta(\pi, P) \le \rho_0(\pi, P)$ for any reversible (π, P).[5] If we allow non-factorizable (π, P), then it is easy to give an example where their ratio is infinite: take $\mathcal{X} = [2]$, $\pi = \mathrm{Unif}(\mathcal{X})$, and $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. The following example shows that the ratio can be arbitrarily large even if we restrict to factorizable (π, P).

Example 3.11 (A three-state Markov chain)**.** This example is the same as Example 3.7. Prop. 3.12 shows that we have $\frac{\rho_0(\pi,P)}{\delta(\pi,P)} \to \infty$ as $M \to \infty$.

Proposition 3.12 (A three-state Markov chain)**.** *Let (π, P) be as in Example 3.7. Then $\delta(\pi, P) = O\left(\frac{1}{\log M}\right)$ and $\rho_0(\pi, P) = \Theta\left(\frac{\log\log M}{\log M}\right)$. In particular, $\frac{\rho_0(\pi,P)}{\delta(\pi,P)} = \Omega(\log\log M)$.*

Proof. **Upper bound on $\delta(\pi, P)$.** We have established in Prop. 3.8 that $\delta(\pi, P) = O\left(\frac{1}{\log M}\right)$.

Lower bound on $\rho_0(\pi, P)$. We will show that there exists a sufficiently large universal constant $C > 0$ such that, for any positive function $f : [3] \to \mathbb{R}^+$, it holds

$$\mathrm{Ent}_\pi(f) \le \frac{C \log M}{\log\log M} \mathcal{E}_{\pi,P}(f, \log f). \tag{3.30}$$

This implies $\rho_0(\pi, P) = \Omega\left(\frac{\log\log M}{\log M}\right)$. By applying suitable scaling, we assume that $f(1) = x$, $f(2) = 1$, and $f(3) = y$ for some $x, y > 0$. Furthermore, we may assume WLOG that $x \ge \frac{\beta}{M}$ where $\beta > 0$ is some tiny absolute constant, due to [37, Lemma 2.1]. More specifically, [37, Lemma 2.1] shows that to establish a modified log-Sobolev inequality Eq. (3.30), it suffices to consider a restricted class of functions such that, among other restrictions, $f(1) \ge \beta' \mathbb{E}_\pi f$ for an absolute constant $\beta' > 0$; such a restriction immediately implies

$$x = f(1) \ge \beta' \left(\frac{Mx}{M+2} + \frac{1}{M+2} + \frac{y}{M+2} \right) \ge \frac{\beta'}{M+2} \ge \frac{\beta'}{3M}. \tag{3.31}$$

Hence, we can safely assume $x \ge \frac{\beta}{M}$ where $\beta = \beta'/3$.

A direct calculation yields

$$(M+2) \cdot \mathrm{Ent}_\pi(f) = Mx \log x + y \log y - (Mx + y + 1) \log\left(\frac{Mx + y + 1}{M+2} \right), \tag{3.32}$$

and

$$4(M+2) \cdot \mathcal{E}_{\pi,P}(f, \log f) = (x-1)\log x + (y-1)\log y. \tag{3.33}$$

The following claim is helpful.

Claim 3.13. *We have*

$$\frac{\mathrm{Ent}_\pi(f)}{4\,\mathcal{E}_{\pi,P}(f, \log f)} \le \frac{(2x - y - 1) + y\log y + (y+1)\log\left(\frac{1}{x}\right)}{(x-1)\log x + (y-1)\log y}. \tag{3.34}$$

Proof. We rewrite the entropy as

$$\begin{aligned} (M+2) \cdot \mathrm{Ent}_\pi(f) &= Mx \log x + y \log y - (Mx + y + 1) \log\left(\frac{Mx + y + 1}{M+2} \right) \\ &= (Mx + y + 1) \log\left(\frac{(M+2)x}{Mx + y + 1} \right) + y \log y + (y+1) \log\left(\frac{1}{x} \right). \end{aligned} \tag{3.35}$$

[5] They in fact prove a more general statement that works for non-reversible (π, P).

Notice that the first term above can be controlled by

$$(Mx+y+1)\log\left(\frac{(M+2)x}{Mx+y+1}\right) = (Mx+y+1)\log\left(1+\frac{2x-y-1}{Mx+y+1}\right) \leq (Mx+y+1)\cdot\frac{2x-y-1}{Mx+y+1} = 2x-y-1. \tag{3.36}$$

The claim then follows. □

We consider two separate cases of (x,y) to establish Eq. (3.30).
Case 1: $(x,y)\notin(\frac{1}{2},\frac{3}{2})\times(\frac{1}{2},\frac{3}{2})$**.** In this case, we have

$$(x-1)\log x+(y-1)\log y \geq \frac{1}{10}. \tag{3.37}$$

Since we have

$$2x-y-1\leq 2x-1\leq 2(x-1)\log x+3\leq 32\left((x-1)\log x+(y-1)\log y\right) \tag{3.38}$$

and also

$$y\log y\leq 2(y-1)\log y+1\leq 12\left((x-1)\log x+(y-1)\log y\right), \tag{3.39}$$

we deduce from Claim 3.13 that

$$\frac{\mathrm{Ent}_\pi(f)}{4\,\mathcal{E}_{\pi,P}(f,\log f)}\leq 44+\frac{(y+1)\log\left(\frac{1}{x}\right)}{(x-1)\log x+(y-1)\log y}. \tag{3.40}$$

Consider three subcases.

(i) If $x\geq\frac{3}{2}$, then $\log(1/x)<0$ and hence

$$\frac{\mathrm{Ent}_\pi(f)}{4\,\mathcal{E}_{\pi,P}(f,\log f)}\leq 44. \tag{3.41}$$

(ii) If $x\leq\frac{1}{2}$, then consider how large y is. If $y\leq\frac{\log M}{\log\log M}$, then by $(x-1)\log x\geq\frac{1}{2}\log(1/x)$ and $(y-1)\log y\geq 0$ we deduce that

$$\frac{(y+1)\log\left(\frac{1}{x}\right)}{(x-1)\log x+(y-1)\log y}\leq 2(y+1)\leq\frac{4\log M}{\log\log M}. \tag{3.42}$$

If $y>\frac{\log M}{\log\log M}$, then by $(x-1)\log x\geq 0$, $\frac{y+1}{y-1}\leq 3$, and $x\geq\frac{\beta}{M}$ we deduce that

$$\frac{(y+1)\log\left(\frac{1}{x}\right)}{(x-1)\log x+(y-1)\log y}\leq\frac{(y+1)\log\left(\frac{1}{x}\right)}{(y-1)\log y}\leq\frac{3\log(\frac{M}{\beta})}{\log(\frac{\log M}{\log\log M})}\leq\frac{C_0\log M}{\log\log M}, \tag{3.43}$$

for some $C_0=C_0(\beta)>0$ when M is sufficiently large.

(iii) If $x\in(\frac{1}{2},\frac{3}{2})$, then by our assumption it must hold $y\notin(\frac{1}{2},\frac{3}{2})$. Since $\log(1/x)\leq 1$ when $x>\frac{1}{2}$, we have

$$\frac{(y+1)\log\left(\frac{1}{x}\right)}{(x-1)\log x+(y-1)\log y}\leq\frac{y+1}{(y-1)\log y}\leq 10, \tag{3.44}$$

when $y\notin(\frac{1}{2},\frac{3}{2})$.

Therefore, in all three subcases we have

$$\frac{\mathrm{Ent}_\pi(f)}{4\,\mathcal{E}_{\pi,P}(f,\log f)} \le \frac{C\log M}{\log\log M} \tag{3.45}$$

where $C = C(\beta) > 0$ is constant, whenever M is sufficiently large.

Case 2: $(x,y) \in (\frac{1}{2},\frac{3}{2}) \times (\frac{1}{2},\frac{3}{2})$**.** In this case, we have

$$(x-1)\log x + (y-1)\log y \ge \frac{1}{2}(x-1)^2 + \frac{1}{2}(y-1)^2, \tag{3.46}$$

and also

$$\begin{aligned}
&(2x-y-1) + y\log y + (y+1)\log\left(\frac{1}{x}\right) \qquad (3.47)\\
&\le 2(x-1) - (y-1) + y(y-1) + (y+1)\left(\frac{1}{x}-1\right)\\
&= \frac{1}{x}(x-1)\,(2x-y-1) + (y-1)^2\\
&= \frac{2}{x}(x-1)^2 - \frac{1}{x}(x-1)(y-1) + (y-1)^2\\
&\le 4(x-1)^2 + 2|(x-1)(y-1)| + (y-1)^2\\
&\le 5(x-1)^2 + 5(y-1)^2.
\end{aligned}$$

By Claim 3.13,

$$\frac{\mathrm{Ent}_\pi(f)}{4\,\mathcal{E}_{\pi,P}(f,\log f)} \le 10. \tag{3.48}$$

Combining the two cases, we conclude that

$$\rho_0(\pi,P) = \Omega\left(\frac{\log\log M}{\log M}\right). \tag{3.49}$$

In fact, this lower bound on $\rho_0(\pi,P)$ is asymptotically tight and can be achieved by, for example, $f(1) = 1/M$, $f(2) = 1$, and $f(3) = \log M$ as given in [29]. Therefore,

$$\rho_0(\pi,P) = \Theta\left(\frac{\log\log M}{\log M}\right) \tag{3.50}$$

as $M \to \infty$.

□

3.4 Modified log-Sobolev constant ρ_0 vs Poincaré constant λ

[8] shows that $\rho_0(\pi,P) \le 2\lambda(\pi,P)$, and provides an example showing that the ratio can be arbitrarily large. We record their example here.

Example 3.14 (Expander graphs)**.** Let $G = (V,E)$ be a expander graph with bounded degree. Consider the lazy random walk on G (Definition 2.4). [8] shows that $\rho_0(\pi,P) = \Theta\left(\frac{1}{\log|V|}\right) = \Theta\left(\frac{\lambda_1(\pi,P)}{\log|V|}\right)$. Therefore as $|V| \to \infty$, we have $\frac{\lambda_1(\pi,P)}{\rho_0(\pi,P)} \to \infty$.

3.5 Other comparisons

[18] shows that the spectral gap λ and the log-Sobolev constant ρ differ by at most a factor of $O(\log(1/\pi_{\min}))$ where

$$\pi_{\min} = \min_{x\in\mathcal{X}:\,\pi(x)>0} \pi(x). \tag{3.51}$$

More precisely, [18, Corollay A.4] shows that, assuming $\pi_{\min} \le 1/2$, it holds

$$\frac{\lambda}{2+\log(1/\pi_{\min})} \le \frac{(1-2\pi_{\min})\lambda}{\log(1/\pi_{\min}-1)} \le \rho \le \frac{\lambda}{2}. \tag{3.52}$$

This in particular shows that all constants discussed in this paper, including also the half-step entropy contraction α, the full-step entropy contraction δ, and the modified log-Sobolev constant ρ_0, differ by at most a factor of $O(\log(1/\pi_{\min}))$ from each other.

In a recent work [34], Salez, Tikhomirov, and Youssef establish a surprising and remarkable comparison between the modified log-Sobolev constant ρ_0 and the log-Sobolev constant ρ. For a reversible Markov kernel P with respect to a probability measure π, define the sparsity parameter as

$$p_{\min} = \min_{(x,y)\in\mathcal{X}^2:\, P(x,y)>0} P(x,y). \tag{3.53}$$

Then, [34, Theorem 1] shows that

$$\frac{\rho_0}{20\log(1/p_{\min})} \le \rho \le \frac{\rho_0}{4}. \tag{3.54}$$

Hence, all entropy-related constants discussed in this paper, including also the half-step entropy contraction α and the full-step entropy contraction δ, differ by at most a factor of $O(\log(1/p_{\min}))$ from each other.

3.6 Comments on several previous works

[17, Prop. 5.1] claims that

$$c\rho_0 \le \alpha \tag{3.55}$$

for some universal constant $c > 0$. In fact, Eq. (3.55) fails for the random transposition model, showing that the claim must be incorrect.

Example 3.15 (Random transposition model)**.** Let n be a positive integer. Let $G = (V, E)$ be the Cayley graph on the symmetric group S_n generated by transpositions. That is, V is the set of permutations of $[n]$, and there is an edge between $\sigma, \tau \in V$ if and only if they differ by exactly two entries. The random transposition model is the lazy random walk on G (Definition 2.4). Considering the point measure at any $x \in \mathcal{X}$ gives $\alpha \le \frac{\log 2}{\log(n!)} = \Theta\left(\frac{1}{n\log n}\right)$. In fact, [9, Theorem 1.9] shows that this is tight, i.e, $\alpha = \frac{\log 2}{\log(n!)}$. On the other hand, [21] shows that $\rho_0 = \Theta\left(\frac{1}{n}\right)$. Therefore, as $n \to \infty$, we have $\frac{\rho_0}{\alpha} \to \infty$.

Our separations of α vs δ (Section 3.2) and δ vs ρ_0 (Section 3.3) also provide alternative counterexamples to the claim. We now explain briefly the issue in the proof of Eq. (3.55) in [17]. In their proof of Prop. 5.1, the authors apply a technical result, Lemma 5.2, which represents the (relative) entropy of a function f with expectation $\mathbb{E}_n(f) = 1$ (where n denotes the underlying probability measure) as an integral of the covariance between f_t and $\log(f_t)$ where $f_t = e^{-t}f + (1-e^{-t})$, $t \in \mathbb{R}_{\ge 0}$ represents an interpolation between f and 1. However, in the actual application of Lemma 5.2, the measure n is a conditional probability measure under which the expectation of f is no longer 1, and hence the interpolation function f_t should be replaced by $f_t = e^{-t}f + (1-e^{-t})\mathbb{E}_n(f)$; this would require the function f_t to depend on the conditioning and the proofs following afterwards no longer work.

[32, Prop. 4.3] claims that

$$\rho_0 \le 1 - c(1-\alpha) \tag{3.56}$$

for some universal constant $c > 0$. Our examples do not disprove the claim. However, the proof of [32, Theorem 4.4] is a generalization of that of [17, Prop. 5.1], so it has the same error. In particular, the last display of the proof of [32, Prop. 4.3] implies that $c\rho_0 \le \alpha$ for some constant $c > 0$, which we have shown to be incorrect. Therefore the proof of [32, Prop. 4.3] is incorrect and does not establish Eq. (3.56). It is unclear whether Eq. (3.56) as stated is correct.

4 Extremal functions

In this section we discuss another difference between the continuous-time entropy contraction constants and the discrete-time entropy contraction constants. It is known ([8]) that for irreducible (π, P), the log-Sobolev constant $\rho(\pi, P)$ and the modified log-Sobolev constant $\rho_0(\pi, P)$ satisfy a dichotomy: they are either equal to twice the Poincaré constant $\lambda(\pi, P)$ or achieved at a full-support function. We show that this is no longer true for the discrete-time entropy contraction constants $\alpha(\pi, K)$ and $\delta(\pi, P)$ by providing explicit examples whose extremal functions have non-full support.

4.1 Log-Sobolev constant ρ, modified log-Sobolev constant ρ_0, Poincaré constant λ

The extremal functions for the Poincaré constant $\lambda(\pi, P)$ are easy to describe (e.g., [8]). They are the (right) eigenfunctions of $-L$ corresponding to the eigenvalue λ. In particular, λ is always achieved at some non-constant function $f : \mathcal{X} \to \mathbb{R}$.

For the log-Sobolev constant ρ, [35, 8] show that for any reversible (π, P), either

(i) $\rho(\pi, P) = 2\lambda(\pi, P)$, or

(ii) there exists a non-constant function $f : \mathcal{X} \to \mathbb{R}_{\ge 0}$ with $\pi[f] = 1$ such that

$$\rho \operatorname{Ent}_\pi(f) = \mathcal{E}_{\pi,P}(\sqrt{f}, \sqrt{f}). \tag{4.1}$$

Furthermore, such f satisfies the equation

$$-L\sqrt{f} = \rho\sqrt{f}\log f. \tag{4.2}$$

If (π, P) is irreducible, then f has full support.

For the modified log-Sobolev constant $\rho_0(\pi, P)$, [8] shows that for any reversible (π, P), either

(i) $\rho_0(\pi, P) = 2\lambda(\pi, P)$, or

(ii) there exists a non-constant function $f : \mathcal{X} \to \mathbb{R}_{\ge 0}$ with $\pi[f] = 1$ such that

$$\rho_0 \operatorname{Ent}_\pi(f) = \mathcal{E}_{\pi,P}(f, \log f). \tag{4.3}$$

Furthermore, such f satisfies the equation

$$-Lf - fL(\log f) = \rho_0 f \log f. \tag{4.4}$$

If (π, P) is irreducible, then f has full support.

4.2 Half-step entropy contraction α and full-step entropy contraction δ

In this section we study the extremal distributions for η_{KL}, which includes both α and δ.

Lemma 4.1 (Extremal distributions for η_{KL})**.** *Let π be a distribution on $\mathcal{X}$ with full support and $K : \mathcal{X} \to \mathcal{Y}$ be a Markov kernel. Then either*

(i) $\eta_{\mathrm{KL}}(\pi, K) = \eta_{\chi^2}(\pi, K)$, or

(ii) there exists distribution ν on $\mathcal{X}$ such that

$$\eta_{\mathrm{KL}}(\pi, K) = \frac{D(\nu K \| \pi K)}{D(\nu \| \pi)}. \tag{4.5}$$

Proof. If (ii) does not happen, then there exists a sequence $\{f_n\}_n$ ($f_n : \mathcal{X} \to \mathbb{R}_{\geq 0}$, $\pi[f_n] = 1$, $f_n \neq \mathbb{1}$) satisfying $\|f_n - \mathbb{1}\|_\infty \to 0$ and $\frac{\mathrm{Ent}_{\pi K}(K_\pi^* f_n)}{\mathrm{Ent}_\pi(f_n)} \to \eta_{\mathrm{KL}}(\pi, K)$ as $n \to \infty$.

Write $f_n = \mathbb{1} + \epsilon_n g_n$, where $\epsilon_n \geq 0$, $\pi[g_n^2] = 1$. Note that the space $\mathcal{G} := \{g : \mathcal{X} \to \mathbb{R} : \pi[g] = 0, \pi[g^2] = 1\}$ is compact. By replacing the sequence $\{f_n\}_n$ with a subsequence, we can WLOG assume that there exists $g^* \in \mathcal{G}$ such that $\|g_n - g^*\|_\infty \to 0$ as $n \to \infty$.

Now let us prove that

$$\lim_{n\to\infty} \frac{\mathrm{Ent}_{\pi K}(K_\pi^* f_n)}{\mathrm{Ent}_\pi(f_n)} = \frac{\mathrm{Var}_{\pi K}(K_\pi^* g^*)}{\mathrm{Var}_\pi(g^*)}. \tag{4.6}$$

If Eq. (4.6) holds, then

$$\eta_{\chi^2}(\pi, K) = \sup_{g \in \mathcal{G}} \frac{\mathrm{Var}_{\pi K}(K_\pi^* g)}{\mathrm{Var}_\pi(g)} \geq \eta_{\mathrm{KL}}(\pi, K), \tag{4.7}$$

and (i) holds.

By Lemma 4.2,

$$\begin{aligned} &\|(1 + \epsilon_n g_n)\log(1 + \epsilon_n g_n) - (1 + \epsilon_n g^*)\log(1 + \epsilon_n g^*) - \epsilon_n(g_n - g^*)\|_\infty \\ &= O\left(\epsilon_n^2(\epsilon_n + \|g_n - g^*\|_\infty)\right). \end{aligned} \tag{4.8}$$

Taking expectation and using triangle inequality, we get

$$|\mathrm{Ent}_\pi(f_n) - \mathrm{Ent}_\pi(1 + \epsilon_n g^*)| = O\left(\epsilon_n^2(\epsilon_n + \|g_n - g^*\|_\infty)\right). \tag{4.9}$$

It is known that

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon^2} \mathrm{Ent}_\pi(1 + \epsilon g^*) = \frac{1}{2} \mathrm{Var}_\pi(g^*). \tag{4.10}$$

So

$$\lim_{n\to\infty} \frac{1}{\epsilon_n^2} \mathrm{Ent}_\pi(f_n) = \frac{1}{2} \mathrm{Var}_\pi(g^*). \tag{4.11}$$

Similarly

$$\lim_{n\to\infty} \frac{1}{\epsilon_n^2} \mathrm{Ent}_{\pi K}(K_\pi^* f_n) = \frac{1}{2} \mathrm{Var}_{\pi K}(K_\pi^* g^*). \tag{4.12}$$

This finishes the proof of Eq. (4.6). □

Lemma 4.2. *There exists $C > 0$ such that for $\epsilon, \epsilon' > 0$ small enough, for $x, y, z, w \in \mathbb{R}$ satisfying $|x|, |y|, |z|, |w| \leq \epsilon$, $|x - z|, |y - w| \leq \epsilon\epsilon'$, we have*

$$|(1 + x)\log(1 + y) - (1 + z)\log(1 + w) - (y - w)| \leq C\epsilon^2(\epsilon + \epsilon'). \tag{4.13}$$

Proof. First note that $|\log(1+y) - (y - y^2/2)| = O(\epsilon^3)$. Then

$$\begin{aligned} & |(1+x)\log(1+y) - (1+z)\log(1+w) - (y-w)| \\ &= |(1+x)(y-y^2/2) - (1+z)(w-w^2/2) - (y-w)| + O(\epsilon^3) \\ &= |xy - zw| + |y^2 - w^2|/2 + |xy^2|/2 + |zw^2|/2 + O(\epsilon^3) \\ &= |(x-z)y + z(y-w)| + |(y-w)(y+w)|/2 + O(\epsilon^2(\epsilon+\epsilon')) \\ &= O(\epsilon^2(\epsilon+\epsilon')). \end{aligned} \tag{4.14}$$

□

Unlike ρ and ρ_0 where the extremal functions (if they exist) have full support, the extremal distributions for η_{KL} may have non-full support.

Example 4.3 (Complete graph)**.** Let $n \geq 3$ be an integer. Let $\mathcal{X} = [n]$ and $\pi = \mathrm{Unif}(\mathcal{X})$. Let $K : \mathcal{X} \to \mathcal{X}$ be the (non-lazy) random walk on the complete graph. That is,

$$K(x,y) = \frac{1}{n-1}\mathbb{1}\{x \neq y\} \tag{4.15}$$

for $x, y \in \mathcal{X}$. [22, Prop. 33] proves that $\eta_{\mathrm{KL}}(\pi, K) = \frac{\log n - \log(n-1)}{\log n}$, and is achieved at and only at point distributions.

In the above example, K is not factorizable, so it corresponds to the half-step contraction coefficient α. The next example shows the extremal distributions for the full-step contraction coefficient δ can have non-full support.

Example 4.4 (Complete bipartite graph)**.** Let $n \geq 3$ be an integer. Let $G = K_{n,n}$ be the complete bipartite graph. That is, $V = [2] \times [n]$, and there is an edge between $(1,i)$ and $(2,j)$ for all $i, j \in [n]$. Let (π, P) be the random walk on G (Definition 2.4). Numerical computation suggests that $\eta_{\mathrm{KL}}(\pi, P) = \frac{\log n}{2\log(2n)}$, and equality is achieved at and only at point distributions. Prop. 4.5 proves this observation for $n = 3$.

Proposition 4.5 (Complete bipartite graph)**.** *Let π, P be as in Example 4.4. For $n = 3$, $\eta_{\mathrm{KL}}(\pi, P) = \frac{\log n}{2\log(2n)}$, and equality is achieved at and only at point distributions.*

Proof. Let ν be the point distribution on any $x \in \mathcal{X}$. Then

$$D(\nu\|\pi) = \log(2n), \tag{4.16}$$

$$D(\nu P\|\pi) = \frac{1}{2}\log n. \tag{4.17}$$

So

$$\eta_{\mathrm{KL}}(\pi, P) \geq \frac{D(\nu P\|\pi)}{D(\nu\|\pi)} = \frac{\log n}{2\log(2n)}. \tag{4.18}$$

Note that this holds for any $n \geq 3$.

We prove that for $n = 3$, the point distributions are the only maximizers. We represent a distribution ν using a tuple (t, ν_1, ν_2), where $t \in [0,1]$ and ν_i ($i = 1, 2$) is a distribution on $\mathcal{X}_i = \{i\} \times [n] \subseteq \mathcal{X}$. Given ν, the corresponding tuple $\phi(\nu) = (\nu(\mathcal{X}_1), \nu|_{\mathcal{X}_1}, \nu|_{\mathcal{X}_2})$, where $\nu|_{\mathcal{X}_i}$ denotes the conditional distribution (if $\nu(\mathcal{X}_i) = 0$, then choose an arbitrary distribution on $\mathcal{X}_i$). Given a tuple (t, ν_1, ν_2), the corresponding distribution ν is $\psi(t, \nu_1, \nu_2) = t\nu_1 + (1-t)\nu_2$. Let $\pi_i = \mathrm{Unif}(\mathcal{X}_i)$ for $i = 1, 2$. By symmetry, we only need to consider the case $\frac{1}{2} \leq t \leq 1$.

Under the tuple parametrization, we have

$$D(\psi(t, \nu_1, \nu_2)\|\pi) = d_{\mathrm{KL}}\left(t\|\frac{1}{2}\right) + tD(\nu_1\|\pi_1) + (1-t)D(\nu_2\|\pi_2), \tag{4.19}$$

$$\phi(\psi(t, \nu_1, \nu_2)P) = \left(\frac{1}{2}, t\nu_1 + (1-t)\pi_1, (1-t)\nu_2 + t\pi_2\right), \tag{4.20}$$

where $d_{\mathrm{KL}}(x\|y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ is the binary KL divergence function. So for $\nu = \psi(t,\nu_1,\nu_2)$, we have

$$\frac{D(\nu P\|\pi)}{D(\nu\|\pi)} = \frac{1}{2}\cdot\frac{D(t\nu_1+(1-t)\pi_1\|\pi_1) + D((1-t)\nu_2+t\pi_2\|\pi_2)}{d_{\mathrm{KL}}\left(t\|\frac{1}{2}\right) + tD(\nu_1\|\pi_1) + (1-t)D(\nu_2\|\pi_2)}. \tag{4.21}$$

When $t=1$, Eq. (4.21) simplifies to

$$\frac{D(\nu P\|\pi)}{D(\nu\|\pi)} = \frac{1}{2}\cdot\frac{D(\nu_1\|\pi_1)}{\log 2 + D(\nu_1\|\pi_1)} \tag{4.22}$$

which is maximized when and only when ν_1 is a point measure, taking value $\frac{\log n}{2\log(2n)}$. Note that when $t=1$ and ν_1 is a point measure, $\nu = \psi(t,\nu_1,\nu_2)$ does not depend on ν_2 and is always a point measure.

Now assume that $\frac{1}{2}\le t<1$. By [22, Prop. 17], for fixed $D(\nu_1\|\pi_1)$ (resp. $D(\nu_2\|\pi_2)$), the maximum value of $D(t\nu_1+(1-t)\pi_1\|\pi_1)$ (resp. $D(\nu_2+t\pi_2\|\pi_2)$) is uniquely achieved (up to permutation of the alphabet) at a distribution of form $\left(x,\frac{1-x}{n-1},\dots,\frac{1-x}{n-1}\right)$ where $\frac{1}{n}\le x\le 1$. Therefore, we can assume WLOG that $\nu_1 = \left(x,\frac{1-x}{n-1},\dots,\frac{1-x}{n-1}\right)$, $\nu_2 = \left(y,\frac{1-y}{n-1},\dots,\frac{1-y}{n-1}\right)$ for some $x,y\in\left[\frac{1}{n},1\right]$. In this case, we can simplify Eq. (4.21) as

$$\frac{D(\nu P\|\pi)}{D(\nu\|\pi)} = \frac{1}{2}\cdot\frac{f_n\left(tx+\frac{1-t}{n}\right)+f_n\left((1-t)y+\frac{t}{n}\right)}{f_2(t)+tf_n(x)+(1-t)f_n(y)} =: F_n(t,x,y), \tag{4.23}$$

where

$$f_n(x) = D\left(\left(x,\frac{1-x}{n-1},\dots,\frac{1-x}{n-1}\right)\|\,\mathrm{Unif}([n])\right) = \log n + x\log x + (1-x)\log\frac{1-x}{n-1}. \tag{4.24}$$

Therefore, computing $\eta_{\mathrm{KL}}(\pi,P)$ is equivalent to computing

$$\sup_{\substack{\frac{1}{2}\le t\le 1\\ \frac{1}{n}\le x,y\le 1\\ (t,x,y)\ne\left(\frac{1}{2},\frac{1}{n},\frac{1}{n}\right)}} F_n(t,x,y) \tag{4.25}$$

We have reduced the original problem of computing $\eta_{\mathrm{KL}}(\pi,K)$, which is a priori an $(2n-1)$-dimensional optimization problem, to a optimization problem with three real variables $t\in[0,1], x,y\in\left[\frac{1}{n},1\right]$. The following lemma helps us reduce it further to two real variables.

Lemma 4.6. *For any $n\ge 3$, there exists $t^*>\frac{1}{2}$ such that*

$$\sup_{\substack{0<t<t^*\\ \frac{1}{n}<x\le 1}} \frac{f_n\left(tx+\frac{1-t}{n}\right)}{tf_n(x)} < \frac{\log n}{\log(2n)}. \tag{4.26}$$

Proof. By [22, Eqs. (23) and (27)], for fixed $0\le t\le 1$, we have

$$\sup_{\frac{1}{n}<x\le 1} \frac{f_n\left(tx+\frac{1-t}{n}\right)}{tf_n(x)} \le t^{\frac{1}{\log n}}. \tag{4.27}$$

For any $t<\exp\left((\log n)\log\frac{\log n}{\log(2n)}\right)$, we have

$$t^{\frac{1}{\log n}} < \frac{\log n}{\log(2n)}. \tag{4.28}$$

Furthermore, notice that $\exp\left((\log n)\log\frac{\log n}{\log(2n)}\right) > \frac{1}{2}$ for all $n\ge 3$. So we can take t^* to be any number smaller than $\exp\left((\log n)\log\frac{\log n}{\log(2n)}\right)$. □

Note that all arguments until this point work for any $n \geq 3$. From now on we will use the assumption $n = 3$. For $n = 3$, we take $t^* = 0.58 < \exp\left((\log 3) \log \frac{\log 3}{\log 6}\right)$ and apply Lemma 4.6 to Eq. (4.25). For $\frac{1}{2} \leq t \leq t^*$, we can apply Lemma 4.6 to (t, x) and $(1-t, y)$ and get

$$\sup_{\substack{\frac{1}{2} \leq t \leq t^* \\ \frac{1}{3} \leq x, y \leq 1 \\ (t,x,y) \neq \left(\frac{1}{2}, \frac{1}{3}, \frac{1}{3}\right)}} F_3(t, x, y) < \frac{\log 3}{2 \log 6} \tag{4.29}$$

For $t^* < t < 1$, we have

$$\sup_{\substack{t^* < t < 1 \\ \frac{1}{3} \leq x, y \leq 1}} F_3(t, x, y) \geq \lim_{t \to 1^-} F_3\left(t, 1, \frac{1}{3}\right) = \frac{\log 3}{2 \log 6}. \tag{4.30}$$

Applying Lemma 4.6 to $(1-t, y)$ we see that for any (t, x, y) with $t^* < t < 1$ and $F_3(t, x, y) \geq \frac{\log 3}{2 \log 6}$, we have $F_3(t, x, y) \leq F_3\left(t, x, \frac{1}{3}\right)$. So

$$\sup_{\substack{t^* < t < 1 \\ \frac{1}{3} \leq x, y \leq 1}} F_3(t, x, y) = \sup_{\substack{t^* < t < 1 \\ \frac{1}{3} \leq x \leq 1}} F_3\left(t, x, \frac{1}{3}\right) = \sup_{\substack{t^* < t < 1 \\ \frac{1}{3} \leq x \leq 1}} \frac{1}{2} \cdot \frac{f_3\left(tx + \frac{1-t}{3}\right)}{f_2(t) + t f_3(x)}. \tag{4.31}$$

In the following, we prove that for $t^* < t < 1$, $\frac{1}{3} \leq x \leq 1$, we have

$$G_3(t, x) = \frac{\log 6}{\log 3} \cdot f_3\left(tx + \frac{1-t}{3}\right) - (f_2(t) + t f_3(x)) < 0. \tag{4.32}$$

Note that Eq. (4.32) implies the desired result by Eqs. (4.29) and (4.31).

Case 1: $t \geq 0.999$**,** $x \geq 0.999$**.** For simplicity of notation, write $a = 1 - t$ and $b = 1 - x$. Then $0 < a \leq 0.001$ and $0 \leq b \leq 0.001$. Let $c = 1 - \left(tx + \frac{1-t}{3}\right) = \frac{2}{3}a + b - ab$. Write $g_n(u) = -(1-u)\log(1-u) - u \log \frac{u}{n} \geq 0$ for $u \in [0, 1]$. Then

$$\begin{aligned} & G_3(1-a, 1-b) \\ &= \frac{\log 6}{\log 3} \cdot (\log 3 - g_2(c)) - (\log 2 - g_1(a) + t(\log 3 - g_2(b))) \\ &\leq \frac{\log 6}{\log 3} \cdot (\log 3 - g_2(c)) - (\log 2 - g_1(a) - a \log 3 + \log 3 - g_2(b)) \\ &= g_3(a) + g_2(b) - \frac{\log 6}{\log 3} \cdot g_2(c). \end{aligned} \tag{4.33}$$

For $0 \leq w \leq 0.001$, we have

$$0.999w \leq -(1-w)\log(1-w) \leq w. \tag{4.34}$$

So

$$g_3(a) \leq a \log \frac{3e}{a}, \qquad g_2(b) \leq b \log \frac{2e}{b}, \qquad g_2(c) \geq c\left(0.999 + \log \frac{2}{c}\right) =: h(c). \tag{4.35}$$

Because $h(c)$ is concave and increasing for $0 \leq c \leq 0.002$, we have

$$h(c) \geq \frac{1}{2} h\left(\frac{4}{3}a\right) + \frac{1}{2} h(2b(1-a)) \geq \frac{1}{2} h\left(\frac{4}{3}a\right) + \frac{1}{2} h(1.998b). \tag{4.36}$$

For $0 < a \le 0.001$ and $0 \le b \le 0.001$, we have

$$a \log \frac{3e}{a} < \frac{1}{2} h\left(\frac{4}{3}a\right), \qquad b \log \frac{2e}{b} \le \frac{1}{2} h(1.998b). \tag{4.37}$$

So $G_3(t,x) < 0$ for $0.999 \le t < 1$, $0.999 \le x \le 1$. This finishes the proof for Case 1.

Case 2: $t \le 0.999$ or $x \le 0.999$. Let $A = \left([t^*, 1] \times \left[\frac{1}{3}, 1\right]\right) \setminus ([0.999, 1] \times [0.999, 1])$. Our goal is to prove that $G_3(t,x) < 0$ for all $(t,x) \in A$. The proof strategy is as follows. We choose $\epsilon, \delta > 0$ and a finite set $A^* \subseteq A$ such that the following are true.

(a) $G_3(t,x) < -\epsilon$ for all $(t,x) \in A^*$;

(b) For any $(t,x) \in A$, there exists $(t^*, x^*) \in A^*$ such that $\max\{|t - t^*|, |x - x^*|\} \le \delta$;

(c) For any $(t,x), (t',x') \in A$, if $\max\{|t - t'|, |x - x'|\} \le \delta$, then $|G_3(t,x) - G_3(t',x')| \le \epsilon$.

If all three items hold, then they imply our goal.

We take $\epsilon = 0.00078$, $\delta = 10^{-5}$, $A^* = A \cap \left(10^{-5}\mathbb{Z} \times 10^{-5}\mathbb{Z}\right)$. (a) is verified using a computer program by iterating over all points in A^*.[6] (b) is immediate by our choice of A^*. It remains to prove (c). Note that on A, $f_3\left(tx + \frac{1-t}{3}\right)$ and $f_2(t) + t f_3(x)$ are non-decreasing in both t and x. By convexity and monotonicity,

$$\sup_{\substack{\frac{1}{3} \le u, u' \le 1 \\ |u - u'| \le \delta}} |f_3(u) - f_3(u')| = |f_3(1) - f_3(1 - \delta)| \le 0.00014, \tag{4.38}$$

$$\sup_{\substack{\frac{1}{2} \le u, u' \le 1 \\ |u - u'| \le \delta}} |f_2(u) - f_2(u')| = |f_2(1) - f_2(1 - \delta)| \le 0.00013. \tag{4.39}$$

So for $(t,x), (t',x') \in A$ with $\max\{|t - t^*|, |x - x^*|\} \le \delta$, we have

$$\begin{aligned} |G_3(t,x) - G_3(t,x')| &\le \max\left\{\frac{\log 6}{\log 3} \cdot \left| f_3\left(tx + \frac{1-t}{3}\right) - f_3\left(tx' + \frac{1-t}{3}\right)\right|, \right. \\ &\qquad \left. |(f_2(t) + t f_3(x)) - (f_2(t) + t f_3(x'))|\right\} \\ &\le \max\left\{\frac{\log 6}{\log 3} \cdot 0.00014, 0.00014\right\} \le 0.00023, \end{aligned} \tag{4.40}$$

$$\begin{aligned} |G_3(t,x') - G_3(t',x')| &\le \max\left\{\frac{\log 6}{\log 3} \cdot \left| f_3\left(tx' + \frac{1-t}{3}\right) - f_3\left(t'x' + \frac{1-t'}{3}\right)\right|, \right. \\ &\qquad \left. |(f_2(t) + t f_3(x')) - (f_2(t') + t' f_3(x'))|\right\} \\ &\le \max\left\{\frac{\log 6}{\log 3} \cdot 0.00014, 0.00013 + \delta \log 3\right\} \le 0.00023. \end{aligned} \tag{4.41}$$

Therefore

$$|G_3(t,x) - G_3(t',x')| \le 0.00023 + 0.00023 = 0.00046 < \epsilon. \tag{4.42}$$

This proves (c), thus finishing the proof for Case 2.

Eq. (4.32) follows by combining the two cases. This finishes the proof that $\eta_{\mathrm{KL}}(\pi, P) = \frac{\log 3}{2 \log 6}$ and equality is achieved at and only at point distributions. □

We conjecture that the statement of Prop. 4.5 holds for all $n \ge 3$. Our proof strategy for the $n = 3$ case could potentially be generalized to larger n. However, it is unclear to us how to prove the result for all $n \ge 3$.

[6] See supplementary material for the program used for verification.

Finally, we provide sufficient conditions for the extremal distribution for η_{KL} to have full support. This result can be contrasted with Examples 4.3 and 4.4, and [9] where it is shown that the half-step entropy contraction for many natural chains (e.g., the random transposition model) has point measures as their only extremizers.

Lemma 4.7. *Let (π, P) be a reversible pair where P is irreducible. Suppose that either*

(i) $P(x,x) \geq \frac{1}{2}$ for all $x \in \mathcal{X}$, and $\eta_{\mathrm{KL}}(\pi, P) > \frac{1}{2}$, or

(ii) $P(x,y) > 0$ for all $x, y \in \mathcal{X}$.

Let ν be a distribution on $\mathcal{X}$ satisfying $\eta_{\mathrm{KL}}(\pi, P) = \frac{D(\nu P \| \pi P)}{D(\nu \| \pi)}$. Then ν has full support.

Proof. Let $f = \frac{d\nu}{d\pi}$ be the Radon-Nikodym derivative. For the sake of contradiction assume that $\operatorname{supp} f \neq \mathcal{X}$. Choose $a \in \mathcal{X} - \operatorname{supp} f$ and $b \in \operatorname{supp} f$ such that $P(a,b) > 0$. Because P is irreducible, such (a,b) always exists.

Let $h = \mathbb{1}_a - \frac{\pi(a)}{\pi(b)} \mathbb{1}_b$ and $f_\epsilon = f + \epsilon h$. For small enough $\epsilon > 0$, we have $f_\epsilon \geq 0$ and $\pi[f_\epsilon] = 1$. We prove that for $\epsilon > 0$ small enough, we have

$$\frac{d}{d\epsilon} \frac{\operatorname{Ent}_\pi(Pf_\epsilon)}{\operatorname{Ent}_\pi(f_\epsilon)} > 0. \tag{4.43}$$

Computation shows that

$$\frac{d}{d\epsilon} \frac{\operatorname{Ent}_\pi(Pf_\epsilon)}{\operatorname{Ent}_\pi(f_\epsilon)} = \frac{1}{\operatorname{Ent}_\pi(f_\epsilon)^2} \left(\pi[Ph \log(Pf_\epsilon)] \pi[f_\epsilon \log f_\epsilon] - \pi[h \log f_\epsilon] \pi[Pf_\epsilon \log(Pf_\epsilon)] \right). \tag{4.44}$$

Expanding near $\epsilon = 0$ gives

$$\pi[h \log f_\epsilon] = \pi(a) \log \epsilon + O(1), \tag{4.45}$$

$$\begin{aligned} \pi[Ph \log(Pf_\epsilon)] &= \sum_{j \in \mathcal{X}} \pi(j) \left(P(j,a) - P(j,b) \frac{\pi(a)}{\pi(b)} \right) \log(Pf_\epsilon(j)) \\ &= \sum_{j \in \mathcal{X} - \operatorname{supp}(Pf)} \pi(j) P(j,a) \log(P(j,a)\epsilon) + O(1) \\ &= \pi(a) P(a, \mathcal{X} - \operatorname{supp}(Pf)) \log \epsilon + O(1), \end{aligned} \tag{4.46}$$

$$\pi[f_\epsilon \log f_\epsilon] = \operatorname{Ent}_\pi(f) + o(1), \tag{4.47}$$

$$\pi[Pf_\epsilon \log(Pf_\epsilon)] = \operatorname{Ent}_\pi(Pf) + o(1). \tag{4.48}$$

Case (i). Because $P(a,b) > 0$, we have $a \in \operatorname{supp}(Pf)$. Because P is lazy, we have

$$P(a, \mathcal{X} - \operatorname{supp}(Pf)) \leq 1 - P(a,a) \leq \frac{1}{2}. \tag{4.49}$$

By assumption, $\frac{\operatorname{Ent}_\pi(Pf)}{\operatorname{Ent}_\pi(f)} = \eta_{\mathrm{KL}}(\pi, P) > \frac{1}{2}$. Therefore Eq. (4.43) holds for $\epsilon > 0$ small enough.

Case (ii). In this case, $P(a, \mathcal{X} - \operatorname{supp}(Pf)) = 0$. So Eq. (4.43) holds for $\epsilon > 0$ small enough. □

Supplementary Material

Numerical verification of Proof of Lemma 4.6, Case 2, Item (a). We provide `C++` code that verifies Proof of Lemma 4.6, Case 2, Item (a).

References

[1] Rudolf Ahlswede and Peter Gács, *Spreading of sets in product spaces and hypercontraction of the Markov operator*, The Annals of Probability (1976), 925–939.

[2] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong, *Entropic independence: optimal mixing of down-up random walks*, Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC), 2022, pp. 1418–1430.

[3] Nima Anari, Kuikui Liu, Shayan Oveis Gharan, and Cynthia Vinzant, *Log-concave polynomials II: high-dimensional walks and an FPRAS for counting bases of a matroid*, Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC), 2019, pp. 1–12.

[4] Abraham Berman and Naomi Shaked-Monderer, *Completely positive matrices*, World Scientific, 2003.

[5] Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda, *On mixing of Markov chains: coupling, spectral independence, and entropy factorization*, Electronic Journal of Probability **27** (2022), 1–42.

[6] Antonio Blanca, Pietro Caputo, Daniel Parisi, Alistair Sinclair, and Eric Vigoda, *Entropy decay in the Swendsen-Wang dynamics on $\mathbb{Z}^d$*, Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC), 2021, pp. 1551–1564.

[7] Sergej G. Bobkov and Friedrich Götze, *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, Journal of Functional Analysis **163** (1999), no. 1, 1–28.

[8] Sergey G. Bobkov and Prasad Tetali, *Modified logarithmic Sobolev inequalities in discrete settings*, Journal of Theoretical Probability **19** (2006), no. 2, 289–336.

[9] Alexandre Bristiel and Pietro Caputo, *Entropy inequalities for random walks and permutations*, Annales de l'Institut Henri Poincare (B) Probabilites et statistiques **60** (2024), no. 1, 54–81.

[10] Pietro Caputo, *Lecture notes on entropy and Markov chains*, Preprint (2023), Available at: http://www.mat.uniroma3.it/users/caputo/entropy.pdf.

[11] Pietro Caputo, Paolo Dai Pra, and Gustavo Posta, *Convex entropy decay via the Bochner-Bakry-Emery approach*, Annales de l'IHP Probabilités et statistiques **45** (2009), no. 3, 734–753.

[12] Pietro Caputo, Florentin Münch, and Justin Salez, *Entropy and curvature: beyond the Peres-Tetali conjecture*, arXiv preprint arXiv:2401.17148 (2024).

[13] Mu-Fa Chen, *Variational formulas of Poincaré-type inequalities for birth-death processes*, Acta Mathematica Sinica **19** (2003), no. 4, 625–644.

[14] _____, *Poincaré-type inequalities in dimension one*, Eigenvalues, Inequalities, and Ergodic Theory (2005), 113–130.

[15] Yuansi Chen and Ronen Eldan, *Localization schemes: A framework for proving mixing bounds for Markov chains*, 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2022, pp. 110–122.

[16] Zongchen Chen, Kuikui Liu, and Eric Vigoda, *Optimal mixing of Glauber dynamics: Entropy factorization via high-dimensional expansion*, Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC), 2021, pp. 1537–1550.

[17] Pierre Del Moral, Michel Ledoux, and Laurent Miclo, *On contraction properties of Markov kernels*, Probability theory and related fields **126** (2003), no. 3, 395–420.

[18] Persi Diaconis and Laurent Saloff-Coste, *Logarithmic Sobolev inequalities for finite Markov chains*, The Annals of Applied Probability **6** (1996), no. 3, 695–750.

[19] Roland L. Dobrushin, *Central limit theorem for nonstationary Markov chains. I*, Theory of Probability & Its Applications **1** (1956), no. 1, 65–80.

[20] James Allen Fill, *Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process*, The annals of applied probability (1991), 62–87.

[21] Fuqing Gao and Jeremy Quastel, *Exponential decay of entropy in the random transposition and Bernoulli-Laplace models*, The Annals of Applied Probability **13** (2003), no. 4, 1591–1600.

[22] Yuzhou Gu and Yury Polyanskiy, *Non-linear log-Sobolev inequalities for the Potts semigroup and applications to reconstruction problems*, Communications in Mathematical Physics **404** (2023), no. 2, 769–831.

[23] Tali Kaufman and David Mass, *High dimensional random walks and colorful expansion*, 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2017.

[24] David A. Levin and Yuval Peres, *Markov chains and mixing times*, vol. 107, American Mathematical Soc., 2017.

[25] Marvin Marcus and Morris Newman, *The permanent of a symmetric matrix*, Notices Amer. Math. Soc **8** (1961), 595.

[26] John E. Maxfield and Henryk Minc, *On the matrix equation $X'X = A$*, Proceedings of the Edinburgh Mathematical Society **13** (1962), no. 2, 125–129.

[27] Laurent Miclo, *Remarques sur l'hypercontractivité et l'évolution de l'entropie pour des chaînes de Markov finies*, Séminaire de Probabilités XXXI, Springer, 1997, pp. 136–167.

[28] _____, *An example of application of discrete hardy's inequalities*, Markov Process. Related Fields **5** (1999), no. 3, 319–330.

[29] Florentin Münch, *Ollivier curvature, isoperimetry, concentration, and log-Sobolev inequalitiy*, arXiv preprint arXiv:2309.06493 (2023).

[30] Yury Polyanskiy and Yihong Wu, *Strong data-processing inequalities for channels and bayesian networks*, Convexity and Concentration, Springer, 2017, pp. 211–249.

[31] _____, *Information theory: From coding to learning*, Cambridge university press, 2024.

[32] Maxim Raginsky, *Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels*, IEEE Transactions on Information Theory **62** (2016), no. 6, 3355–3389.

[33] Cyril Roberto, *Inégalités de Hardy et de Sobolev logarithmiques*, Ph.D. thesis, PhD thesis, 2001.

[34] Justin Salez, Konstantin Tikhomirov, and Pierre Youssef, *Upgrading MLSI to LSI for reversible Markov chains*, Journal of Functional Analysis **285** (2023), no. 9, 110076.

[35] Laurent Saloff-Coste, *Lectures on finite markov chains*, Lecture notes in mathematics (1997), 301–413.

[36] Richard Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, The annals of mathematical statistics **35** (1964), no. 2, 876–879.

[37] Konstantin Tikhomirov and Pierre Youssef, *Regularized modified log-Sobolev inequalities and comparison of Markov chains*, The Annals of Probability **52** (2024), no. 4, 1201–1224.

Acknowledgments. The work of Y.G. was supported in part by the National Science Foundation under Grant No. DMS-1926686. The work of Y.P. was supported in part by the MIT-IBM Watson AI Lab and by the National Science Foundation under Grant No CCF-2131115. We thank anonymous reviewers for helpful comments and suggestions.