E_γ-Mixing Time

Behnoosh Zamanlooy*, Shahab Asoodeh*, Mario Diaz† and Flavio P. Calmon‡
* Department of Computing and Software, McMaster University, Canada, *{zamanlob, asoodeh}@mcmaster.ca*
† IIMAS, Universidad Nacional Autónoma de México, Mexico, *mario.diaz@sigma.iimas.unam.mx*
‡ School of Engineering and Applied Sciences, Harvard University, USA, *flavio@seas.harvard.edu*

***Abstract*—We investigate the mixing times of Markov kernels under E_γ-divergence. We demonstrate that the zero-error E_γ-mixing time, for any $\gamma > 1$, of irreducible and aperiodic Markov chains, is bounded, a property that is not shared by the TV-mixing time. We further obtain upper bounds on the E_γ-mixing times for a broad family of contractive Markov kernels via a new *non-linear* strong data processing inequality for the E_γ-divergence. We apply our results to derive new bounds for the local differential privacy guarantees offered by the sequential application of a privacy mechanism to data.**

I. Introduction

Let $G = (\mathcal{X}, \mathcal{E})$ be a connected graph with the vertex set $\mathcal{X}$ and the edge set $\mathcal{E} \subseteq \mathcal{X} \times \mathcal{X}$. Consider a discrete-time Markov chain $\{X_n\}_n$ with states in $\mathcal{X}$ with the corresponding one-step transition probability matrix specified by a Markov kernel K. Let P_0 be the distribution of the initial state X_0 and P_n denote the distribution of X_n, the state at time n, so that $P_n = P_0\mathsf{K}^n$, where K^n represents the n-step transition probability matrix. It is widely known that if the Markov chain is irreducible and aperiodic, then P_n converges to the (unique) stationary distribution Q^*, that is, $P_0\mathsf{K}^n$ approaches Q^* as n tends to infinity. The rate of such convergence is captured by the *mixing time* $t_{\mathsf{D}}^{\mathsf{K}}(\varepsilon)$, which is the smallest $n \geq 1$ such that $\mathsf{D}(P_n\|Q^*) \leq \varepsilon$ for a given divergence measure D. Mixing time has been extensively studied in the literature under different divergence measures, such as total variation distance TV [1], Rényi and KL divergences [2], and χ^2-divergence [3]. We refer interested readers to [1, 4] for a comprehensive exposition of existing mixing time results. It is customary to fix some value of ε (for discrete-time chains, a common choice is $1/4$), and to investigate the scaling of the mixing time in terms of properties of K.

In this paper, we examine the mixing time of E_γ-divergence, denoted by $t_\gamma^{\mathsf{K}}(\varepsilon)$. A formal definition of E_γ-divergence is given in Section II; note that total variation distance corresponds to E_1 (see [5] for more properties of E_γ-divergence). Our motivation for studying mixing time under E_γ-divergence is twofold. First, the widely recognized standard for privacy in machine learning, namely, differential privacy [6], can be expressed in terms of E_γ-divergence. Properties of E_γ-divergence have been successfully used to analyze differentially private algorithms in several machine learning settings, see, e.g., [7–12]. Here, we use this connection to study properties of *locally* differentially private mechanisms, viewed as Markov kernels, through their mixing time[1] (see Section VI for more details). Second, a large family of f-divergences can be represented by E_γ-divergence, see [16, Corollary 3.7]. Thus, by studying the mixing time under E_γ-divergence and bounding $t_\gamma^{\mathsf{K}}(\varepsilon)$, we can bound the mixing time for a range of f-divergences as a by-product.

Our framework for characterizing $t_\gamma^{\mathsf{K}}(\varepsilon)$ relies on the theory of strong data processing inequalities (SDPI), originally developed by Ahlswede and Gács [17]. The key quantity in this framework is $\eta_{\mathsf{D}}(\mathsf{K})$, the contraction coefficient of a Markov kernel K under a divergence measure D, which quantifies the extent to which the data processing inequality for K can be improved (see Section II for the definition). The connection between mixing time and SDPI has been already established, see, e.g., [18, 19]. In particular, a direct application of SDPI implies that

$$t_{\mathsf{TV}}^{\mathsf{K}}(\varepsilon) \leq \frac{\log(\varepsilon)}{\log(\eta_{\mathsf{TV}}(\mathsf{K}))}, \tag{1}$$

where $\eta_{\mathsf{TV}}(\mathsf{K})$ is the contraction coefficient of K under total variation distance.

Our contributions can be summarized as follows:

- We first demonstrate that $t_\gamma^{\mathsf{K}}(\varepsilon) < \infty$ for any $\gamma > 1$ and any aperiodic and irreducible kernels *even* for $\varepsilon = 0$. This result, which is in stark contrast with $t_{\mathsf{TV}}^{\mathsf{K}}(\varepsilon)$, serves as the main motivation for the ensuing technical results. In particular, we consider $\varepsilon = 0$ for studying t_γ^{K}.
- In Theorem 2, we derive an upper bound on $t_\gamma^{\mathsf{K}}(0)$ for any kernel K satisfying $\mathsf{K}(y|x) > 0$ for all $x, y \in \mathcal{X}$, which we call *contractive*. We achieve this by combining the two-point characterization of the contraction coefficient of K under E_γ-divergence, recently proved in [20] with some functional properties of E_γ-divergence. This bound involves $\min_{x\in\mathcal{X}} Q^*(x)$ which, although strictly positive, can be arbitrarily small.
- We then develop a framework for the *non-linear* SDPI under E_γ-divergence and use it to establish a tighter bound for $t_\gamma^{\mathsf{K}}(0)$ independent of Q^* (Theorem 4). Such Non-linear SDPI provids a sharp bound on $\mathsf{E}_\gamma(P\mathsf{K}\|Q\mathsf{K})$, the E_γ-divergence between $P\mathsf{K}$ and $Q\mathsf{K}$, in terms of

The work of B. Zamanlooy was supported in part by the Ontario Graduate Scholarship. The work of S. Asoodeh was supported in part by the NSERC of Canada. The work of M. Diaz was supported in part by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) under grant IN103224. The work of F. Calmon was supported in part by the NSF awards CAREER-1845852, CIF-1900750, CIF-2231707, and CIF-2312667.

[1] It is worth noting that mixing time under Rényi divergence has been instrumental in some recent discoveries in differential privacy applications in machine learning, see [13–15].

$\mathsf{E}_\gamma(P\|Q)$ for any distributions P and Q and contractive kernels K. As stated earlier, this framework allows us to obtain bounds on mixing time under a large family of f-divergences. For instance, we show that our bound on $t_\gamma^{\mathsf{K}}(0)$ directly leads to a bound on $t_{\mathsf{TV}}^{\mathsf{K}}(\varepsilon)$ for a certain choice of γ.

- In Section VI, we use the connection between E_γ-divergence and differential privacy to study *privacy amplification by composition* [7] for the local differential privacy (LDP) setting. In particular, we answer the following two questions in Theorem 5 and Lemma 3, respectively: (1) If K_i is an ε_i-locally differentially private (ε_i-LDP for short) mechanism for $i \in [n] \coloneqq \{1, 2, \ldots, n\}$, then what is the privacy guarantee of their composition $\mathsf{K}_1 \circ \cdots \circ \mathsf{K}_n$? and (2) If K is (ε, δ)-LDP, what is the smallest n such that K^n is ε'-LDP for a given $\varepsilon' < \varepsilon$?

Notation. We use upper-case letters (e.g., X) to denote random variables and calligraphic letters to represent their support sets (e.g., $\mathcal{X}$). A Markov kernel (or channel) $\mathsf{K} : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ is specified by a collection of distributions $\{\mathsf{K}(\cdot|x) \in \mathcal{P}(\mathcal{Z}) : x \in \mathcal{X}\}$. Given such kernel K and $P \in \mathcal{P}(\mathcal{X})$, we denote by $P\mathsf{K}$ the output distribution of K when the input is distributed according to P, given by $P\mathsf{K}(A) \coloneqq \int P(\mathrm{d}x)\mathsf{K}(A|x)$ for $A \subset \mathcal{Z}$. We define $[n] \coloneqq \{1, \ldots, n\}$, $(a)_+ \coloneqq \max\{a, 0\}$, and $a \vee b = \max\{a, b\}$. Finally, for $a \in \mathbb{R}$, we define

$$\zeta(a) \coloneqq \frac{a-1}{a+1}. \tag{2}$$

II. Preliminaries

A. E_γ Divergence

Let $\mathcal{X}$ be an arbitrary set. Given any pair of distributions $P, Q \in \mathcal{P}(\mathcal{X})$, we define their E_γ-divergence for $\gamma \in (0, \infty)$ as

$$\mathsf{E}_\gamma(P\|Q) = \int_{\mathcal{X}} \left(\frac{\mathrm{d}P_0}{\mathrm{d}Q} - \gamma\right)_+ \mathrm{d}Q + P_\perp\big(\mathrm{Supp}(Q)^c\big) - (1-\gamma)_+,$$

where $P = P_0 + P_\perp$ with $P_0 \ll Q$ and $P_\perp \perp Q$ (i.e., the Lebesgue decomposition of P with respect to Q). When $\mathcal{X}$ is a discrete set, the definition of the E_γ-divergence becomes

$$\mathsf{E}_\gamma(P\|Q) = \sum_{x \in \mathcal{X}} \big(P(x) - \gamma Q(x)\big)_+ - (1-\gamma)_+.$$

Since E_γ-divergence satisfies the reciprocity relation (see, e.g., [21]), that is,

$$\mathsf{E}_{1/\gamma}(Q\|P) = \frac{\mathsf{E}_\gamma(P\|Q)}{\gamma} \qquad \text{for all } \gamma \in (0, \infty),$$

we focus on the E_γ-divergence for $\gamma \geq 1$. It is important to remark that the E_γ-divergences with $\gamma = 1$ coincides with the total variation distance, i.e., $\mathsf{E}_1(P\|Q) = \mathsf{TV}(P, Q)$.

B. Strong Data Processing Inequalities

The E_γ-divergence (as any other f-divergence) satisfies the data processing inequality [22], i.e., $\mathsf{E}_\gamma(P\mathsf{K}\|Q\mathsf{K}) \leq \mathsf{E}_\gamma(P\|Q)$ for any pair of probability distributions (P, Q), any Markov kernel K, and any $\gamma \geq 0$. The *contraction coefficient* of a Markov kernel K under E_γ-divergence $\eta_\gamma(\mathsf{K})$ is the smallest number $\eta \leq 1$ such that $\mathsf{E}_\gamma(P\mathsf{K}\|Q\mathsf{K}) \leq \eta_\gamma(\mathsf{K})\mathsf{E}_\gamma(P\|Q)$ for any pair of (P, Q). More precisely, we have

$$\eta_\gamma(\mathsf{K}) \coloneqq \sup_{\substack{P,Q:\\ \mathsf{E}_\gamma(P\|Q)\neq 0}} \frac{\mathsf{E}_\gamma(P\mathsf{K}\|Q\mathsf{K})}{\mathsf{E}_\gamma(P\|Q)}. \tag{3}$$

Since E_γ-divergence reduces to the total variation distance for $\gamma = 1$, we denote $\eta_1(\mathsf{K})$ by $\eta_{\mathsf{TV}}(\mathsf{K})$. Interestingly, $\eta_\gamma(\mathsf{K}) \leq \eta_{\mathsf{TV}}(\mathsf{K})$ for all $\gamma \geq 0$ and kernels K [23]. It has recently been shown in [20] that $\eta_\gamma(\mathsf{K})$ with $\gamma \in [1, \infty)$ enjoys a simple characterization:

$$\eta_\gamma(\mathsf{K}) = \sup_{x, x' \in \mathcal{X}} \mathsf{E}_\gamma\big(\mathsf{K}(\cdot|x)\|\mathsf{K}(\cdot|x')\big). \tag{4}$$

We say a kernel is contractive if $\mathsf{K}(y|x) > 0$ for all $x, y \in \mathcal{X}$. This definition in particular implies that $\eta_{\mathsf{TV}}(\mathsf{K}) < 1$, and thus $\eta_\gamma(\mathsf{K}) < 1$ for all $\gamma > 0$.

C. Properties of Markov Kernels

Consider a discrete-time Markov chain $\{X_n\}_n$ with states in a finite alphabet $\mathcal{X}$ and a one-step transition probability matrix specified by K, if $X_0 \sim P$, then $X_n \sim P\mathsf{K}^n$, where K^n denotes the n-step transition probility. For instance, $\mathsf{K}^2(x|x') = \sum_{y \in \mathcal{X}} \mathsf{K}(x|y)\mathsf{K}(y|x')$. We say that a Markov chain is irreducible if for any $x, x' \in \mathcal{X}$, there exists an integer n_0 such that $\mathsf{K}^{n_0}(x|x') > 0$. Furthermore, let $\mathcal{T}(x) \coloneqq \{t \geq 1 : \mathsf{K}^t(x, x) > 0\}$ be the set of times when the chain can return to starting position x. A Markov chain is considered aperiodic if the period of all its states $x \in \mathcal{X}$, which is defined as the greatest common divisor of $\mathcal{T}(x)$, is equal to one. Recall that for any aperiodic and irreducible Markov chain, there exists a unique distribution Q^*, typically referred to as the *stationary distribution*, such that $Q^*\mathsf{K} = Q^*$.

D. Mixing Times

Given an irreducible and aperiodic Markov chain with the associated kernel K, it is widely known that $P\mathsf{K}^n$ converges to its stationary distribution Q^*. To capture the rate of such convergence, we measure the distance of $P\mathsf{K}^n$ from the stationarity distribution Q^* via E_γ-divergence:

$$\begin{aligned} d_\gamma^{\mathsf{K}}(n) &\coloneqq \sup_{P \in \mathcal{P}(\mathcal{X})} \mathsf{E}_\gamma\big(P\mathsf{K}^n\|Q^*\big) \\ &= \max_{x \in \mathcal{X}} \mathsf{E}_\gamma\big(\mathsf{K}^n(\cdot|x)\|Q^*\big), \end{aligned}$$

where the last equality follows from the joint covexity of $(P, Q) \mapsto \mathsf{E}_\gamma(P\|Q)$ (similar to any other f-divergences). For each $\varepsilon \geq 0$, the E_γ-mixing time of a Markov kernel K is defined as

$$t_\gamma^{\mathsf{K}}(\varepsilon) \coloneqq \min\{n \geq 1 : d_\gamma^{\mathsf{K}}(n) \leq \varepsilon\}.$$

III. Finite Mixing time for Irreducible and Aperiodic Markov Chains

In this section, we aim to prove that $t_\gamma^{\mathsf{K}}(\varepsilon) < \infty$ for all $\varepsilon \geq 0$ provided that $\gamma > 1$. This rather surprising feature of E_γ-mixing time is in sharp contrast with the known property

of the TV-mixing time (which corresponds to $t_1^{\mathsf{K}}(\varepsilon)$), which asserts that it is infinite for $\varepsilon = 0$.

Recall that $Q^*\mathsf{K}^n = Q^*$. Thus, we can write

$$\begin{aligned}\mathsf{E}_\gamma\big(P\mathsf{K}^n\|Q^*\big) &= \mathsf{E}_\gamma\big(P\mathsf{K}^n\|Q^*\mathsf{K}^n\big)\\ &\le \eta_\gamma(\mathsf{K}^n)\mathsf{E}_\gamma(P\|Q^*),\end{aligned}$$

where the inequality follows from the definition of η_γ. Since $\mathsf{E}_\gamma(P\|Q^*)$ is always smaller than 1, it follows that

$$d_\gamma^{\mathsf{K}}(n) \le \eta_\gamma(\mathsf{K}^n) \le \eta_\gamma(\mathsf{K})^n.$$

This in turn implies

$$t_\gamma^{\mathsf{K}}(\varepsilon) \le \frac{\log(\varepsilon)}{\log\big(\eta_\gamma(\mathsf{K})\big)},$$

recovering the typical behavior of the TV-mixing time in (1). Note, however, that the previous bound is vacuous for $\varepsilon = 0$. While the latter behavior is typical for $\gamma = 1$, the following theorem shows that $t_\gamma^{\mathsf{K}}(0)$ is finite for $\gamma > 1$.

Theorem 1. *Let $\mathsf{K} : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ be a Markov kernel over a finite alphabet $\mathcal{X}$. If K describes an irreducible and aperiodic Markov chain, then we have $t_\gamma^{\mathsf{K}}(0) < \infty$ for all $\gamma \in (1, \infty)$.*

Given this theorem, it is natural to aim to determine upper bounds for $t_\gamma^{\mathsf{K}}(0)$. In the following sections, we develop tools to address this problem for a broad family of *contractive* kernels. We note that if K is contractive, then it describes an irreducible and aperiodic Markov chain and, from the above theorem, $t_\gamma^{\mathsf{K}}(0) < \infty$.

IV. Mixing Time Through Linear SDPI

In this section, we further examine the connection between SDPI and mixing times to develop a framework for bounding $t_\gamma^{\mathsf{K}}(0)$ for a family of contractive kernels. In particular, we define the following parametric family of kernels parametrized by a parameter $\alpha \ge 0$:

$$\mathcal{F}_\alpha \coloneqq \Big\{\mathsf{K} : \mathcal{X} \to \mathcal{P}(\mathcal{X}) : \sup_{x,x'\in\mathcal{X},y\in\mathcal{X}} \log\frac{\mathsf{K}(y|x)}{\mathsf{K}(y|x')} \le \alpha\Big\}. \tag{5}$$

First, we show that for $\mathsf{K} \in \mathcal{F}_\alpha$, it is sufficient to consider the E_γ-mixing time for $1 < \gamma < e^\alpha$.

Lemma 1. *Let $\mathcal{X}$ be a finite set and $\mathsf{K} : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ be a Markov kernel in $\mathcal{F}_\alpha$. Then, for $\gamma \ge e^\alpha$, $\eta_\gamma(\mathsf{K}) = 0$ and $t_\gamma^{\mathsf{K}}(0) = 1$.*

Proof. First, note that since $\mathsf{K}(y|x) \le e^\alpha \mathsf{K}(y|x')$ for any $x, x', y \in \mathcal{X}$, we have $\mathsf{E}_{e^\alpha}(\mathsf{K}(\cdot|x)\|\mathsf{K}(\cdot|x')) = 0$ for any $x, x' \in \mathcal{X}$. Thus, in light of the characterization given in (4), we have $\eta_{e^\alpha}(\mathsf{K}) = 0$. It then follows from the monotonicity of the mapping $\gamma \to \mathsf{E}_\gamma(P\|Q)$ [21] that $\eta_\gamma(\mathsf{K}) = 0$ for $\gamma \ge e^\alpha$.

Now, let Q^* be the stationary distribution of K. We can assert that for all $\gamma \ge e^\alpha$,

$$\begin{aligned}d_\gamma^{\mathsf{K}}(1) &= \sup_{P\in\mathcal{P}(\mathcal{X})} \mathsf{E}_\gamma\big(P\mathsf{K}\|Q^*\big)\\ &= \sup_{P\in\mathcal{P}(\mathcal{X})} \mathsf{E}_\gamma\big(P\mathsf{K}\|Q^*\mathsf{K}\big)\\ &\le \eta_\gamma(\mathsf{K}) \sup_{P\in\mathcal{P}(X)} \mathsf{E}_\gamma\big(P\|Q^*\big)\\ &\le \eta_\gamma(\mathsf{K}).\end{aligned}$$

Since $\eta_\gamma(\mathsf{K}) = 0$ for $\gamma \ge e^\alpha$, we conclude that $t_\gamma^{\mathsf{K}}(0) = 1$. ∎

Next, we focus on characterizing the E_γ-mixing time for $\gamma \in (1, e^\alpha)$. The following theorem presents an upper bound for $t_\gamma^{\mathsf{K}}(0)$ that depends on Q^* the stationary distribution of K. Let $Q^*_{\min}$ denote $\min_{x\in\mathcal{X}} Q^*(x)$. Notice that $Q^*(x) > 0$ for all $x \in \mathcal{X}$, and thus $Q^*_{\min} > 0$.

Theorem 2. *Let $\mathcal{X}$ be a finite set and $\mathsf{K} \in \mathcal{F}_\alpha$ with the stationary distribution Q^*. If $(\gamma-1)Q^*_{\min} \ge 1$, then $t_\gamma^{\mathsf{K}}(0) = 1$. Otherwise, if $1 < \gamma < e^\alpha$, then*

$$t_\gamma^{\mathsf{K}}(0) \le \frac{\log[(\gamma-1)Q^*_{\min}]}{\log\zeta(e^\alpha)},$$

where $\zeta(\cdot)$ was defined in (2)*.*

The proof of Theorem 2 depends on two results, each of which might be of independent interest. The first one establishes an optimal bound on $\mathsf{E}_{\gamma'}(P\|Q)$ in terms of $\mathsf{E}_\gamma(P\|Q)$ for $\gamma' \le \gamma$. The second result states that if $\mathsf{E}_{\gamma'}(P\|Q)$ is sufficiently small, then $\mathsf{E}_\gamma(P\|Q) = 0$ for $\gamma' \le \gamma$.

Proposition 1. *Let $\mathcal{X}$ be an arbitrary set. If $1 \le \gamma' \le \gamma$, then, for all $P, Q \in \mathcal{P}(\mathcal{X})$,*

$$\mathsf{E}_{\gamma'}(P\|Q) \le 1 - \frac{\gamma'+1}{\gamma+1}\big(1 - \mathsf{E}_\gamma(P\|Q) \vee \mathsf{E}_\gamma(Q\|P)\big).$$

We remark that this bound is a generalization of [24, Proposition 4] that states

$$\mathsf{TV}(P, Q) \le 1 - \frac{1}{\gamma}\big(1 - \mathsf{E}_\gamma(P\|Q)\big).$$

Combined with the characterization of $\eta_\gamma(\mathsf{K})$ given in (4), Proposition 1 establishes the following result which is a key component in proving Theorem 2:

$$\eta_{\gamma'}(\mathsf{K}) \le 1 - \frac{\gamma'+1}{\gamma+1}\big(1 - \eta_\gamma(\mathsf{K})\big). \tag{6}$$

It is worth noting that this inequality is tight in general, meaning it cannot be improved for general kernels K. For instance, if K is a binary symmetric channel with crossover probability $\kappa \in (0, 1/2)$, then it can be verified that $\eta_\gamma(\mathsf{K}) = (1 - (\gamma+1)\kappa)_+$ and thus $\eta_{\gamma'}(\mathsf{K}) = 1 - \frac{\gamma'+1}{\gamma+1}\big(1 - \eta_\gamma(\mathsf{K})\big)$ for $1 \le \gamma' \le \gamma \le 1/\kappa - 1$.

The second component required for the proof of Theorem 2 is the following.

Proposition 2. *Let $\mathcal{X}$ be a finite set and $P, Q \in \mathcal{P}(\mathcal{X})$. If $1 \le \gamma' \le \gamma$ and $\mathsf{E}_{\gamma'}(P\|Q) \le (\gamma-\gamma')\min_{x\in\mathcal{X}} Q(x)$, then $\mathsf{E}_\gamma(P\|Q) = 0$.*

We can now provide a sketch of the proof of Theorem 2.

Proof of Theorem 2. First note that from (6), we have that

$$\eta_{\mathsf{TV}}(\mathsf{K}) \le 1 - \frac{2}{e^\alpha+1}\big(1 - \eta_{e^\alpha}(\mathsf{K})\big).$$

Since $\eta_{e^\alpha}(\mathsf{K}) = 0$ and $Q^*\mathsf{K}^n = Q^*$, it follows from the definition of the contraction coefficient that for all $P \in \mathcal{P}(\mathcal{X})$,

$$\mathsf{TV}(P\mathsf{K}^n, Q^*) \le \eta_{\mathsf{TV}}(\mathsf{K})^n \le \left(1 - \frac{2}{e^\alpha + 1}\right)^n.$$

Take

$$n = \frac{\log[(\gamma-1)Q^*_{\min}]}{\log\left[1 - 2/(e^\alpha+1)\right]}.$$

Then, we have that

$$\mathsf{TV}(P\mathsf{K}^n, Q^*) \le (\gamma - 1)\min_{x\in\mathcal{X}} Q^*(x).$$

By Proposition 2, we conclude that $\mathsf{E}_\gamma(P\mathsf{K}^n\|Q^*) = 0$. Since $P \in \mathcal{P}(\mathcal{X})$ is arbitrary, we have that $d_\gamma^{\mathsf{K}}(n) = 0$ and we have the result. ∎

We must note that the upper bound in Theorem 2 for the E_γ-mixing time of any contractive Markov kernel depends on its stationary distribution, which can potentially be unknown. This may restrict the versatility of this theorem. In the following section, we develop a framework, based on the *non-linear* SDPI, that enables us to establish bounds on the E_γ-mixing time of contractive kernels in $\mathcal{F}_\alpha$ without relying on their stationary distributions. Interestingly, the resulting bounds might even be tighter than Theorem 2.

V. Mixing Time Through Non-linear SDPI

In this section, we first mathematically formulate a new type of SDPI for general Markov kernels which is complementary to the one discussed in Section II-B. We then tightly characterize such SDPI for $\mathsf{K} \in \mathcal{F}_\alpha$.

Define $\mathsf{F}_\gamma : [0,1] \to [0,1]$ for any kernel K and any $\gamma \ge 1$ as follows:

$$\mathsf{F}_\gamma(\mathsf{K}, t) \coloneqq \sup\left\{\mathsf{E}_\gamma(P\mathsf{K}\|Q\mathsf{K}) : \mathsf{E}_\gamma(P\|Q) \le t\right\}. \tag{7}$$

We write $\mathsf{F}_\gamma(t)$ for $\mathsf{F}_\gamma(\mathsf{K}, t)$ when the Markov kernel K is clear from the context. We note that this function is closely related to the "Dobrushin curve" defined in [21] which essentially corresponds to $\mathsf{F}_1(\mathsf{K}, t)$. As such, we call $t \mapsto \mathsf{F}_\gamma(\mathsf{K}, t)$ as the generalized Dobrushin curve.

Notice that the mapping $t \mapsto \mathsf{F}_\gamma(\mathsf{K}, t)$ identifies a better upper bound for $\mathsf{E}_\gamma(P\mathsf{K}\|Q\mathsf{K})$ in terms of $\mathsf{E}_\gamma(P\|Q)$ than the one obtained by the contraction coefficient. In other words, we directly have $\mathsf{F}_\gamma(\mathsf{K}, t) \le \eta_\gamma(\mathsf{K})t$. However, this inequality is strict for most non-trivial kernels. For instance, it is possible to have $\mathsf{E}_\gamma(P\mathsf{K}\|Q\mathsf{K}) = 0$ for some $\mathsf{K} \in \mathcal{F}_\alpha$ and $\gamma > 1$, while $\mathsf{E}_\gamma(P\|Q) > 0$, as delineated by the next lemma.

Lemma 2. *Let $\gamma \in (1, e^\alpha]$, and P and Q be any two distributions such that $\mathsf{E}_\gamma(P\|Q) \le \frac{\gamma-1}{e^\alpha-1}$. Then, $\mathsf{E}_\gamma(P\mathsf{K}\|Q\mathsf{K}) = 0$ for any $\mathsf{K} \in \mathcal{F}_\alpha$. In particular, for all $t \le \frac{\gamma-1}{e^\alpha-1}$, we have*

$$\mathsf{F}_\gamma(t) = 0.$$

Fig. 1. Comparison of the non-linear SDPI from [7] and ours in Theorem 3 where P and Q are full-support random distributions on $[k]$ and $\mathsf{K} : [k] \to \mathcal{P}([k])$ is given by $\mathsf{K}(x|x) \propto e^\alpha$ and $\mathsf{K}(y|x) \propto 1$ for $y \ne x \in [k]$ and some $\alpha \ge 0$. It is straightforward to see that $\mathsf{K} \in \mathcal{F}_\alpha$. Here, we pick $\alpha = 1$ and $k = 5$.

Lemma 2 highlights the potential weakness of linear SDPI, suggesting $\mathsf{F}_\gamma(\mathsf{K}, t)$ as a better tool for establishing bounds on the mixing time. Balle et al. [7] recently established an example of such non-linear upper bounds:

$$\mathsf{E}_\gamma(P\mathsf{K}\|Q\mathsf{K}) \le \eta_{\gamma_t}(\mathsf{K})\mathsf{E}_\gamma(P\|Q), \tag{8}$$

where $\gamma_t = 1 + \frac{\gamma-1}{t}$. Since $\gamma_t > \gamma$ for $t < 1$, this upper bound strictly improves over its linear counterpart. While this result is indeed remarkable, its formulation may not immediately lend itself to bounding F_γ of the composition of Markov kernels, which is a crucial tool in the characterization of E_γ-mixing time. Therefore, we seek an alternative non-linear SDPI that can more effectively address the composition of Markov kernels. The following theorem presents our non-linear SDPI which can directly be employed to establish an upper bound on the E_γ-mixing time. This result also solves a conjecture put forth in [25], demonstrating the tightness of this non-linear SDPI.

Theorem 3. *Let $\mathsf{K} \in \mathcal{F}_\alpha$, and $\gamma \ge 1$ then*

$$\mathsf{E}_\gamma\left(P\mathsf{K}\|Q\mathsf{K}\right) \le \frac{1}{1+e^\alpha}\left(\left(e^\alpha - 1\right)\mathsf{E}_\gamma(P\|Q) + (1-\gamma)\right)_+. \tag{9}$$

We remark that the upper bound in this theorem is tight, see [25, Theorem 2] for an example of P, Q and K for which the inequality in (9) becomes equality. Figure 1 compares our non-linear SDPI with that of [7] given in (8) where P and Q are full-support random distributions on $[k]$ and $\mathsf{K} : [k] \to \mathcal{P}([k])$ is given by $\mathsf{K}(x|x) = \frac{e^\alpha}{e^\alpha+k-1}$ and $\mathsf{K}(y|x) = \frac{1}{e^\alpha+k-1}$ for $y \ne x \in [k]$ and some $\alpha \ge 0$. It is straightforward to see that $\mathsf{K} \in \mathcal{F}_\alpha$. While our bound is not consistently superior, it accurately captures the well-known behavior of E_γ-divergence, reaching zero after a certain threshold γ.

Next, we utilize our non-linear SDPI in Theorem 3 to characterize our bound on the E_γ-mixing time.

Theorem 4. *Let $\mathsf{K} \in \mathcal{F}_\alpha$ and $1 < \gamma < e^\alpha$. Then, we have*

$$t_\gamma^{\mathsf{K}}(0) \le \frac{\log\zeta(\gamma)}{\log\zeta(e^\alpha)},$$

where $\zeta(\cdot)$ was defined in (2)*.*

Proof. Using Theorem 3, we have that

$$\mathsf{F}_\gamma(t) \le \left(t\zeta(e^\alpha) - \frac{(\gamma-1)}{2}(1-\zeta(e^\alpha))\right)_+.$$

Recall that $\mathsf{F}_\gamma(\mathsf{K}_1 \circ \mathsf{K}_2, t) \le \mathsf{F}_\gamma(\mathsf{K}_1, \mathsf{F}_\gamma(\mathsf{K}_2, t))$ for any Markov kernels K_1 and K_2. Hence, the generalized Dobrushin curve

for K^n is given by the n-fold composition of F_γ, i.e.,

$$\begin{aligned}\mathsf{F}_\gamma(\mathsf{K}^n, t) &\leq \left(t\zeta(e^\alpha)^n - \frac{(\gamma-1)}{e^\alpha+1}\left(\frac{1-\zeta(e^\alpha)^n}{1-\zeta(e^\alpha)}\right)\right)_+ \\ &= \left(t\zeta(e^\alpha)^n - \frac{(\gamma-1)\left(1-\zeta(e^\alpha)^n\right)}{2}\right)_+ \\ &\leq \left(\zeta(e^\alpha)^n - \frac{(\gamma-1)\left(1-\zeta(e^\alpha)^n\right)}{2}\right)_+.\end{aligned}$$

The result is obtained by setting the RHS to 0. ∎

Theorem 4 improves upon Theorem 2 in two significant ways. First of all, the bound in Theorem 4 is independent of the stationary distribution of the Markov kernel under investigation. Second, for choices of $\gamma > \frac{1}{Q^*_{\min}} - 1$, Theorem 4 provides a strict improvement over its counterpart in Theorem 2.

It is worth noting that Theorem 4, together with Proposition 1, can be used to determine bounds on the mixing time under other divergences. For instance, it reduces to the standard bound on the TV-mixing time given in (1). To see this, first note that from Proposition 1 we have $t^{\mathsf{K}}_{\mathsf{TV}}(\zeta(\gamma)) \leq t^{\mathsf{K}}_\gamma(0)$ for any $\gamma > 1$. Applying Theorem 4, we therefore obtain $t^{\mathsf{K}}_{\mathsf{TV}}(\zeta(\gamma)) \leq \frac{\log\zeta(\gamma)}{\log\zeta(e^\alpha)}$. Thus, choosing $\gamma = \zeta^{-1}(\varepsilon)$ yields $t^{\mathsf{K}}_{\mathsf{TV}}(\varepsilon) \leq \frac{\log\varepsilon}{\log\zeta(e^\alpha)}$. This implies the bound in (1) after noticing $\eta_{\mathsf{TV}}(\mathsf{K}) \leq \zeta(e^\alpha)$ for all $\mathsf{K} \in \mathcal{F}_\alpha$ (see [26, Cor. 11] for a proof).

VI. By-product: Privacy Amplification by Composition

In this section, we apply the results proved in the previous section to the differential privacy settings. A well-known result in the privacy literature is that the post-processing of a privacy mechanism preserves its privacy guarantees. In many practical scenarios, it is desirable to precisely quantify how much a post-processing can in fact *amplify* privacy.

We focus on the family of locally differentially private (LDP) mechanisms [27, 28]. A Markov kernel $\mathsf{K} : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ is said to be (ε, δ)-LDP for $\varepsilon \geq 0$ and $\delta \in [0, 1]$ if

$$\sup_{x,x' \in \mathcal{X}} \mathsf{E}_{e^\varepsilon}\left(\mathsf{K}(\cdot|x) \| \mathsf{K}(\cdot|x')\right) = 0$$

If K is $(\varepsilon, 0)$-LDP, we say that it is ε-LDP. We categorize an LDP mechanism as an approximate LDP mechanism when $\delta > 0$, and as a pure LDP mechanism when $\delta = 0$. It is important to note that $\mathcal{F}_\varepsilon$ is in fact the set of all ε-LDP mechanisms.

The connection between SDPI and LDP has recently become (more) clear in [11] which showed: K is (ε, δ)-LDP *if and only if* $\eta_{e^\varepsilon}(\mathsf{K}) \leq \delta$. Thus, the privacy amplification problem naturally fits into the SDPI framework, because post-processing a privacy mechanism can be viewed as composing a Markov kernel with the privacy mechanism.

We consider the following two scenarios:

Scenario I: Suppose $\mathsf{K}_i : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ is ε_i-LDP for $i \in [n]$. We want to identify the privacy guarantees of their composition as a function of ε_i's. In particular, we seek to determine ε' such that $\mathsf{K}_1 \circ \cdots \circ \mathsf{K}_n$ is ε'-LDP.

Scenario II: Suppose K is an (ε, δ)-LDP mechanism. We are interested in identifying the smallest n such that K^n is a pure LDP.

The privacy amplification described in Scenario I can be converted into the following problem related to SDPI: Given $\gamma_i = e^{\varepsilon_i}$ and kernels K_i satisfying $\eta_{\gamma_i}(\mathsf{K}_i) = 0$ for $i \in [n]$, what is the γ such that $\eta_\gamma(\mathsf{K}_1 \circ \cdots \circ \mathsf{K}_n) = 0$? This problem was investigated in [29] for the special case of $n = 2$. Specifically, it was shown that the composition of an ε_1-LDP mechanism with an ε_2-LDP mechanism is $\tilde{\varepsilon}$-LDP where $\tilde{\varepsilon} = \frac{e^{\varepsilon_1+\varepsilon_2}+1}{e^{\varepsilon_1}+e^{\varepsilon_2}}$. In the following theorem, we extend this result by leveraging the non-linear SDPI proved in Theorem 3.

Theorem 5. *Let K_i be ε_i-LDP for $i \in [n]$. Then, their composition is ε'-LDP where*

$$\varepsilon' = \log\left(\frac{1 + \prod_{j=1}^n \zeta\left(e^{\varepsilon_j}\right)}{1 - \prod_{j=1}^n \zeta\left(e^{\varepsilon_j}\right)}\right).$$

We observe that the result in Theorem 5 recovers [29, Theorem 4.3] when $n = 2$.

Next, we consider Scenario II. The privacy amplification described in this scenario is equivalent to the following problem: Given a kernel K satisfying $\eta_{e^\varepsilon}(\mathsf{K}) \leq \delta$, what are the smallest n and γ such that $\eta_\gamma(\mathsf{K}^n) = 0$? In the following lemma, we partially answer this question.

Lemma 3. *Any (ε, δ)-LDP mechanism K defined over a finite alphabet can be converted to an ε'-LDP with $\varepsilon' < \varepsilon$ by composing K with itself n times with*

$$n = \max\left\{\frac{\log\left[(e^{\varepsilon'}-1)Q^*_{\min}/2\right]}{\log(\eta_{\mathsf{TV}}(\mathsf{K}))}, \frac{\log(Q^*_{\min}/2)}{\log(\eta_{\mathsf{TV}}(\mathsf{K}))}\right\},$$

where Q^* is the stationary distribution of K, and $Q^*_{\min} = \min_{x\in\mathcal{X}} Q^*(x)$.

It is worth noting that similar problems were explored in [7] and [30]. In [7], the authors demonstrated that composing an (ε, δ)-LDP mechanism K with a $\log(1 + \frac{e^\varepsilon - 1}{\delta})$-LDP mechanism results in a mechanism that is ε-LDP. Additionally, [30, Theorem 1] offered a method for converting approximate LDP mechanisms into pure ones. Specifically, given an (ε, δ)-LDP mechanism, they introduced an approach to obtain an 8ε-LDP mechanism. Therefore, in both cases, privacy is amplified in terms of δ, but not in terms of ε, which makes our result remarkable as it amplifies privacy in both ε and δ.

REFERENCES

[1] R. Montenegro and P. Tetali, "Mathematical aspects of mixing times in markov chains," *Foundations and Trends® in Theoretical Computer Science*, vol. 1, no. 3, pp. 237–354, 2006. [Online]. Available: http://dx.doi.org/10.1561/0400000003

[2] J. M. Altschuler and K. Talwar, "Resolving the mixing time of the langevin algorithm to its stationary distribution for log-concave sampling," in *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023,*, vol. 195, 2023, pp. 2509–2510.

[3] K. Temme, M. J. Kastoryano, M. B. Ruskai, M. M. Wolf, and F. Verstraete, "The χ^2-divergence and mixing times of quantum Markov processes," *Journal of Mathematical Physics*, vol. 51, no. 12, p. 122201, 2010.

[4] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*. American Mathematical Soc., 2017, vol. 107.

[5] I. Sason and S. Verdú, "f -divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory of Cryptography (TCC)*, 2006, pp. 265–284.

[7] B. Balle, G. Barthe, M. Gaboardi, and J. Geumlek, "Privacy amplification by mixing and diffusion mechanisms," in *NeurIPS*, 2019, pp. 13 277–13 287.

[8] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled Rényi differential privacy and analytical moments accountant," in *AISTAT*, vol. 89, 16–18 Apr 2018, pp. 1226–1235.

[9] S. Asoodeh and H. Zhang, "Contraction of locally differentially private mechanisms," 2022. [Online]. Available: https://arxiv.org/abs/2210.13386

[10] S. Asoodeh, J. Liao, F. P. Calmon, O. Kosut, and L. Sankar, "Three variants of differential privacy: Lossless conversion and applications," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 208–222, 2021.

[11] S. Asoodeh, M. Aliakbarpour, and F. P. Calmon, "Local differential privacy is equivalent to contraction of an E_γ-divergence," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 545–550.

[12] S. Asoodeh, W.-N. Chen, F. P. Calmon, and A. Özgür, "Differentially private federated learning: An information-theoretic perspective," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 344–349.

[13] J. Ye and R. Shokri, "Differentially private learning needs hidden state (or much faster convergence)," in *Advances in Neural Information Processing Systems*, 2022.

[14] R. Chourasia, J. Ye, and R. Shokri, "Differential privacy dynamics of langevin diffusion and noisy gradient descent," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[15] J. Altschuler and K. Talwar, "Privacy of noisy stochastic gradient descent: More iterations without more privacy loss," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 3788–3800.

[16] J. Cohen, J. Kemperman, and G. Zbăganu, *Comparisons of Stochastic Matrices, with Applications in Information Theory, Economics, and Population Sciences*. Birkhäuser, 1998.

[17] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, 12 1976.

[18] M. Raginsky, "Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3355–3389, June 2016.

[19] O. Faust and H. Fawzi, "Sum-of-squares proofs of logarithmic sobolev inequalities on finite markov chains," *IEEE Transactions on Information Theory*, vol. 70, no. 2, pp. 803–819, 2024.

[20] S. Asoodeh, M. Diaz, and F. P. Calmon, "Privacy analysis of online learning algorithms via contraction coefficients," in *International Symposium on Information Theory*, 2020, p. 1.

[21] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 35–55, 2016.

[22] I. Csiszar, "Generalized cutoff rates and renyi's information measures," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 26–34, Jan 1995.

[23] J. E. Cohen, Y. Iwasa, G. Rautu, M. Beth Ruskai, E. Seneta, and G. Zbaganu, "Relative entropy under mappings by stochastic matrices," *Linear Algebra and its Applications*, vol. 179, pp. 211 – 235, 1993.

[24] J. Liu, P. Cuff, and S. Verdú, "E_γ-resolvability," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 2629–2658, 2016.

[25] B. Zamanlooy and S. Asoodeh, "Strong data processing inequalities for locally differentially private mechanisms," in *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2023, pp. 1794–1799.

[26] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *Journal of Machine Learning Research*, vol. 17, no. 17, pp. 1–51, 2016.

[27] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. ACM symp. Principles of Database Systems (PODS)*. ACM, 2003, pp. 211–222.

[28] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, Jun. 2011.

[29] M. Naor and N. Vexler, "Can two walk together: Privacy enhancing methods and preventing tracking of users," in *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[30] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev, "Distributed differential privacy via shuffling," in *Advances in Cryptology–EUROCRYPT 2019: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19–23, 2019, Proceedings, Part I 38*. Springer, 2019, pp. 375–403.