Doeblin Curves

Dongmin Lee, William Lu, Anuran Makur, and Japneet Singh

Abstract

Recent research on Doeblin coefficients has shed light on their usefulness as a multi-way generalization of the Dobrushin contraction coefficient for TV distance, in a separate vein from their classic role in the theory of Markov chain ergodicity. However, strong conditions, such as being bounded away from 0, are typically necessary for Doeblin coefficients to establish the existence of information contraction. Building on the recently formulated concept of Dobrushin curves, we aim to propose a finer-grained Doeblin-based characterization of multi-way contraction behavior which yields non-vacuous contraction guarantees even for channels whose Doeblin coefficient is 0. To this end, we introduce the notion of a *Doeblin curve*—a nonlinear function which quantifies the contraction behavior of a Markov kernel on collections of input distributions at specific levels of divergence and power. Through the course of our analysis, we develop a new variational characterization of Doeblin coefficients, present several properties of Doeblin curves, define several versions of power-constrained Doeblin curves, and derive upper and lower bounds using our aforementioned variational characterization. We then utilize these results in diverse areas, including generalization bounds for noisy iterative optimization, error bounds for reliable computation with noisy circuits, and differential privacy guarantees for online iterative algorithms. In particular, we extend results in these areas to broader domains or group settings, leveraging Doeblin curves to reveal finer-grained contraction phenomena than Doeblin coefficients.

Index Terms

Doeblin coefficient, Markov kernel, information contraction.

I. Introduction

Analyzing the contraction properties of channels or Markov kernels is a fundamental problem in information theory stemming back to the celebrated *data processing inequality*, which states that the f-divergence between two distributions does not increase after they are passed through a Markov kernel [2], [3]. Data processing inequalities can be strengthened using contraction coefficients [4]–[9], which quantify the degree to which two distributions contract after being pushed forward through a Markov kernel (also see more recent work [10]–[18] and the references therein). In some cases, the quantification of this contraction behavior can be extended to an arbitrary collection of distributions represented by another Markov kernel, e.g., using Doeblin coefficients [19].

Historically, such contraction analyses were closely tied to investigating the convergence rate of a Markov chain to its steady-state distribution. Initial work in this vein by Doeblin [20], [21] established exponential convergence rates for all Markov matrices with strictly positive Doeblin coefficients, while more general weak ergodicity results for inhomogeneous Markov chains have been obtained using similar ideas, e.g., [22, Lemma 3]. Furthermore, [19] recently showed that Doeblin coefficients may be perceived as an n-way generalization of total variation (TV) distance, a view in which the submultiplicativity of Doeblin coefficients corresponds to the data processing inequality for n input distributions.

However, the utility of Doeblin coefficients remains limited in the regime of channels whose Doeblin coefficient is 0, for which information contraction properties cannot be easily established using existing Doeblin-based techniques. For example, prior work has demonstrated that Markov chains exhibit convergence to stationarity even under much weaker conditions such as drift and local minorization, cf. [23], and a strictly positive Doeblin coefficient may be considered a strong assumption. Consequently, contemporary results in Markov process analysis often shed the dependence on Doeblin coefficients altogether, instead being rooted in alternative methods such as spectral gap (or Poincaré) and logarithmic Sobolev inequalities [24].

The issue of contraction coefficients taking on trivial values exists in broader information-theoretic settings too [14], [25]. For example, Dobrushin's contraction coefficient for TV distance often takes the value 1 which does not demonstrate any meaningful information contraction [25]. To circumvent this issue, [25] develops nonlinear functions called *Dobrushin curves* that capture TV contraction even if the contraction coefficient is 1. Propelled by this line of inquiry, we seek to understand Doeblin-based information contraction properties of Markov kernels under general conditions, including the case of zero Doeblin coefficient. To this end, we introduce the notion of a *Doeblin curve* in this work. This nonlinear function quantifies the degree of contraction exhibited by its associated Markov kernel K, when applied to an arbitrary collection of input distributions represented by another kernel W. In contrast to conventional Doeblin coefficients, the nonlinearity of the Doeblin curve captures a more nuanced view of the contraction behavior of K.

The author ordering is alphabetical. This work was supported by the National Science Foundation (NSF) CAREER Award under Grant CCF-2337808. An earlier version of this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT) 2024 [1].

Dongmin Lee is with the Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA (e-mail: lee4818@purdue.edu).

William Lu is with the Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA (e-mail: lu909@purdue.edu).

Anuran Makur is with the Department of Computer Science and the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA (e-mail: amakur@purdue.edu).

Japneet Singh is with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA (e-mail: sing1041@purdue.edu).

A. Main Contributions and Outline

We briefly outline the structure of our paper and list our main contributions. Following a discussion of preliminaries, in Section II-A, we present a new variational characterization of Doeblin coefficients as an infimum over arbitrary partitions of the output space. Next, we introduce and formally define the Doeblin curve of a general Markov kernel and enumerate basic properties in Section II-B. We discuss power-constrained versions in Section II-C, derive upper and lower bounds using the aforementioned variational characterization in Section II-D, and present examples of kernels with closed-form Doeblin curves. Notably, in contrast to much of the literature on information contraction and Doeblin coefficients, this work is developed in the context of Markov kernels over general Polish spaces (as opposed to Markov matrices over finite sets).

Building on these theoretical results, we discuss several applications of Doeblin curves in Section III. In Section III-A, we employ Doeblin curves to derive generalization error bounds for noisy iterative algorithms operating on feasible sets with infinite diameter, a situation in which Doeblin coefficients fail to produce non-trivial contraction bounds. In Section III-B, we discuss lower bounds for reliable computation using circuits of noisy q-ary gates. Lastly, in Section III-C, we utilize our variational characterization of Doeblin coefficients to extend the definition of (ϵ, δ)-local differential privacy (LDP) to a group setting, and provide differential privacy guarantees for noisy iterative algorithms in terms of the Doeblin curve of the privacy mechanism.

We provide proofs of our main results in Section IV and proofs for the aforementioned applications in Section V. We defer further technical miscellany to Appendices A to C and relate Doeblin curves back to the classic setting of Markov chain ergodicity in Appendix D.

B. Related Literature

We summarize the prior literature on Doeblin coefficients, multi-way divergence metrics, and strong data processing inequalities (SDPIs). At the outset, Doeblin's seminal work [20], [21] introduced coupling as a technique for establishing uniform exponential convergence rates of Markov chains with respect to TV distance (also see [26]–[28]), and minorization as a technique to prove the weak ergodicity of inhomogeneous Markov chains (also see [22]). Notably, Doeblin coefficients were extremal minorization constants in such techniques. Subsequent results [27, Theorem 3.1], [29, Lemma 5], [30, Section IV-D] established the equivalence between Doeblin minorization and degradation by erasure channels, akin to the view of contraction coefficients for Kullback-Leibler (KL) divergence as domination by erasure channels under the less noisy preorder [13]. Doeblin coefficients also find relevance in various machine learning and applied probability problems, such as change detection algorithms [31], regret bounds in multi-armed bandit problems [32], Markov chain Monte Carlo methods [33], analysis of mixing times [34], and estimation of entropy rates [35]. Previous work on strong data processing inequalities [12, Remark 3.2], [36, Section I-D] has also shown how the Doeblin minorization condition yields upper bounds on contraction coefficients for f-divergences.

Extending the notion of f-divergence to compare three or more distributions simultaneously is another key avenue of research. For example, [37], [38] develop some general approaches for such extensions. Moreover, [19] established several properties of Doeblin coefficients (including geometric aspects, simultaneous and maximal coupling characterizations, and contraction properties over Bayesian networks), and noted that Doeblin coefficients may be perceived as a multi-way generalization of TV distance. (Interestingly, max-Doeblin coefficients [19], or maximal leakage [39], also share many of these properties [40].)

Much like Doeblin coefficients, the Dobrushin contraction coefficient for TV distance also plays a seminal role in the analysis of Markov kernels. Both the Doeblin and Dobrushin coefficients share key properties such as submultiplicativity [13], [19], and the latter enjoys utility in applications such as generalization error bounds [41], differential privacy guarantees [42], and analysis of reinforcement learning algorithms [43]. As noted earlier, [25] introduced Dobrushin curves to quantify information dissipation in channels with average input cost constraint even when their Dobrushin coefficient is trivially 1. Akin to this nonlinear perspective, other studies [13], [14] developed strong data processing inequalities which capture similar fine-grained contraction behavior. In this paper, we extend these ideas to the realm of Doeblin coefficients, demonstrating that analogous nonlinear enhancements can be devised to broaden the efficacy of Doeblin-based approaches for characterizing contraction behavior in general settings.

C. Preliminaries

We define relevant notation, review technical preliminaries, and provide several pertinent characterizations of Doeblin coefficients used in our paper. Let $\mathbb{N} \triangleq \{1, 2, \ldots\}$ denote the natural numbers starting from 1. Let $\mathbb{R}_+$ denote the non-negative real numbers. Let $[n] \triangleq \{1, \ldots, n\}$ denote integer intervals. Given a function $f : \mathcal{X} \to \mathbb{R}$ defined on a convex domain $\mathcal{X} \subseteq \mathbb{R}$, let $\hat{f} : \mathcal{X} \to \mathbb{R}$ denote its upper concave envelope (i.e., the pointwise infimum of all its concave upper bounds, or the "convex hull of f" [44, Chapter B, Proposition 2.5.1, Definition 2.5.3]). Given a Banach space $(\mathcal{X}, \|\cdot\|)$, denote the *diameter* and *Chebyshev radius* of a (multi-)subset $\mathcal{S} \subseteq \mathcal{X}$ as

$$|||\mathcal{S}|||_\infty \triangleq \sup_{x,y \in \mathcal{S}} \|x - y\|, \quad \mathrm{rad}(\mathcal{S}) \triangleq \inf_{a \in \mathcal{X}} \sup_{x \in \mathcal{S}} \|x - a\|. \tag{1}$$

All measures considered in this paper are finite. Let $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ be (measurable) Polish spaces equipped with Borel σ-algebras, and let $\mathscr{P}$ denote the set of probability measures on $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$. (Throughout this paper, all Polish spaces

considered are assumed to have cardinality at least 2.) A *Markov kernel* (or channel) K acting from $\mathcal{X}$ to $\mathcal{Y}$ is a mapping $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ such that $K(\cdot \mid x)$ is a probability measure for each $x \in \mathcal{X}$ and $K(A \mid \cdot)$ is a measurable function for each $A \in \mathcal{F}_{\mathcal{Y}}$. For any scalar $\alpha \in \mathbb{R}$, kernel K, and measure π, we write $K \geq \alpha\pi$ to denote that

$$\forall x \in \mathcal{X},\, \forall A \in \mathcal{F}_{\mathcal{Y}},\;\; K(A \mid x) \geq \alpha \cdot \pi(A)\,.$$

Given probability distributions P and Q, let $P \otimes Q$ denote their product distribution and let $P * Q$ denote their convolution. Given σ-algebras $\mathcal{F}_{\mathcal{X}}$ and $\mathcal{F}_{\mathcal{Y}}$, let $\mathcal{F}_{\mathcal{X}} \otimes \mathcal{F}_{\mathcal{Y}}$ denote their product σ-algebra. Given measures $P_i : \mathcal{F}_{\mathcal{Y}} \to \mathbb{R}_+$ for $i \in [n]$, let $[P_1, \ldots, P_n] : [n] \times \mathcal{F}_{\mathcal{Y}} \to \mathbb{R}_+$ denote the kernel formed by collecting these measures, i.e.,

$$\forall i \in [n],\, \forall A \in \mathcal{F}_{\mathcal{Y}},\;\; [P_1, \ldots, P_n]\,(A \mid i) \triangleq P_i(A)\,.$$

Let δ_x denote the Dirac measure (i.e., point mass) concentrated at x. Let Ξ denote the identity kernel, i.e.,

$$\forall x \in \mathcal{X},\;\; \Xi(\cdot \mid x) \triangleq \delta_x(\cdot)\,.$$

Given measures $\pi, \mu : \mathcal{F}_{\mathcal{Y}} \to \mathbb{R}_+$, we write $\pi \ll \mu$ to indicate that π is dominated by (or *absolutely continuous with respect to*) μ, i.e., $\mu(A) = 0 \Rightarrow \pi(A) = 0$ for all $A \in \mathcal{F}_{\mathcal{Y}}$. When $\pi \ll \mu$ and μ is σ-finite, we let $\frac{d\pi}{d\mu} : \mathcal{Y} \to \mathbb{R}_+$ denote the Radon-Nikodym derivative of π with respect to μ. We say that a Markov kernel $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ is *absolutely continuous* if there exists a σ-finite measure $\mu : \mathcal{F}_{\mathcal{Y}} \to \mathbb{R}_+$ such that $K(\cdot \mid x) \ll \mu$ for all $x \in \mathcal{X}$.[1] For example, note that if $\mathcal{X} = \{x_1, x_2, \ldots\}$ is countably infinite, one such dominating finite measure μ is

$$\mu(\cdot) = \sum_{i=1}^{\infty} \frac{K(\cdot \mid x_i)}{2^i}\,. \tag{2}$$

By Doob's variant of the Radon-Nikodym theorem [45, Chapter V, Theorem 4.44], absolute continuity of K implies the existence of a kernel function of Radon-Nikodym derivatives $\frac{dK}{d\mu}(y \mid x) : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}_+$, such that

$$\forall x \in \mathcal{X},\, \forall A \in \mathcal{F}_{\mathcal{Y}},\;\; K(A \mid x) = \int_A \frac{dK}{d\mu}(\cdot \mid x)\, d\mu$$

and $\frac{dK}{d\mu}$ is jointly measurable with respect to the product σ-algebra $\mathcal{F}_{\mathcal{Y}} \otimes \mathcal{F}_{\mathcal{X}}$. We impose this mild regularity condition on K as a precondition for several of our results; it is standard in measure-theoretic treatments of information theory (cf. [46, Theorem 2.12]).

Given a collection of measures $\{\pi_i\}_{i \in \mathcal{I}}$ on a Polish space $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$, where $\mathcal{I}$ is an arbitrary index set, define their *greatest common component* $\bigwedge_{i \in \mathcal{I}} \pi_i$ [47, Chapter 3, Section 7.1] as the unique measure on $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ given by

$$\forall A \in \mathcal{F}_{\mathcal{Y}},\;\; \bigwedge_{i \in \mathcal{I}} \pi_i(A) \triangleq \sup\{\nu(A) : \forall i \in \mathcal{I},\, \pi_i \geq \nu\}\,, \tag{3}$$

where the supremum is taken over all measures $\nu : \mathcal{F}_{\mathcal{Y}} \to \mathbb{R}_+$ such that $\pi_i \geq \nu$ for each $i \in \mathcal{I}$. We remark that $\bigwedge_{i \in \mathcal{I}} \pi_i(A) \neq \inf_{i \in \mathcal{I}} \pi_i(A)$ in general, because the pointwise infimum of measures is not a measure in general (as it may fail to satisfy countable additivity). We refer readers to [47, Chapter 3, Theorem 7.1] for a proof that the pointwise supremum in (3) defines a valid measure.

When we want to emphasize the discrete, finite-dimensional nature of a setting, we denote vectors and matrices with lowercase and uppercase bold letters, respectively. Let $\mathscr{P}_{d-1} \subset \mathbb{R}^d$ be the $(d-1)$-dimensional probability simplex of *row* pmf vectors, and let $\mathbb{R}^{d \times d}_{\mathsf{sto}}$ be the set of all $d \times d$ row stochastic matrices. Let $\mathbf{e}_i \in \mathscr{P}_{d-1}$ be the ith standard basis *row* vector. Given a matrix $\mathbf{A}$, let $[\mathbf{A}]_{i,j}$ denote its entry at row i and column j, and let $[\mathbf{A}]_{\langle i \rangle} = \mathbf{e}_i\mathbf{A}$ denote its ith row, represented as a *row* vector. If $\mathbf{A}$ is square, let $\lambda_i(\mathbf{A})$ denote its ith eigenvalue (counting algebraic multiplicity and ordered by descending magnitude). Let $\|\cdot\|_p$ denote the ℓ^p-norm in Euclidean space.

Given a Markov kernel K, define its *Doeblin coefficient* $\tau(K) \in [0,1]$ and *complementary Doeblin coefficient* $\rho(K) \in [0,1]$ as

$$\begin{aligned} \tau(K) &\triangleq \sup\{\alpha \in \mathbb{R} : \exists \pi \in \mathscr{P},\, K \geq \alpha\pi\}\,, \\ \rho(K) &\triangleq 1 - \tau(K)\,. \end{aligned} \tag{4}$$

The Doeblin coefficient of K may be characterized in terms of its greatest common component as

$$\tau(K) = \bigwedge_{x \in \mathcal{X}} K(\mathcal{Y} \mid x)\,, \tag{5}$$

[1] Note that such a common dominating measure μ always exists for any K, but we additionally require σ-finiteness of μ so that [45, Chapter V, Theorem 4.44] can be invoked.

because the scalar $\alpha = \bigwedge_{x\in\mathcal{X}} K(\mathcal{Y} \mid x)$ and the measure $\pi(\cdot) = (1/\alpha) \bigwedge_{x\in\mathcal{X}} K(\cdot \mid x)$ achieve the supremum in (4) by [22, Theorem 1].[2] For finite spaces (i.e., row stochastic matrices $\mathbf{K} \in \mathbb{R}_{\text{sto}}^{m\times n}$), $\tau(\mathbf{K})$ reduces to *Doeblin's coefficient of ergodicity* [48], [19, Definition 1], [22, Eq. 11], defined as

$$\tau(\mathbf{K}) = \sum_{j=1}^{n} \min_{i\in[m]} [\mathbf{K}]_{i,j} \,. \tag{6}$$

II. Main Results on Doeblin Curves

We begin by recalling the *maximal coupling characterization* of Doeblin coefficients [19, Theorem 2] defined as follows. Consider a collection of probability measures $\{P_i\}_{i\in\mathcal{I}}$ where $\mathcal{I}$ is a Polish index set and each probability measure P_i is on a respective Polish space $(\mathcal{Y}_i, \mathcal{F}_{\mathcal{Y}_i})$. A *coupling* of probability measures $\{P_i\}_{i\in\mathcal{I}}$ is a collection $\{Y_i\}_{i\in\mathcal{I}}$ of random variables all defined on the common measure space $\otimes_{i\in\mathcal{I}}(\mathcal{Y}_i, \mathcal{F}_{\mathcal{Y}_i})$, equipped with a (joint) probability measure $\mathbb{P}$ such that for all $i \in \mathcal{I}$, the marginal probability law of Y_i on $(\mathcal{Y}_i, \mathcal{F}_{\mathcal{Y}_i})$ is equal to P_i. (Since each $(\mathcal{Y}_i, \mathcal{F}_{\mathcal{Y}_i})$ is Polish, results such as the Kolmogorov extension theorem guarantee the existence of measures $\mathbb{P}$ on the infinite collection of random variables $\{Y_i\}_{i\in\mathcal{I}}$ which are consistent with finite dimensional distributions [49].) With this groundwork established, the following proposition states the desired maximal coupling characterization for Polish spaces.

Proposition 1 (Maximal Coupling Characterization of Doeblin Coefficients)**.** *Let $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0, 1]$ be a Markov kernel defined over Polish spaces $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$. Then,*

$$\tau(K) = \sup_{\mathbb{P}: Y_x \sim K(\cdot|x)} \mathbb{P}(\forall x, x' \in \mathcal{X},\, Y_x = Y_{x'}) \,, \tag{7}$$

where the supremum is taken over all couplings $\{Y_x\}_{x\in\mathcal{X}}$ of the probability measures $\{K(\cdot \mid x)\}_{x\in\mathcal{X}}$.

Proposition 1 is proved in Section IV-A for completeness. We remark that for any Markov kernel K, the supremum in (7) is achieved by the construction in (34), which we refer to as the "maximal coupling" for the remainder of this paper. Next, we present our main results, beginning with a new variational characterization of Doeblin coefficients.

A. Variational Characterization of Doeblin Coefficients

Our first main result is a new variational characterization of Doeblin coefficients, expressed as an infimum over arbitrary partitions of the underlying space. In this sense, our result is comparable to the Gel'fand-Yaglom-Peres variational characterization of KL divergence [50] or the extensions to f-divergences in [51], [46, Theorem 7.6].

Theorem 1 (Variational Characterization of Doeblin Coefficient)**.** *Let $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0, 1]$ be an absolutely continuous Markov kernel defined over Polish spaces $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$. Then,*

$$\tau(K) = \inf_{n\in\mathbb{N}} \inf_{x_1,\ldots,x_n\in\mathcal{X}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} K(A_i \mid x_i) \,,$$

where the innermost infimum is taken with respect to all measurable n-partitions $A_1, \ldots, A_n \in \mathcal{F}_{\mathcal{Y}}$, i.e., $A_i \cap A_j = \varnothing$ for all $i \neq j$ and $\bigcup_{i=1}^{n} A_i = \mathcal{Y}$.

Theorem 1 is proved in Section IV-A. The proof uses the characterization of Doeblin coefficients in (5) and the Radon-Nikodym theorem to express $\tau(K)$ in terms of the density of the greatest common component of K. To relate this quantity to the individual densities $\frac{dK}{d\mu}(\cdot \mid x)$ for $x \in \mathcal{X}$, we utilize the notion of *lattice infimum* as a measurable analogue of pointwise infimum, so that the lattice infimum of uncountably many measurable functions $\frac{dK}{d\mu}(\cdot \mid x)$ (indexed by $x \in \mathcal{X}$) remains measurable. Approximating this uncountable lattice infimum with a countable sequence gives rise to the infima over $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathcal{X}$ in Theorem 1, and we complete the proof by leveraging the following proposition to convert the integral of the pointwise minimum of finitely many densities into an infimum over partitions.

Proposition 2 (Integral Characterization of Infimum)**.** *Let $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0, 1]$ be an absolutely continuous Markov kernel defined over Polish spaces $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$. For any fixed $x_1, \ldots, x_n \in \mathcal{X}$ and any fixed constants $\gamma_1, \ldots, \gamma_n \geq 0$, we have*

$$\inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} \gamma_i K(A_i \mid x_i) = \int_{\mathcal{Y}} \left(\min_{i\in[n]} \gamma_i \frac{dK}{d\mu}(\cdot \mid x_i) \right) d\mu \,.$$

Proposition 2 is proved in Section IV-A and is also used in the proofs of subsequent results. We remark that Proposition 2 may be interpreted as a special case of Theorem 1 where $\mathcal{X}$ is finite. To see this, observe that whereas the infima over $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathcal{X}$ in Theorem 1 define the finite subset of $\mathcal{X}$ included in the sum (and thus the infinite subset of $\mathcal{X}$ omitted

[2] If $\alpha = 0$, we may take π to be any probability measure on $\mathcal{Y}$.

from the sum), they are unnecessary when $\mathcal{X}$ is finite, as any $x_i \in \mathcal{X}$ may be "omitted" from the sum by simply taking the respective A_i to be $\emptyset$. We note that by imposing stronger regularity conditions on K, we may prove Theorem 1 without the use of lattice infima; we refer interested readers to Appendix B for an analytically simpler argument under equicontinuity assumptions on $\frac{dK}{d\mu}$.

Throughout the rest of the paper, we find it helpful to define a notion of *constrained Doeblin coefficient*[3] for a Markov kernel $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to [0,1]$ and a measurable set $\mathcal{S} \in \mathcal{F}_\mathcal{X}$ as

$$\tau_\mathcal{S}(K) \triangleq \sup\big\{\alpha \in \mathbb{R} : \exists \pi \in \mathscr{P},\, \forall x \in \mathcal{S},\, K(\cdot \mid x) \geq \alpha\pi\big\}\,,$$
$$\rho_\mathcal{S}(K) \triangleq 1 - \tau_\mathcal{S}(K)\,.$$

We remark that $\rho_\mathcal{S}(K)$ is monotonically non-decreasing in $\mathcal{S}$, i.e., for any two sets $\mathcal{R}, \mathcal{S} \in \mathcal{F}_\mathcal{X}$ such that $\mathcal{R} \subseteq \mathcal{S}$, we have $\rho_\mathcal{R}(K) \leq \rho_\mathcal{S}(K)$.

B. Doeblin Curves

As our centerpiece contribution, we introduce the notion of a Doeblin curve to capture contraction properties of Markov kernels whose Doeblin coefficient is 0. For any Polish space $(\mathcal{U}, \mathcal{F}_\mathcal{U})$, define the composition of two kernels $W : \mathcal{U} \times \mathcal{F}_\mathcal{X} \to \mathbb{R}_+$ and $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to \mathbb{R}_+$ as the kernel $WK : \mathcal{U} \times \mathcal{F}_\mathcal{Y} \to \mathbb{R}_+$ given by

$$\forall u \in \mathcal{U},\, \forall A \in \mathcal{F}_\mathcal{Y},\;\; WK(A \mid u) \triangleq \int_\mathcal{X} K(A \mid x)\, W(dx \mid u)\,. \tag{8}$$

Now, we define the *Doeblin curve* of a Markov kernel $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to [0,1]$ as the function $\mathrm{F}_K(\cdot\,; \mathcal{G}) : [0,1] \to [0,1]$ given by

$$\mathrm{F}_K(t; \mathcal{G}) \triangleq \sup\left\{\rho(WK) : \rho(W) \leq t,\, W \in \mathcal{G}\right\}, \tag{9}$$

where the supremum is taken over all Markov kernels $W : \mathcal{U} \times \mathcal{F}_\mathcal{X} \to [0,1]$ from all Polish spaces $(\mathcal{U}, \mathcal{F}_\mathcal{U})$, such that W belongs to some non-empty constraint set $\mathcal{G}$. We define the joint range of the input and output complementary Doeblin coefficients as $\mathfrak{F}(K; \mathcal{G}) \triangleq \mathrm{cl}\left\{(\rho(W), \rho(WK)) : W \in \mathcal{G}\right\} \subseteq [0,1]^2$, where cl denotes closure. Specifically, the joint range $\mathfrak{F}(K; \mathcal{G})$ is contained within the set $\{(t,y) \in [0,1]^2 : y \leq \mathrm{F}_K(t; \mathcal{G})\}$. For any kernel K, the Doeblin curve $\mathrm{F}_K(t; \mathcal{G})$ is monotonically non-decreasing in t for any fixed constraint set $\mathcal{G}$. Also, the Doeblin curve $\mathrm{F}_K(t; \mathcal{G})$ and joint range $\mathfrak{F}(K; \mathcal{G})$ are monotonically non-decreasing in $\mathcal{G}$, i.e., for any two sets of kernels $\mathcal{G}, \mathcal{H}$ such that $\mathcal{G} \subseteq \mathcal{H}$, we have $\mathrm{F}_K(\cdot\,; \mathcal{G}) \leq \mathrm{F}_K(\cdot\,; \mathcal{H})$ and $\mathfrak{F}(K; \mathcal{G}) \subseteq \mathfrak{F}(K; \mathcal{H})$.

By submultiplicativity of complementary Doeblin coefficients [22, Section 4], [19, Theorem 1], we have $\rho(WK) \leq \rho(W)\rho(K)$. Hence, the Doeblin curve and joint range satisfy the outer bounds

$$\mathrm{F}_K(t; \mathcal{G}) \leq \rho(K)\, t\,,\;\; \mathfrak{F}(K; \mathcal{G}) \subseteq \left\{(t,y) \in [0,1]^2 : y \leq \rho(K)\, t\right\}, \tag{10}$$

and since $\rho(K) \leq 1$, we have

$$\mathrm{F}_K(t; \mathcal{G}) \leq t\,,\;\; \mathfrak{F}(K; \mathcal{G}) \subseteq \left\{(t,y) \in [0,1]^2 : y \leq t\right\}. \tag{11}$$

Figure 1 presents the numerically simulated joint range for a discrete memoryless channel $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{5\times 5}$, comprising 1 million instances of $\mathbf{W}$ sampled uniformly from the set of 5×5 row stochastic matrices. The blue dashed line depicts the Doeblin curve of $\mathbf{K}$ obtained analytically from Proposition 3, Part 2. The blue dots represent numerically sampled points $(\rho(\mathbf{W}), \rho(\mathbf{WK})) \in \mathfrak{F}(\mathbf{K}; \mathbb{R}_{\mathsf{sto}}^{5\times 5})$ in the joint range of $\mathbf{K}$. The figure shows that when $\mathbf{W}$ is uniformly sampled, such as in settings where channel inputs do not inherently tend towards degenerate regions of the probability simplex, the corresponding points in the joint range indicate noticeably more information contraction than the worst-case bound given by the Doeblin curve would suggest.

We enumerate several additional properties of Doeblin curves in the following proposition. For brevity, we call a Markov kernel $W : \mathcal{U} \times \mathcal{F}_\mathcal{X} \to [0,1]$ a *constant kernel* if there exists a fixed probability measure $\pi : \mathcal{F}_\mathcal{X} \to [0,1]$ such that $W(\cdot \mid u) = \pi(\cdot)$ for all $u \in \mathcal{U}$.

Proposition 3 (Properties of Doeblin Curves)**.** *The Doeblin curve defined in (9) satisfies the following properties:*

1) (Data processing property) *Given Markov kernels K_1 and K_2 and constraint sets $\mathcal{G}_1$ and $\mathcal{G}_2$, the Doeblin curve of the composite channel represented by the diagram $\mathcal{X}_1 \xrightarrow{K_1} \mathcal{Y}_1 = \mathcal{X}_2 \xrightarrow{K_2} \mathcal{Y}_2$, with constraint set $\mathcal{G}$ consisting of all kernels W such that $W \in \mathcal{G}_1$ and $WK_1 \in \mathcal{G}_2$, satisfies $\mathrm{F}_{K_1K_2}(t; \mathcal{G}) \leq \mathrm{F}_{K_2}(\mathrm{F}_{K_1}(t; \mathcal{G}_1); \mathcal{G}_2)$.*
2) (Sharpness) *For any Markov kernel $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to [0,1]$ and any convex constraint set $\mathcal{G}$ containing the identity kernel $\Xi : \mathcal{X} \times \mathcal{F}_\mathcal{X} \to [0,1]$ from $\mathcal{X}$ to $\mathcal{X}$, i.e., $\Xi(\cdot \mid u) = \delta_u(\cdot)$ for all $u \in \mathcal{X}$, and a constant kernel $\bar{W} : \mathcal{X} \times \mathcal{F}_\mathcal{X} \to [0,1]$, i.e., $\bar{W}(\cdot \mid u) = \pi(\cdot)$ for all $u \in \mathcal{X}$ for some fixed probability measure $\pi : \mathcal{F}_\mathcal{X} \to [0,1]$, the Doeblin curve $\mathrm{F}_K(t; \mathcal{G})$ achieves the bound in (10) with equality, i.e., $\mathrm{F}_K(t; \mathcal{G}) = \rho(K)\, t$.*

[3]The constrained Doeblin coefficient $\tau_\mathcal{S}(K)$ is essentially the (unconstrained) Doeblin coefficient $\tau(K')$ of the kernel K' obtained by restricting the input space of K to $\mathcal{S}$. Hence, Theorem 1 also holds for constrained Doeblin coefficients by taking the middle infimum over $x_1, \ldots, x_n \in \mathcal{S}$ instead of $\mathcal{X}$.

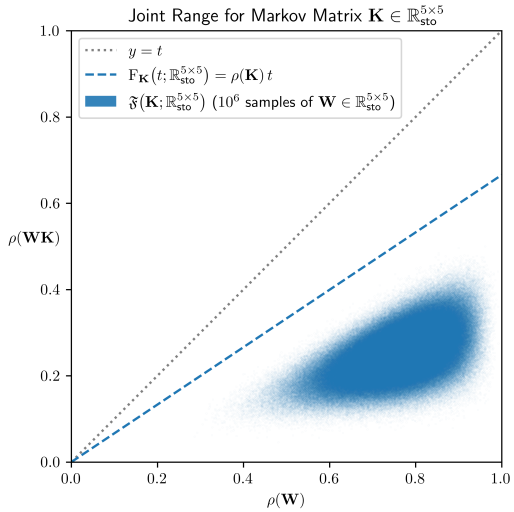

Fig. 1. Randomly sampled joint range $\mathfrak{F}(\mathbf{K};\mathbb{R}^{5\times5}_{\mathsf{sto}})$ for a 5×5 Markov matrix $\mathbf{K}$, shaded in blue.

3) (Super-homogeneity) *For any Markov kernel $K:\mathcal{X}\times\mathcal{F}_{\mathcal{Y}}\to[0,1]$ and any convex constraint set $\mathcal{G}$ containing a constant kernel $\bar{W}:\mathcal{U}\times\mathcal{F}_{\mathcal{X}}\to[0,1]$, the Doeblin curve $\mathrm{F}_K(t;\mathcal{G})$ satisfies $\mathrm{F}_K(\lambda t;\mathcal{G})\geq\lambda\mathrm{F}_K(t;\mathcal{G})$ for all $\lambda\in[0,1]$ and $t\in[0,1]$, or equivalently, the function $t\mapsto\mathrm{F}_K(t;\mathcal{G})/t$ is non-increasing.*
4) (Lipschitz continuity) *For any Markov kernel $K:\mathcal{X}\times\mathcal{F}_{\mathcal{Y}}\to[0,1]$ and any convex constraint set $\mathcal{G}$ containing a constant kernel $\bar{W}:\mathcal{U}\times\mathcal{F}_{\mathcal{X}}\to[0,1]$, the Doeblin curve $\mathrm{F}_K(t;\mathcal{G})$ is 1-Lipschitz continuous in t.*

Proposition 3 is proved in Section IV-B, primarily by using the fact that convex combinations of any Markov kernel with a constant kernel linearly interpolate ρ. We remark that Doeblin curves generalize the notion of Dobrushin curves in [25], just as Doeblin coefficients generalize contraction coefficients for TV distance. When $\mathcal{G}$ only contains Markov kernels $W:\mathcal{U}\times\mathcal{F}_{\mathcal{X}}\to[0,1]$ where $|\mathcal{U}|=2$, the Doeblin curve $\mathrm{F}_K(\cdot\,;\mathcal{G})$ reduces to the Dobrushin curve of K. We emphasize that the super-homogeneity property (Proposition 3, Part 3) does *not* necessarily imply that F_K is concave, and the concavity of F_K is also unknown in the Dobrushin case [25, Remark 1].

Although the joint range $\mathfrak{F}(K;\mathcal{G})$ is contained within the area under the Doeblin curve $\mathrm{F}_K(\cdot\,;\mathcal{G})$ due to the definition of F_K as a supremum, it is *not* necessarily the entire area under F_K, as shown by the following counterexample.

Proposition 4 (Joint Range of $2\times n$ Discrete Channels)**.** *The joint range of any discrete memoryless channel $\mathbf{K}\in\mathbb{R}^{2\times n}_{\mathsf{sto}}$, considering inputs $\mathbf{W}\in\mathbb{R}^{m\times2}_{\mathsf{sto}}$ with any number of rows $m\geq2$, is the line $\mathfrak{F}\big(\mathbf{K};\bigcup_{m\geq2}\mathbb{R}^{m\times2}_{\mathsf{sto}}\big)=\{(t,\rho(\mathbf{K})\,t):t\in[0,1]\}$.*

Proposition 4 is proved in Section IV-B.

C. Power-Constrained Doeblin Curves

Inspired by earlier developments of nonlinear information contraction [25], we note that the amount by which input distributions contract after passing through a channel depends on the power level of the input distributions. Hence, power constraints on the input kernel W admit natural and useful classes of constraint sets $\mathcal{G}$ which we consider in our subsequent analysis. To commence this discussion, we define two notions of power level for W taking values in a normed output space.

Definition 1 (Power Level)**.** *Let $(\mathcal{X},\|\cdot\|)$ be a separable Banach space equipped with the Borel σ-algebra $\mathcal{F}_{\mathcal{X}}$ induced by the norm topology. Given some convex and strictly increasing power function $M:\mathbb{R}_+\to\mathbb{R}_+$ with $M(0)=0$, we define the following notions of power level for a Markov kernel $W:\mathcal{U}\times\mathcal{F}_{\mathcal{X}}\to[0,1]$.*

- Uniform average power*:*
$$|||W|||_{\mathsf{UA}}\triangleq\sup_{u\in\mathcal{U}}\int_{\mathcal{X}}M(\|x\|)\,W(dx\mid u)\,. \tag{12}$$
- Average extremal power*: Let $\mathbb{P}$ be the maximal coupling of random variables $\{X_u\}_{u\in\mathcal{U}}$ with $X_u\sim W(\cdot\mid u)$ for each $u\in\mathcal{U}$, as defined in (34). Then,*
$$|||W|||_{\mathsf{AE}}\triangleq\mathbb{E}\left[\sup_{u\in\mathcal{U}}M(\|X_u\|)\right], \tag{13}$$
where the expectation is taken with respect to the maximal coupling $\{X_u\}_{u\in\mathcal{U}}\sim\mathbb{P}$, and we only define this notion of power for kernels W such that $\sup_{u\in\mathcal{U}}M(\|X_u\|)$ is measurable.

Note that the uniform average power is an extremal special case of the *average power* $|||W|||_{\mathsf{A}} \triangleq \mathbb{E}\big[\int_{\mathcal{X}} M(\|x\|)\, W(dx \mid U)\big]$ where the expectation is taken over some marginal distribution of $U \in \mathcal{U}$. The three power levels satisfy the relations $|||W|||_{\mathsf{A}} \le |||W|||_{\mathsf{UA}} \le |||W|||_{\mathsf{AE}}$. Since the marginal distribution of U is usually unknown, we hereafter consider only the uniform average (12) and average extremal (13) formalizations in our analysis. Note that the canonical notion of power (cf. [46, Section 20.1]) corresponds to taking $|||W|||_{\mathsf{UA}}$ with $M(z) = z^2$. We provide examples of non-trivial kernels satisfying the average extremal power constraint in Appendix C.

We write $\mathrm{F}_K^{\mathsf{UA}}(t;p)$ and $\mathrm{F}_K^{\mathsf{AE}}(t;p)$ to denote the Doeblin curves where the set $\mathcal{G}$ in (9) consists of all kernels satisfying the respective power constraints $|||W|||_{\mathsf{UA}} \le p$ and $|||W|||_{\mathsf{AE}} \le p$ for some $p \in [0,\infty)$. Since Doeblin curves are monotonically non-decreasing in the constraint set, we have $\mathrm{F}_K^{\mathsf{UA}}(t;p) \ge \mathrm{F}_K^{\mathsf{AE}}(t;p)$. Furthermore, we note that each of the power-constrained Doeblin curves satisfies the following homogeneity property (akin to Dobrushin curves).

Proposition 5 (Homogeneity of Power-Constrained Doeblin Curves)**.** *For any Markov kernel $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$, the power-constrained Doeblin curves $\mathrm{F}_K^{\mathsf{UA}}$ and $\mathrm{F}_K^{\mathsf{AE}}$ satisfy $\mathrm{F}_K^{\mathsf{UA}}(\lambda t;\lambda p) = \lambda \mathrm{F}_K^{\mathsf{UA}}(t;p)$ and $\mathrm{F}_K^{\mathsf{AE}}(\lambda t;\lambda p) = \lambda \mathrm{F}_K^{\mathsf{AE}}(t;p)$ for all $\lambda \in \mathbb{R}_+$ such that $\lambda t \le 1$.*

Proposition 5 is proved in Section IV-B using similar arguments based on convex combinations with constant kernels as those used in the proof of Proposition 3. In addition, the Doeblin curves of certain classes of kernels such as additive noise channels are invariant under scaling (when the power constraint is scaled accordingly) as the following proposition shows.

Proposition 6 (Scale Invariance of Power-Constrained Doeblin Curves for Additive Noise Channels)**.** *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$. Suppose $K_\sigma : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ is an additive noise channel parameterized by σ such that $K_\sigma(A \mid \mathbf{x}) = \int_A \frac{1}{\sigma^d} f\big(\frac{\mathbf{y}-\mathbf{x}}{\sigma}\big) d\mathbf{y}$ where $f : \mathbb{R}^d \to \mathbb{R}_+$ is a probability density function. Then, if the power function is $M(z) = z^2$, $\mathrm{F}_{K_\sigma}^{\mathsf{UA}}(t;p) = \mathrm{F}_{K_1}^{\mathsf{UA}}(t;p/\sigma^2)$ for all $\sigma > 0$ and $p \ge 0$.*

Proposition 6 is proved in Section IV-B.

D. Bounds on Power-Constrained Doeblin Curves

In this subsection, we present upper and lower bounds on the power-constrained Doeblin curves for a general kernel and provide examples of kernels whose Doeblin curves may be computed in closed form. Throughout this analysis, let $(\mathcal{X}, \|\cdot\|)$ be a separable Banach space equipped with the Borel σ-algebra $\mathcal{F}_{\mathcal{X}}$ induced by the norm topology, let $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ be an arbitrary Polish space, and let $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ be an absolutely continuous Markov kernel. Let $\mathcal{B}(a,r) \subset \mathcal{X}$ denote the closed ball centered at $a \in \mathcal{X}$ with radius r in the $\|\cdot\|$-norm. Define functions $\theta : \mathcal{X} \times \mathbb{R}_+ \to [0,1]$ and $\Theta : \mathbb{R}_+ \to [0,1]$ as $\theta(a,r) = \rho_{\mathcal{B}(a,r)}(K)$ and $\Theta(r) = \sup_{a \in \mathcal{X}} \theta(a,r)$. We remark that Θ (and hence $\hat{\Theta}$, by Lemma 9, Part 2) are non-decreasing, because $\rho_{\mathcal{S}}$ is non-decreasing in $\mathcal{S}$ as previously mentioned. Since $\mathcal{X}$ is a Banach space, there exists an element $\mathbf{0} \in \mathcal{X}$ such that $\|\mathbf{0}\| = 0$; we denote this element in bold font to distinguish it from the scalar $0 \in \mathbb{R}$. Let

$$\gamma \triangleq \sup\left\{ \frac{\mathrm{rad}(\mathcal{S})}{|||\mathcal{S}|||_\infty} : \mathcal{S} \subset \mathcal{X},\, 0 < |||\mathcal{S}|||_\infty < \infty \right\} \tag{14}$$

denote the *Jung constant* of $\mathcal{X}$ [52], [53], i.e., the tightest multiplicative factor between the Chebyshev radius and diameter of any bounded subset of $\mathcal{X}$. In Euclidean space $\mathcal{X} = \mathbb{R}^d$, we have $\gamma \le d/(d+1)$ for a general norm $\|\cdot\|$ [54, Theorem 6], and $\gamma = 1/2$ for the ℓ^∞-norm [46, Section 5.3].

First, we present upper and lower bounds on the average extremal power-constrained Doeblin curve of K (similar to [25]).

Theorem 2 (Bounds on Average Extremal Doeblin Curve)**.** *The power-constrained Doeblin curve $\mathrm{F}_K^{\mathsf{AE}}$ satisfies the upper and lower bounds $t\,\theta\big(\mathbf{0}, M^{-1}\big(\frac{p}{t}\big)\big) \le \mathrm{F}_K^{\mathsf{AE}}(t;p) \le t\,\hat{\Theta}\big(2\gamma\, M^{-1}\big(\frac{p}{t}\big)\big)$ for all $t \in (0,1]$.*

The proof is provided in Section IV-C, utilizing the variational characterization of Doeblin coefficients from Theorem 1 and the maximal coupling characterization from Proposition 1.

If K acts as a convolution operator on $\mathbb{R}^d$ (e.g., additive noise), then $\theta(a,s)$ is independent of a, namely $\theta(a,s) = \Theta(s)$ for all $a \in \mathcal{X}$. This immediately leads to the following counterpart to [25, Corollary 5].

Corollary 1 (Average Extremal Doeblin Curves of Convolution Kernels)**.** *For any kernel K which acts as a convolution operator on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, and for which Θ is concave, we have $t\,\Theta\big(M^{-1}\big(\frac{p}{t}\big)\big) \le \mathrm{F}_K^{\mathsf{AE}}(t;p) \le t\,\Theta\big(\frac{2d}{d+1} M^{-1}\big(\frac{p}{t}\big)\big)$. Furthermore, if $d = 1$ or $\|\cdot\|$ is the ℓ^∞-norm, then $\mathrm{F}_K^{\mathsf{AE}}(t;p) = t\,\Theta\big(M^{-1}\big(\frac{p}{t}\big)\big)$.*

The equality in the $d = 1$ or ℓ^∞-norm case trivially follows, since the upper and lower bounds from Theorem 2 match. The next corollary provides three examples of closed-form Doeblin curves for convolution kernels on $\mathbb{R}$.

Corollary 2 (Examples of Average Extremal Doeblin Curves)**.** *Consider the Gaussian, Laplace, and q-Gaussian [55, Section 2.3] ($q = 2$) additive noise kernels on $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, given by $K_1(A \,|\, x\,;\, \sigma^2) = \int_A \frac{1}{\sigma\sqrt{2\pi}} \exp\big(-\frac{(y-x)^2}{2\sigma^2}\big)\, dy$, $K_2(A \,|\, x\,;\, b) = \int_A \frac{1}{2b} \exp\big(-\frac{|y-x|}{b}\big)\, dy$, and $K_3(A \,|\, x\,;\, \beta) = \int_A \frac{\sqrt{\beta}}{\pi}\big(\frac{1}{1+\beta(y-x)^2}\big) dy$, respectively. Then, under the norm $\|\cdot\| = |\cdot|$ and power*

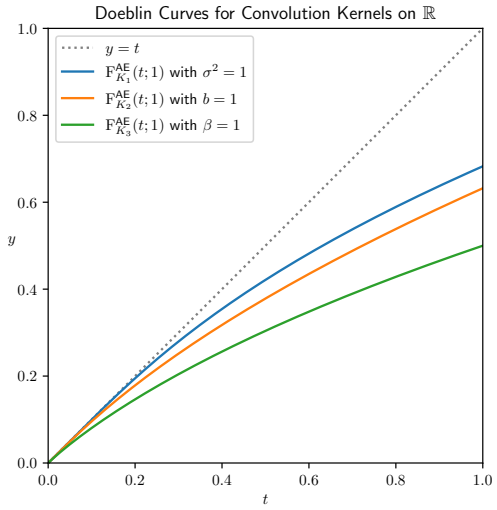

Fig. 2. Average extremal power-constrained Doeblin curves for the Gaussian, Laplace, and q-Gaussian ($q = 2$) additive noise kernels on $\mathbb{R}$.

function $M(z) = z^2$, the average extremal power-constrained Doeblin curves for these kernels are $\mathrm{F}^{\mathsf{AE}}_{K_1}(t;p,\sigma^2) = t(1 - 2\Phi(-\frac{1}{\sigma}\sqrt{p/t}))$, $\mathrm{F}^{\mathsf{AE}}_{K_2}(t;p,b) = t(1-\exp(-\frac{1}{b}\sqrt{p/t}))$, and $\mathrm{F}^{\mathsf{AE}}_{K_3}(t;p,\beta) = \frac{2t}{\pi}\arctan(\sqrt{\beta p/t})$, where $\Phi : \mathbb{R} \to (0,1)$ denotes the standard Gaussian CDF $\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{t^2}{2}\big)\, dt$.

Corollary 2 is proved in Section IV-C. By following the steps in this proof up to (97), we obtain that Θ is concave for any convolution kernel on $\mathbb{R}$ whose density function $g(z) = g(x-y) = \frac{dK}{dy}(y \mid x)$ is symmetric about $z = 0$ and non-increasing on $z \in [0,\infty)$. For such convolution kernels, the Doeblin curve $\mathrm{F}^{\mathsf{AE}}_K$ is concave in t, since Corollary 1 establishes that $\mathrm{F}^{\mathsf{AE}}_K$ is the *perspective* of the composition of a concave function Θ and a concave non-decreasing function M^{-1} (cf. [25, Remark 4]).

Figure 2 presents Doeblin curves for the convolution kernels in Corollary 2 with $\sigma^2 = 1$, $b = 1$, and $\beta = 1$. The trivial upper bound (11) due to submultiplicativity of ρ is depicted by all three curves remaining below the gray dotted line $y = t$. Since the derivatives of all three curves approach 1 as $t \to 0$, the complementary Doeblin coefficients for all three kernels are $\rho(K_1) = \rho(K_2) = \rho(K_3) = 1$. Hence, the contraction behavior of these kernels is only captured by their Doeblin curves through the analysis in Theorem 2 and the ensuing corollaries.

Lastly, we present bounds on the uniform average power-constrained Doeblin curve of a general kernel K. To do so, we impose additional regularity structure by requiring all input kernels W to have a fixed source space $\mathcal{U}$ with finiteness or boundedness assumptions. We also impose sub-Gaussianity assumptions on W as required.

Proposition 7 (Upper Bound on Uniform Average Doeblin Curve)**.** *Let $\mathcal{U}$ be a finite set. Let $\mathcal{G}$ be the set of all Markov kernels $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ satisfying the following assumptions:*

a) (Uniform average power constraint) $|||W|||_{\mathsf{UA}} \le p$.

b) (Sub-Gaussianity) Consider the set of random variables $\{Z_u\}_{u\in\mathcal{U}}$ given by $Z_u = M(\|X_u\|)$, $X_u \sim W(\cdot \mid u)$ for each $u \in \mathcal{U}$. Then, there exists $\sigma > 0$ such that for any $u \in \mathcal{U}$, $\mathbb{E}\big[e^{\lambda(Z_u - \mathbb{E}[Z_u])}\big] \le \exp(\frac{\sigma^2\lambda^2}{2})$ for all $\lambda \in \mathbb{R}$.

Then, the constrained Doeblin curve $\mathrm{F}_K(t;\mathcal{G})$ satisfies the upper bound $\mathrm{F}_K(t;\mathcal{G}) \le t\,\hat{\Theta}(2\gamma\, M^{-1}((p + \sigma\sqrt{2\log_e|\mathcal{U}|})/t))$ for all $t \in (0,1]$.

Proposition 8 (Upper Bound on Uniform Average Doeblin Curve)**.** *Let $(\mathcal{U}, d_{\mathcal{U}})$ be a Polish space endowed with a metric $d_{\mathcal{U}} : \mathcal{U}\times\mathcal{U} \to \mathbb{R}_+$ such that $(\mathcal{U}, d_{\mathcal{U}})$ is totally bounded, i.e., the ϵ-covering number $N(\epsilon,\mathcal{U},d_{\mathcal{U}}) \triangleq \min\{n \in \mathbb{N} : \exists\{u_1^*,\ldots,u_n^*\} \subseteq \mathcal{U}, \forall u \in \mathcal{U},\, \exists i \in [n],\, d_{\mathcal{U}}(u,u_i^*) \le \epsilon\}$ is finite for all $\epsilon > 0$. Let $\mathcal{G}$ be the set of all Markov kernels $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ satisfying the following assumptions:*

a) (Uniform average power constraint) $|||W|||_{\mathsf{UA}} \le p$.

b) (Sub-Gaussian increments) Consider the random process $\{Z_u\}_{u\in\mathcal{U}}$ given by $Z_u = M(\|X_u\|)$, where $\{X_u\}_{u\in\mathcal{U}} \sim \mathbb{P}$ is the maximal coupling of random variables with $X_u \sim W(\cdot \mid u)$ for each $u \in \mathcal{U}$, as defined in (34). Then, there exists $\sigma > 0$ such that for any $u, v \in \mathcal{U}$, $\mathbb{E}\big[e^{\lambda((Z_u - Z_v) - \mathbb{E}[Z_u - Z_v])}\big] \le \exp\big(\frac{\sigma^2 d_{\mathcal{U}}(u,v)^2\lambda^2}{2}\big)$ for all $\lambda \in \mathbb{R}$.

c) (Measurability) Under the above definition of $\{Z_u\}_{u\in\mathcal{U}}$, the quantity $\sup_{u\in\mathcal{U}}\{Z_u - \mathbb{E}[Z_u]\}$ is measurable.

Then, the constrained Doeblin curve $\mathrm{F}_K(t;\mathcal{G})$ satisfies the upper bound

$$\mathrm{F}_K(t;\mathcal{G}) \le t\,\hat{\Theta}\left(2\gamma\, M^{-1}\left(\frac{p}{t} + \frac{32\sigma}{t}\int_0^{|||\mathcal{U}|||_\infty} \sqrt{\log_e N(\epsilon,\mathcal{U},d_{\mathcal{U}})}\, d\epsilon\right)\right)$$

for all $t \in (0,1]$, where $|||\mathcal{U}|||_\infty \triangleq \sup_{u,v\in\mathcal{U}} d_{\mathcal{U}}(u,v)$ denotes the diameter of $(\mathcal{U}, d_{\mathcal{U}})$.

Proposition 9 (Lower Bound on Uniform Average Doeblin Curve)**.** *The power-constrained Doeblin curve* $\mathrm{F}_K^{\mathsf{UA}}$ *satisfies the lower bound* $\mathrm{F}_K^{\mathsf{UA}}(t;p) \geq t\,\theta(\mathbf{0}, M^{-1}(\frac{p}{t}))$ *for all* $t \in (0,1]$.

Propositions 7 to 9 are proved in Section IV-C by adapting the arguments from the proof of Theorem 2. We emphasize that Proposition 9 does *not* apply to the constrained Doeblin curves $\mathrm{F}_K(t;\mathcal{G})$ from Propositions 7 and 8, which are upper-bounded by $\mathrm{F}_K^{\mathsf{UA}}$ in general due to the stricter constraints imposed on the input kernels W when defining the former. We provide examples of non-trivial kernels satisfying the preconditions of Propositions 7 and 8 in Appendix C.

III. Main Results on Applications

In this section, we present three illustrative applications of Doeblin curves to other areas of information theory and learning theory. Firstly, we derive generalization error bounds for iterative optimization algorithms operating over unbounded feasible sets. Secondly, we establish lower bounds for reliable computation using circuits of noisy q-ary gates. Lastly, we introduce a new definition of (ϵ,δ)-differential privacy based on our variational characterization of Doeblin coefficients, and provide improved privacy bounds for online iterative optimization.

A. Bounds on Generalization Error

In this subsection, we use Doeblin curves to extend existing information-theoretic generalization error bounds for noisy iterative optimization algorithms in [41] to settings where the feasible set has infinite diameter or the optimization problem is unconstrained altogether (also see [56]). We utilize the information-theoretic framework introduced in [41], which provides generalization error bounds for compact feasible sets by leveraging the Dobrushin contraction coefficient of additive noise channels. Their results indicate that data points used in earlier iterations have a decaying contribution to the generalization error due to the cumulative effect of noise injection, with the rate of decay governed by the Dobrushin coefficient of the underlying noise channel. This analysis yields a non-trivial bound only when the feasible set has finite diameter, since the Dobrushin coefficient trivially equals 1 otherwise. To address feasible sets with infinite diameter, we utilize Doeblin[4] curves as a finer-grained tool to capture information decay for data points used in earlier iterations even when Dobrushin coefficients fail to do so, thereby yielding non-trivial generalization error bounds despite the absence of coarser coefficient-based contraction.

Formally, following [41], we begin by considering a general (possibly non-convex) stochastic optimization problem,

$$\min_{w\in\mathcal{W}} G_\mu(w)\,, \quad \text{where} \quad G_\mu(w) \triangleq \mathop{\mathbb{E}}_{Z\sim\mu}[g(w,Z)] = \int_{\mathcal{Z}} g(w,\cdot)\,d\mu\,,$$

where $w \in \mathcal{W} \subseteq \mathbb{R}^d$ are the model parameters constrained to the (potentially unbounded) feasible set $\mathcal{W}$, μ is the underlying data distribution over the samples Z belonging to the sample space $\mathcal{Z}$, and $g : \mathcal{W}\times\mathcal{Z} \to \mathbb{R}_+$ is the loss function. In practice, the true data distribution μ is usually unknown, and only a dataset $\mathcal{S} \triangleq (Z_1,\ldots,Z_n)$ consisting of n independent and identically distributed samples $Z_i \sim \mu$ is available. Thus, we instead consider the empirical risk minimization (ERM) problem

$$\min_{w\in\mathcal{W}} G_{\mathcal{S}}(w)\,, \quad \text{where} \quad G_{\mathcal{S}}(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} g(w,Z_i)\,,$$

with the aim of finding a solution which generalizes well to the original problem of minimizing G_μ. To this end, we consider the following noisy iterative algorithm. The algorithm is initialized with an arbitrary parameter $W_0 \in \mathcal{W}$. Before the optimization process, T disjoint mini-batch index sets $\mathcal{B}_1,\ldots,\mathcal{B}_T$ are chosen deterministically and fixed, where $\mathcal{B}_t \subseteq [n]$ contains the indices of the samples comprising the mini-batch at iteration t. Next, the algorithm performs T iterations by updating the parameters W_t according to the rule

$$W_t \triangleq \mathop{\mathrm{proj}}_{\mathcal{W}}\left(W_{t-1} - \frac{\eta_t}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t}\nabla g(W_{t-1},Z_i) + m_t N_t\right)$$

for each iteration $t \in [T]$, where the projection operator $\mathrm{proj}_{\mathcal{W}}$ ensures that the parameters remain within the feasible set $\mathcal{W}$, η_t is the learning rate, ∇g represents the partial gradient of g with respect to its first argument (i.e., the model parameters), $N_t \sim \mathsf{Normal}(0,I)$ represents independent additive standard Gaussian noise (where I is the identity matrix of appropriate dimension),[5] and m_t controls the noise magnitude. We remark that several algorithms can be expressed in this form, such as Stochastic Gradient Langevin Dynamics (SGLD) and Differentially-Private Stochastic Gradient Descent (DP-SGD).

[4]In all applications of Doeblin curves, suitable constraints may be placed on the Doeblin curve to obtain tighter bounds by incorporating prior knowledge about the input distributions. For example, when deriving our results on generalization error, we only utilize the Doeblin curve $\mathrm{F}_\Phi^{\mathsf{UA}}$ of the additive noise mechanism to quantify the contraction between *two* distributions pushed through the mechanism. Hence, we may instead consider the constrained curve which only includes input kernels with $|\mathcal{U}| = 2$, which reduces to the *Dobrushin curve* of the noise mechanism. Nonetheless, we present all results in this paper in terms of the more general *Doeblin curve* $\mathrm{F}_\Phi^{\mathsf{UA}}$ for conceptual simplicity.

[5]We focus on Gaussian noise for conceptual simplicity, although our results can be easily extended to other noise distributions such as Laplace.

We assume that the loss function is bounded above by a constant $A > 0$, i.e., $g(w, z) \leq A$ for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$, and that the gradient of the loss is bounded in magnitude, i.e.,

$$\forall w \in \mathcal{W},\, \forall z \in \mathcal{Z},\;\; \|\nabla g(w, z)\|_2 \leq L\,. \tag{15}$$

Our objective is to upper-bound the expected generalization error at the final model parameters W_T after T iterations, i.e., $|\mathbb{E}[G_\mu(W_T) - G_{\mathcal{S}}(W_T)]|$, where the expectation is taken with respect to the randomness in the dataset $\mathcal{S}$ and the noise $N_1, \ldots, N_T$ throughout the optimization process.

Let $\mathrm{F}_\Phi^{\mathsf{UA}}$ denote the uniform average power-constrained Doeblin curve of the $\mathsf{Normal}(0, I)$ additive noise channel under the power function $M(z) = z^2$. Our main result is formally stated as follows.

Theorem 3 (Expected Generalization Error)**.** *The expected generalization error satisfies*

$$|\mathbb{E}[G_\mu(W_T) - G_{\mathcal{S}}(W_T)]| \leq A \sum_{t=1}^{T} \frac{|\mathcal{B}_t|}{n}\, \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Bigg(\cdots \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\bigg(\frac{\eta_t \sigma_{t-1}}{m_t\,|\mathcal{B}_t|}; p_{t+1}\bigg) \cdots ; p_T\Bigg)\,, \tag{16}$$

where for each $t \in \{0, \ldots, T-1\}$ we define

$$\sigma_t \triangleq \mathop{\mathbb{E}}_{\substack{(W_t, Z)\\ \sim P_{W_t} \otimes \mu}} \Big[\Big\|\nabla g(W_t, Z) - \mathop{\mathbb{E}}_{\substack{(W_t, Z)\\ \sim P_{W_t} \otimes \mu}} [\nabla g(W_t, Z)]\Big\|_2\Big]\,,$$

for each $t \in \{2, \ldots, T\}$ we define

$$p_t \triangleq \frac{1}{m_t^2}\Bigg(2 \max_{i \in \bigcup_{s=1}^{t-1} \mathcal{B}_s} \sup_{z \in \mathcal{Z}} \mathbb{E}\Big[\|W_{t-1}\|_2^2 \,\Big|\, Z_i = z\Big] + 2\eta_t^2 L^2\Bigg)\,,$$

and the composition of Doeblin curves in (16) reduces to the identity function (i.e., no Doeblin curves are applied) for the $t = T$ term of the sum.

Theorem 3 is proved in Section V-A. The argument utilizes the following lemma to express the expected generalization error in terms of the TV-information between the model parameter W_T and each data sample Z_i.

Lemma 1 (Generalization Error and TV-Information [41, Lemma 1])**.** *Define the TV-information between two random variables X and Y as*

$$I_{\mathsf{TV}}(X; Y) \triangleq \|P_{X,Y} - P_X \otimes P_Y\|_{\mathsf{TV}}\,, \tag{17}$$

where $P_{X,Y}$ is the joint distribution of (X, Y), and P_X and P_Y are the marginal distributions of X and Y, respectively. Then, the expected generalization error satisfies

$$|\mathbb{E}[G_\mu(W_T) - G_{\mathcal{S}}(W_T)]| \leq \frac{A}{n} \sum_{i=1}^{n} I_{\mathsf{TV}}(W_T; Z_i)\,.$$

Our principal contribution is to bound the TV-information between the output W_T of the learning algorithm and the data samples Z_i by utilizing properties of the Gaussian noise channel's Doeblin curve, as formalized in the following lemma which holds for possibly non-compact $\mathcal{W}$.

Lemma 2 (Recursive Bound on TV-Information)**.** *The TV-information between the final model parameter W_T and a data sample Z_i satisfies*

$$I_{\mathsf{TV}}(W_T; Z_i) \leq \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Bigg(\cdots \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\bigg(\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Big(I_{\mathsf{TV}}(W_t; Z_i); p_{t+1}\Big); p_{t+2}\bigg) \cdots ; p_T\Bigg)\,, \tag{18}$$

where t is the iteration on which Z_i is used in the update rule (i.e., $t \in \mathcal{B}_i$), for each $s \in \{t+1, \ldots, T\}$ we define

$$p_s \triangleq \frac{1}{m_s^2}\Bigg(2 \max_{i \in \bigcup_{r=1}^{s-1} \mathcal{B}_r} \sup_{z \in \mathcal{Z}} \mathbb{E}\Big[\|W_{s-1}\|_2^2 \,\Big|\, Z_i = z\Big] + 2\eta_s^2 L^2\Bigg)\,, \tag{19}$$

and the composition of Doeblin curves in (18) reduces to the identity function if $t = T$ (in which case (18) trivially holds with equality).

Lemma 2 is proved in Section V-A. We remark that bounds on the second moment $\mathbb{E}[\|W_t\|_2^2]$ of the iterates W_t have been derived in the literature, e.g., under additive Gaussian noise [56]. These results can often be directly translated into bounds on p_t. In principle, tighter bounds can be obtained by considering the second moment $\mathbb{E}[\|W_t - \tilde{w}\|_2^2]$ of the iterates with respect to a reference point $\tilde{w} \in \mathcal{W}$ chosen as the center of the feasible set or, ideally, the optimal solution w^*. However, since w^* is unknown in practice, we use the worst-case bounds on the second moment of the iterates instead.

B. Reliable Computation Using Noisy q-ary Gates

In this subsection, we use Doeblin curves to establish information-theoretic bounds on the expressive power of circuits of noisy gates for the computation of q-ary functions. This builds on results for Boolean functions in [25]. We consider n-input circuits which compute a single output value using b-input gates, where the result of each gate is perturbed by independent noise. Historically, the study of such architectures began with von Neumann's pioneering work on fault-tolerant Boolean circuits [57]–[59], and has been largely limited to binary operations. We expand this line of study to multivalued logic systems, as motivated by recent advancements in classical and quantum information processing [60].

Formally, we model a noisy circuit over a q-ary alphabet $\mathcal{Q} = \{\xi_1, \dots, \xi_q\} \subset \mathbb{R}^d$ as a directed acyclic graph with n *input vertices* $X_1, \dots, X_n \in \mathcal{Q}$ indexed by $i \in [n]$. The inputs are processed by a collection of m *gate vertices* indexed by $j \in [m]$, each of which takes $b_j \le b$ $\mathbb{R}^d$-valued arguments and computes a $\mathcal{Q}$-valued result $Z_j \in \mathcal{Q}$, which is then corrupted by independent noise from some arbitrary fixed kernel Φ to produce $Y_j \in \mathbb{R}^d$. Formally, letting $\mathcal{N}_j$ and $\mathcal{M}_j$ denote the *incoming edge sets* of input vertices and gate vertices feeding into gate j, respectively, we have

$$Z_j \triangleq \Gamma_j\Big(\{X_i\}_{i\in\mathcal{N}_j}, \{Y_{j'}\}_{j'\in\mathcal{M}_j}\Big), \;\; Y_j \sim \Phi(\cdot \mid Z_j)$$

for some deterministic gate function $\Gamma_j : (\mathbb{R}^d)^{b_j} \to \mathcal{Q}$, where $b_j = |\mathcal{N}_j| + |\mathcal{M}_j| \le b$. Without loss of generality, we assume the gates are indexed in topological order, i.e., $\mathcal{M}_j \subseteq [j-1]$ for each $j \in [m]$. At the end, the circuit produces a single *output vertex* Y_m with no outgoing edges. We assume there exists a directed path from each X_i to Y_m, so none of the circuit inputs are discarded. We note that this model can be seen as a q-ary generalization of the setup considered in [25, Section 5.3].

One goal of such a circuit is to compute a function $G : \mathcal{Q}^n \to \mathcal{Q}$ with error probability at most $\mathrm{P}_{\mathsf{error}} > 0$. Namely, there must exist some fixed decoding function $\hat{G} : \mathbb{R}^d \to \mathcal{Q}$ such that for any input vector $(x_1, \dots, x_n) \in \mathcal{Q}^n$, $\mathbb{P}(\hat{G}(Y_m) = G(x_1, \dots, x_n)) \ge 1 - \mathrm{P}_{\mathsf{error}}$, where the probability is taken with respect to all the noise in the circuit. In the binary case ($q = |\mathcal{Q}| = 2$), by Le Cam's relation, the error probability of the optimal decoder can be controlled by the TV distance between the marginal distributions of Y_m induced by setting some "initial" input vertex to 0 or 1, respectively, while fixing all other input values [25, Eq. 146]. So, several works study the question of how information measures contract over fault-tolerant circuits, e.g., [25] recursively upper-bounds TV distance while [9] bounds mutual information. Propelled by such analyses, and in light of the interpretation of complementary Doeblin coefficients as a multi-way generalization of TV distance [19, Theorem 2], we analyze the complementary Doeblin coefficient of the q induced marginal distributions of Y_m as a first step towards understanding information propagation in noisy q-ary circuits. (Intuitively, perceiving a circuit as a Markov chain, the analysis of TV distance in [25] is related to the coupling time of two copies of the chain with two different values at the "initial" vertex [61]. On the other hand, our analysis of Doeblin coefficients is related to the coupling time of q copies of the chain initiated at all q possible values.) Our main contribution is the following theorem, which relates the Doeblin coefficient to the noise mechanism's Doeblin curve.

Theorem 4 (Upper Bound on Circuit Output Divergence)**.** *Fix $\epsilon > 0$. For all n-input circuits of noisy b-input gates with sufficiently large n, there exists some input vertex, which we take to be X_1 without loss of generality, such that for any fixed values $x_2, \dots, x_n \in \mathcal{Q}$ of the remaining inputs,*

$$\rho\Big(\big[P_{Y_m}^{(1)}, \dots, P_{Y_m}^{(q)}\big]\Big) \le \max\Big\{t \in [0,1] : \hat{\mathrm{F}}_{\Phi}^{\mathsf{UA}}(\min\{1, bt\}; p) = t\Big\} + \epsilon\,,$$

where $P_{Y_m}^{(\ell)}$ is the marginal distribution of Y_m induced by the circuit when setting $X_1 = \ell$ and $X_i = x_i$ for all $i > 1$, $\mathrm{F}_{\Phi}^{\mathsf{UA}}$ is the uniform average power-constrained Doeblin curve of the noise kernel Φ for some norm $\|\cdot\|$ on $\mathbb{R}^d$ and power function $M : \mathbb{R}_+ \to \mathbb{R}_+$, and $p \triangleq \max_{\xi\in\mathcal{Q}} M(\|\xi\|)$.

Theorem 4 represents an information-theoretic limit on the quality of any output decoder, since decoding performance improves as output divergence increases. We defer the proof to Section V-B. The proof constructs a coupling of the induced marginal distributions at each gate output Y_j, using the maximal coupling characterization of Doeblin coefficients (34) to refine the couplings from earlier gates. This is formalized in the following lemma. For notational convenience, we denote collections of superscripted variables as $w^{(1:q)} = (w^{(1)}, \dots, w^{(q)})$.

Lemma 3 (Stepwise Coupling Construction)**.** *Let $\mathcal{U}$ be a finite subset of a separable Banach space. Let $(\mathcal{V}, \mathcal{F}_{\mathcal{V}})$ be a Polish space and let $\Phi : \mathcal{U} \times \mathcal{F}_{\mathcal{V}} \to [0,1]$ be a Markov kernel from $\mathcal{U}$ to $\mathcal{V}$. Let $P_U^{(1)}, \dots, P_U^{(q)}$ be a collection of probability measures on $\mathcal{U}$, and let $P_V^{(1)}, \dots, P_V^{(q)}$ be the respective probability measures on $\mathcal{V}$ induced by pushing $P_U^{(\ell)}$ through Φ for each $\ell \in [q]$, i.e.,*

$$\forall A \in \mathcal{F}_{\mathcal{V}}, \;\; P_V^{(\ell)}(A) \triangleq \int_{\mathcal{U}} \Phi(A \mid u)\, dP_U^{(\ell)}(u)\,. \tag{20}$$

Then, given any coupling π_U of $P_U^{(1)}, \dots, P_U^{(q)}$, there exists a coupling π_V of $P_V^{(1)}, \dots, P_V^{(q)}$ satisfying

$$\mathop{\mathbb{P}}_{V^{(1:q)}\sim\pi_V}\Big(\neg\Big(V^{(1)} = \cdots = V^{(q)}\Big)\Big) \le \hat{\mathrm{F}}_{\Phi}^{\mathsf{UA}}\Big(\mathop{\mathbb{P}}_{U^{(1:q)}\sim\pi_U}\Big(\neg\Big(U^{(1)} = \cdots = U^{(q)}\Big)\Big); p\Big)\,, \tag{21}$$

where we define $p \triangleq \max_{u\in\mathcal{U}} M(\|u\|)$.

Lemma 3 is proved in Section V-B by constructing π_V as the weighted average (with respect to $U^{(1:q)} \sim \pi_U$) of the maximal coupling (34) of $\Phi(\cdot \mid U^{(1)}), \ldots, \Phi(\cdot \mid U^{(q)})$. Iterating this construction for each gate leads to repeated application of the Doeblin curve $\mathrm{F}_\Phi^{\mathsf{UA}}$, which converges to the greatest fixed point of a transformed version of $\mathrm{F}_\Phi^{\mathsf{UA}}$.

C. Relation to Differential Privacy and Online Algorithms

In this subsection, we utilize our variational characterization of Doeblin coefficients to motivate a new definition of group local differential privacy (LDP). The standard information-theoretic[6] definition of LDP asserts that a mechanism $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to [0,1]$, which is nothing but a Markov kernel, is (ϵ,δ)-LDP if for any inputs $x, x' \in \mathcal{X}$, both $K(\cdot \mid x)$ and $K(\cdot \mid x')$ exhibit similar distributions [42], [63]–[65], i.e.,

$$\forall x, x' \in \mathcal{X},\ \forall A \in \mathcal{F}_\mathcal{Y},\ \ K(A \mid x) - e^\epsilon K(A \mid x') \le \delta\,. \tag{22}$$

Subtracting both sides from 1, this is equivalent to

$$\forall x, x' \in \mathcal{X},\ \forall A \in \mathcal{F}_\mathcal{Y},\ \ K(A^\mathrm{c} \mid x) + e^\epsilon K(A \mid x') \ge 1 - \delta\,. \tag{23}$$

This rearrangement has an interesting interpretation: Consider a binary hypothesis testing scenario that aims to distinguish whether the output Y of a mechanism is derived from the distribution $K(\cdot \mid x)$ versus $K(\cdot \mid x')$ for any fixed x and x'. Let $A \subseteq \mathcal{Y}$ be any possible choice of the rejection region for the hypothesis $Y \sim K(\cdot \mid x')$. The reformulated version in (23) essentially highlights the impossibility of achieving a very low weighted sum of type-I and type-II errors from the data Y derived from a differentially private mechanism K (see [66, Theorem 2.1] for more details).

Motivated by our variational characterization of Doeblin coefficients from Theorem 1, we extend the standard definition of LDP to a group setting.[7] Consider any group size $n \ge 2$ and any $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}_+^n$ such that $\epsilon_1 = 0$ and $\epsilon_i \le \epsilon_j$ for all $i < j$. A mechanism K is said to be $(\boldsymbol{\epsilon}, \delta, n)$-*LDP* if for any $x_1, \ldots, x_n \in \mathcal{X}$ and any $\mathcal{F}_\mathcal{Y}$-measurable n-partition $A_1, \ldots, A_n$ of $\mathcal{Y}$,

$$\sum_{i=1}^n e^{\epsilon_i} K(A_i \mid x_i) \ge 1 - \delta\,. \tag{24}$$

This definition may be interpreted in the context of an n-ary hypothesis testing problem. Given a set of n hypotheses $H_i : Y \sim K(\cdot \mid x_i)$ indexed by $i \in [n]$, where Y is the observed variable, the generalized definition in (24) states that the problem of identifying a single false hypothesis has a large weighted sum error for any possible choices of rejection regions. Any reasonable test for this hypothesis testing problem would partition the space $\mathcal{Y}$ into n rejection regions $A_1, \ldots, A_n$, and the test rejects hypothesis H_i if Y takes on a value in A_i. Therefore, $K(A_i \mid x_i)$ represents the conditional probability of error given H_i in this scenario. For $n = 2$, the notion of $(\boldsymbol{\epsilon}, \delta, 2)$-LDP reduces to the standard definition in (22). Furthermore, K being $(\boldsymbol{\epsilon}, \delta, n)$-LDP for $n \ge 2$ is a stronger condition, as it implies that K is also $([\epsilon_1, \ldots, \epsilon_m], \delta, m)$-LDP for all $2 \le m \le n$.

Motivated by this formulation, we define weighted analogues of the Doeblin and complementary Doeblin coefficients for a Markov kernel $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to [0,1]$:

$$\tau_{\boldsymbol{\epsilon}}(K, n) \triangleq \inf_{x_1,\ldots,x_n \in \mathcal{X}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^n e^{\epsilon_i} K(A_i \mid x_i)\,, \tag{25}$$

$$\rho_{\boldsymbol{\epsilon}}(K, n) \triangleq 1 - \tau_{\boldsymbol{\epsilon}}(K, n)\,. \tag{26}$$

We remark that $\tau_{\boldsymbol{\epsilon}}(K, n) \in [0,1]$ (and so $\rho_{\boldsymbol{\epsilon}}(K, n) \in [0,1]$), because for any $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$\tau_{\boldsymbol{\epsilon}}(K, n) \overset{(a)}{\le} e^{\epsilon_1} K(\mathcal{Y} \mid x_1) + \sum_{i=2}^n e^{\epsilon_i} K(\emptyset \mid x_i) \overset{(b)}{=} K(\mathcal{Y} \mid x_1) \overset{(c)}{=} 1\,,$$

where (a) holds by upper-bounding the infima in (25) with a specific instance, (b) holds because $\epsilon_1 = 0$, and (c) holds because K is a Markov kernel. Moreover, akin to how [42], [67] reformulated local differential privacy in terms of the E_γ contraction coefficient,[8] $\rho_{\boldsymbol{\epsilon}}(K, n)$ captures $(\boldsymbol{\epsilon}, \delta, n)$-LDP exactly: $\rho_{\boldsymbol{\epsilon}}(K, n) \le \delta$ if and only if K is $(\boldsymbol{\epsilon}, \delta, n)$-LDP (see proof of Theorem 5 in Section V-C).

In addition, we note a connection between $\rho_{\boldsymbol{\epsilon}}$ and the concept of min-DeGroot distance from [19], which is construed as a generalization of Bayes statistical information. In this context, a hidden random variable $X \in \mathcal{X}$ (where $\mathcal{X} = \{x_1, \ldots, x_n\}$ has cardinality $|\mathcal{X}| = n$) follows a prior distribution $\boldsymbol{\lambda} \in \mathbb{R}^n$, and a random variable $Y \in \mathcal{Y}$ is observed according to an observation

[6] This differs slightly from the learning-theoretic definition, which considers input *datasets of multiple users* differing by only one user [62, Definition 2.4].

[7] Accordingly, this is different from learning-theoretic definitions of group differential privacy based on distance between the input datasets [62, Theorem 2.2].

[8] We note that the E_γ-divergence is a weighted version of TV distance, which is equivalent (when appropriately scaled) to Bayes statistical information or DeGroot distance [68].

model denoted by K. Let $\hat{X} \in \mathcal{X}$ be any (possibly randomized) estimator of X based on Y, such that $X \to Y \to \hat{X}$ forms a Markov chain. [19] defines the min-DeGroot distance $\tau_{\min}(\boldsymbol{\lambda}, K)$ as the reduction in Bayes risk when observing the data Y compared to not observing Y. Specifically,

$$\tau_{\min}(\boldsymbol{\lambda}, K) \triangleq \min_{i\in[n]} \lambda_i - \int_{\mathcal{Y}} \left(\min_{i\in[n]} \lambda_i \frac{dK}{d\mu}(\cdot \mid x_i) \right) d\mu \,,$$

where μ is the common dominating measure for $K(\cdot \mid x_1), \ldots, K(\cdot \mid x_n)$ from (2). By setting $\lambda_i = e^{\epsilon_i} / \sum_{i=1}^n e^{\epsilon_i}$ and applying Proposition 2, the min-DeGroot distance can be interpreted as a rescaled version of $\rho_{\boldsymbol{\epsilon}}(K, n)$, namely,

$$\rho_{\boldsymbol{\epsilon}}(K, n) = \left(\sum_{i=1}^n e^{\epsilon_i} \right) \tau_{\min}(\boldsymbol{\lambda}, K) \,. \tag{27}$$

Therefore, the definition of $(\boldsymbol{\epsilon}, \delta, n)$-LDP is essentially a rescaled version of the generalized notion of Bayes statistical information obtained from the set $\{x_1, \ldots, x_n\}$. Our proposed definition of an $(\boldsymbol{\epsilon}, \delta, n)$-LDP mechanism underscores that, given the processed data of the entire group, obtaining any substantial "information" (quantified in the Bayes statistical information sense) about any individual member is "hard."

In the following theorem, we prove contraction properties of $\rho_{\boldsymbol{\epsilon}}(K, n)$. By (27), these results translate to contraction properties of $\tau_{\min}(\boldsymbol{\lambda}, K)$.

Theorem 5 (Contraction of $\rho_{\boldsymbol{\epsilon}}(K, n)$)**.** *Let $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0, 1]$ be an absolutely continuous Markov kernel. For any group size $n \geq 2$ and any $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}_+^n$ such that $\epsilon_1 = 0$ and $\epsilon_i \leq \epsilon_j$ for all $i < j$, we have the following properties:*

1) (Submultiplicativity) *For any Markov kernel $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0, 1]$,*

$$\rho_{\boldsymbol{\epsilon}}(WK, n) \leq \rho_{\boldsymbol{\epsilon}}(W, n)\, \rho_{\boldsymbol{\epsilon}}(K, n) \,.$$

2) (Contraction behavior) *K is $(\boldsymbol{\epsilon}, \delta, n)$-LDP iff for any Markov kernel $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0, 1]$,*

$$\rho_{\boldsymbol{\epsilon}}(WK, n) \leq \delta \rho_{\boldsymbol{\epsilon}}(W, n) \,.$$

3) (Doeblin curves) *For any Markov kernel $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0, 1]$ satisfying $\|\!|W|\!\|_{\mathsf{UA}} \leq p$,*

$$\rho_{\boldsymbol{\epsilon}}(WK, n) \leq \mathrm{F}_K^{\mathsf{UA}}(\rho_{\boldsymbol{\epsilon}}(W, n); p) \,. \tag{28}$$

Theorem 5 is proved in Section V-C. Parts 1 and 2 imply that $\rho_{\boldsymbol{\epsilon}}$ exhibits contraction-coefficient-like properties akin to ρ, with Part 1 acting as a meta-SDPI for $\rho_{\boldsymbol{\epsilon}}$ (cf. [16, Proposition 3]) and Part 2 characterizing δ as the contraction coefficient of this meta-SDPI. In particular, the *data processing inequality* for $\rho_{\boldsymbol{\epsilon}}$, i.e.,

$$\rho_{\boldsymbol{\epsilon}}(WK, n) \leq \rho_{\boldsymbol{\epsilon}}(W, n) \,, \tag{29}$$

is an immediate consequence due to $\rho_{\boldsymbol{\epsilon}}(K, n) \in [0, 1]$. In addition, considering the connection between $\rho_{\boldsymbol{\epsilon}}$ and the min-DeGroot distance highlighted above, Part 1 can be restated as an SDPI for min-DeGroot distances, i.e., for any Markov kernel $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0, 1]$ with finite $\mathcal{U}$, $\tau_{\min}(\boldsymbol{\lambda}, WK) \leq \rho_{\boldsymbol{\epsilon}}(K, |\mathcal{U}|)\, \tau_{\min}(\boldsymbol{\lambda}, W)$.

Equation (28) and the techniques developed above provide tools for tighter privacy guarantees for various mechanisms. Using these results, we now derive differential privacy guarantees for online iterative algorithms using Doeblin curves. While prior works such as [67] rely on compactness assumptions, our focus is on extending such analyses to potentially unbounded spaces where it can be difficult to capture privacy using classical contraction-coefficient-based techniques. To this end, we consider and build on the online learning framework presented in [67, Section IV-B], where the learner sequentially minimizes a sequence of convex objectives $\{g_t\}_{t=1}^n$ over a parameter space $\mathcal{W} \subseteq \mathbb{R}^d$. The protocol proceeds as follows. The learner initializes with a random parameter $W_0 \in \mathcal{W}$, and at each iteration $t \in [T]$, the cost function g_t is revealed. The learner then performs the update

$$W_t \triangleq \operatorname*{proj}_{\mathcal{W}}(W_{t-1} - \eta_t \nabla g_t(W_{t-1}) + m_t N_t) \,, \tag{30}$$

where the projection operator $\mathrm{proj}_{\mathcal{W}}$ ensures that the parameters remain within the feasible set $\mathcal{W}$, $\eta_t > 0$ is the learning rate, and m_t scales the standard Gaussian noise $N_t \sim \mathsf{Normal}(0, I)$ (where I is the identity matrix of appropriate dimension). By taking $g_t(w) \triangleq \ell(w, Z_t)$ where $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$ is a loss function, this framework subsumes one-pass ERM, where the learner approximates the solution of $\min_{w\in\mathcal{W}} \frac{1}{T} \sum_{t=1}^T \ell(w, Z_t)$ with data points $Z_1, \ldots, Z_T \in \mathcal{Z}$ revealed sequentially. The goal is to ensure that the final iterate W_T satisfies a privacy guarantee. We assume the gradient of each objective g_t is uniformly bounded, i.e.,

$$\forall t \in [T], \forall w \in \mathcal{W}, \quad \|\nabla g_t(w)\|_2 \leq L \,. \tag{31}$$

Existing LDP guarantees for online algorithms [67] rely on the contraction coefficient of E_γ-divergence, which quantifies the decay of information due to additive noise. However, for $\mathcal{W}$ with infinite diameter, the contraction coefficient for E_γ-divergence degenerates to 1. Below, we utilize the contraction bound on $\rho_{\boldsymbol{\epsilon}}$ in terms of Doeblin curves $\mathrm{F}_\Phi^{\mathsf{UA}}$ from Theorem 5 to quantify the contraction induced by the noise N_t.

Theorem 6 (Differential Privacy for Unconstrained Online Learning)**.** *Let $W_0^{(i)} \sim P_{W_0}^{(i)}$ for $i \in [n]$ denote n different initializations of the online learning process. Let $W_T^{(i)} \sim P_{W_T}^{(i)}$ be the respective output parameters after T iterations of the update rule (30). Let $\mathrm{F}_\Phi^{\mathsf{UA}}$ be the uniform average power-constrained Doeblin curve (with power function $M(z) = z^2$) of the $\mathsf{Normal}(0, I)$ additive noise channel. Then,*

$$\rho_{\boldsymbol{\epsilon}}\left(\left[P_{W_T}^{(1)}, \ldots, P_{W_T}^{(n)}\right], n\right) \le \mathrm{F}_\Phi^{\mathsf{UA}}\left(\cdots \mathrm{F}_\Phi^{\mathsf{UA}}\left(\rho_{\boldsymbol{\epsilon}}\left(\left[P_{W_0}^{(1)}, \ldots, P_{W_0}^{(n)}\right], n\right); p_1\right) \cdots ; p_T\right),$$

where for each $t \in [T]$ we define

$$p_t \triangleq \frac{1}{m_t^2}\left(2 \max_{i \in [n]} \mathbb{E}\left[\left\|W_{t-1}^{(i)}\right\|_2^2\right] + 2\eta_t^2 L^2\right). \tag{32}$$

Theorem 6 is proved in Section V-C using Theorem 5, Part 3. We note that the above result can also be used to derive privacy amplification bounds for online learning algorithms as in [67] for more general $\mathcal{W}$. In addition, Theorem 6 can also be perceived as a variation of Lemma 2 that bounds an n-way divergence instead of TV-information.

IV. Proofs of Main Results on Doeblin Curves

In this section, we prove the main results presented in Section II, pertaining to fundamental properties of Doeblin coefficients and curves.

A. Proofs of Doeblin Characterizations

In this subsection, we first prove Propositions 1 and 2. Next, we introduce essential preliminaries (such as lattice infimum) required for generalizing the argument to Polish spaces. Finally, we present a detailed step-by-step proof of Theorem 1.

Proof of Proposition 1. We have

$$\tau(K) \overset{(a)}{=} \bigwedge_{x \in \mathcal{X}} K(\mathcal{Y} \mid x) \overset{(b)}{=} \sup_{\mathbb{P}: Y_x \sim K(\cdot|x)} \mathbb{P}(\forall x, x' \in \mathcal{X},\, Y_x = Y_{x'}), \tag{33}$$

where (a) holds by the greatest common component characterization of Doeblin coefficient (5) and (b) holds by [19, Theorem 2], [47, Chapter 3, Theorem 7.3, p. 107]. □

We remark that for any kernel K, the supremum in (33) is achieved by the *maximal coupling* given by (cf. [19, Eq. 34], [47, Chapter 3, Theorem 7.3, p. 107])

$$\forall x \in \mathcal{X},\;\; Y_x \triangleq \begin{cases} Y^*, & \text{if } I = 1, \\ \tilde{Y}_x, & \text{if } I = 0, \end{cases} \tag{34}$$

where $\alpha = \bigwedge_{x \in \mathcal{X}} K(\mathcal{Y} \mid x)$ and the random variables I, Y^*, and $\{\tilde{Y}_x\}_{x \in \mathcal{X}}$ are sampled independently from the probability measures

$$I \sim \mathsf{Bernoulli}(\alpha), \quad Y^* \sim \frac{\bigwedge_{x \in \mathcal{X}} K(\cdot \mid x)}{\alpha}, \quad \forall x \in \mathcal{X},\;\; \tilde{Y}_x \sim \frac{K(\cdot \mid x) - \bigwedge_{x' \in \mathcal{X}} K(\cdot \mid x')}{1 - \alpha}.$$

(Note that Y^* and $\tilde{Y}_x$ are unused in the case that their respective probability measures are undefined, i.e., $\alpha = 0$ or $\alpha = 1$.)

Proof of Proposition 2. Fix an absolutely continuous Markov kernel $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to [0, 1]$ with common dominating measure $\mu : \mathcal{F}_\mathcal{Y} \to \mathbb{R}_+$. Fix $x_1, \ldots, x_n \in \mathcal{X}$ and $\gamma_1, \ldots, \gamma_n \ge 0$.

First, we will show that

$$\inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1, \ldots, A_n}} \sum_{i=1}^n \gamma_i K(A_i \mid x_i) \le \int_\mathcal{Y} \left(\min_{i \in [n]} \gamma_i \frac{dK}{d\mu}(\cdot \mid x_i)\right) d\mu.$$

Let $A_1^*, \ldots, A_n^*$ be the specific partition of $\mathcal{Y}$ given by

$$A_i^* \triangleq \left\{y \in \mathcal{Y} : \min\left(\arg\min_{j \in [n]} \gamma_j \frac{dK}{d\mu}(y \mid x_j)\right) = i\right\}$$

for each $i \in [n]$, where the min operator is added to break ties. We have

$$\begin{aligned} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1, \ldots, A_n}} \sum_{i=1}^n \gamma_i K(A_i \mid x_i) &\overset{(a)}{\le} \sum_{i=1}^n \gamma_i K(A_i^* \mid x_i) \overset{(b)}{=} \sum_{i=1}^n \int_{A_i^*} \left(\gamma_i \frac{dK}{d\mu}(\cdot \mid x_i)\right) d\mu \\ &\overset{(c)}{=} \sum_{i=1}^n \int_{A_i^*} \left(\min_{j \in [n]} \gamma_j \frac{dK}{d\mu}(\cdot \mid x_j)\right) d\mu \overset{(d)}{=} \int_\mathcal{Y} \left(\min_{i \in [n]} \gamma_i \frac{dK}{d\mu}(\cdot \mid x_i)\right) d\mu \end{aligned}$$

as desired, where (a) holds by upper-bounding the infimum with a particular instance, (b) holds by definition of Radon-Nikodym derivative, (c) holds by definition of A_i^*, and (d) holds because $A_1^*, \ldots, A_n^*$ is a partition of $\mathcal{Y}$.

Next, we will show that

$$\inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} \gamma_i K(A_i \mid x_i) \geq \int_{\mathcal{Y}} \left(\min_{i\in[n]} \gamma_i \frac{dK}{d\mu}(\cdot \mid x_i) \right) d\mu \,. \tag{35}$$

For any partition $A_1, \ldots, A_n$ of $\mathcal{Y}$, we have

$$\begin{aligned} \sum_{i=1}^{n} \gamma_i K(A_i \mid x_i) &\overset{(a)}{\geq} \sum_{i=1}^{n} \bigwedge_{j\in[n]} \gamma_j K(A_i \mid x_j) \overset{(b)}{=} \bigwedge_{j\in[n]} \gamma_j K(\mathcal{Y} \mid x_j) \overset{(c)}{=} \sup\left\{\nu(\mathcal{Y}) : \forall j \in [n],\, \gamma_j K(\cdot \mid x_j) \geq \nu\right\} \\ &\overset{(d)}{\geq} \int_{\mathcal{Y}} \left(\min_{i\in[n]} \gamma_i \frac{dK}{d\mu}(\cdot \mid x_i) \right) d\mu \,, \end{aligned}$$

where (a) holds because the greatest common component of a collection of measures is a lower bound on each individual measure, (b) holds because $A_1, \ldots, A_n$ is a partition of $\mathcal{Y}$, (c) holds by definition of greatest common component (3), and (d) holds by lower-bounding the supremum with the specific measure

$$\forall A \in \mathcal{F}_{\mathcal{Y}}, \;\; \nu^*(A) \triangleq \int_A \left(\min_{i\in[n]} \gamma_i \frac{dK}{d\mu}(\cdot \mid x_i) \right) d\mu \,,$$

which satisfies $\gamma_j K(\cdot \mid x_j) \geq \nu^*$ for each $j \in [n]$. Lastly, since $A_1, \ldots, A_n$ was arbitrary, taking the infimum over all n-partitions of $\mathcal{Y}$ proves (35) as desired. □

Next, we present some preliminaries before carrying out the proof in the general setting of Polish spaces.

Definition 2 (Lattice Infimum [69, Section 2, p. 253])**.** *Let $\mathcal{X}$ be a Polish domain, and let $\mathcal{F}$ be the set of all bounded measurable functions $f : \mathcal{X} \to \mathbb{R}$. Given a measure $\mu : \mathcal{X} \to \mathbb{R}_+$, the lattice infimum of an arbitrary subset of functions $\mathcal{G} \subseteq \mathcal{F}$, denoted $\tilde{\bigwedge}_{g\in\mathcal{G}} g$, is the function in $\mathcal{F}$ such that:*

1) For each $h \in \mathcal{G}$,

$$\tilde{\bigwedge_{g\in\mathcal{G}}} g \leq h \quad (\mu\text{-a.e.}) \tag{36}$$

2) For each $f \in \mathcal{F}$, if

$$\forall h \in \mathcal{G}, \; \Big(f \leq h \quad (\mu\text{-a.e.}) \Big) \,, \tag{37}$$

then

$$f \leq \tilde{\bigwedge_{g\in\mathcal{G}}} g \quad (\mu\text{-a.e.}) \tag{38}$$

For any $\mathcal{G}$, the lattice infimum is μ-a.e. unique. To see this, consider any two functions $f_1, f_2 \in \mathcal{F}$ satisfying Definition 2. Since both functions satisfy the first condition (36), both functions satisfy the antecedent in the second condition (37), and so by the consequent in the second condition (38), we have $f_1 \leq f_2$ (μ-a.e.) and $f_2 \leq f_1$ (μ-a.e.) Moreover, if $\mathcal{G}$ is countable, the lattice infimum is simply the pointwise infimum of functions in $\mathcal{G}$, i.e., $\tilde{\bigwedge}_{g\in\mathcal{G}} g(x) = \inf\{g(x) : g \in \mathcal{G}\}$. However, when $\mathcal{G}$ is uncountable, the lattice infimum and pointwise infimum are different in general. We remark that the notion of lattice infimum serves primarily to avoid measurability issues, and the proof of Theorem 1 can be carried out without such notions for "well-behaved" kernels as shown in Appendix B for completeness.

The following lemma shows that the density of the greatest common component of a collection of measures is the lattice infimum of the individual measures' densities.

Lemma 4 (Density of Greatest Common Component)**.** *Let $\{\pi_i\}_{i\in\mathcal{I}}$ be a collection of measures on a Polish space $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$, where $\mathcal{I}$ is an arbitrary index set. Assume each π_i is dominated by a common σ-finite measure μ (i.e., $\pi_i \ll \mu$ for all $i \in \mathcal{I}$). Then, we have the following properties:*

1) The greatest common component $\bigwedge_{i\in\mathcal{I}} \pi_i$ is dominated by μ, i.e., $\bigwedge_{i\in\mathcal{I}} \pi_i \ll \mu$.

2) The lattice infimum $\tilde{\bigwedge}_{i\in\mathcal{I}} \frac{d\pi_i}{d\mu}$ exists.

3) The density of the greatest common component is $\frac{d}{d\mu} \bigwedge_{i\in\mathcal{I}} \pi_i = \tilde{\bigwedge}_{i\in\mathcal{I}} \frac{d\pi_i}{d\mu}$ (μ-a.e.).

Proof of Lemma 4. Fix measures $\{\pi_i\}_{i\in\mathcal{I}}$ on $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ satisfying $\pi_i \ll \mu$ for each $i \in \mathcal{I}$.

Part 1: For any set $A \in \mathcal{F}_{\mathcal{Y}}$ with $\mu(A) = 0$, we have

$$\bigwedge_{i\in\mathcal{I}} \pi_i(A) \overset{(a)}{\leq} \pi_{i^*}(A) \overset{(b)}{=} 0$$

as desired, where (a) holds for any $i^* \in \mathcal{I}$ because the greatest common component of a collection of measures is a lower bound on each measure, and (b) holds because $\pi_{i^*} \ll \mu$.

Part 2: We refer readers to the analogous argument in [69, Lemma 2.6].

Part 3: For notational convenience, define a measure $\nu^* : \mathcal{F}_\mathcal{Y} \to \mathbb{R}_+$ as

$$\forall A \in \mathcal{F}_\mathcal{Y}, \;\; \nu^*(A) \triangleq \int_A \left(\tilde{\bigwedge_{i \in \mathcal{I}}} \frac{d\pi_i}{d\mu} \right) d\mu \,. \tag{39}$$

First, observe that for any $i \in \mathcal{I}$ and $A \in \mathcal{F}_\mathcal{Y}$, we have

$$\nu^*(A) \overset{(a)}{\leq} \int_A \frac{d\pi_i}{d\mu} \, d\mu \overset{(b)}{=} \pi_i(A) \,,$$

where (a) holds by Definition 2, Part 1 and (b) holds by definition of Radon-Nikodym derivative. Since $\nu^* \leq \pi_i$ for each $i \in \mathcal{I}$, it follows that for each $A \in \mathcal{F}_\mathcal{Y}$,

$$\nu^*(A) \leq \sup \{ \nu(A) : \forall i \in \mathcal{I}, \, \nu \leq \pi_i \} \,. \tag{40}$$

Next, observe that any measure ν satisfying $\nu \leq \pi_i$ for each $i \in \mathcal{I}$ is itself dominated by μ (i.e., $\nu \ll \mu$), because

$$\nu(A) \leq \pi_i(A) \overset{(a)}{=} 0$$

for any set $A \in \mathcal{F}_\mathcal{Y}$ with $\mu(A) = 0$, where (a) holds because $\pi_i \ll \mu$. Hence, by the Radon-Nikodym theorem, $\frac{d\nu}{d\mu}$ is well-defined. Moreover, we have

$$\frac{d\nu}{d\mu} \leq \frac{d\pi_i}{d\mu} \quad (\mu\text{-a.e.}) \tag{41}$$

for each $i \in \mathcal{I}$. To see this, suppose the contrary: Assume that for some $i^* \in \mathcal{I}$, we have $\frac{d\nu}{d\mu} > \frac{d\pi_{i^*}}{d\mu}$ on some set $A^* \in \mathcal{F}_\mathcal{Y}$ with $\mu(A^*) > 0$. Then,

$$\nu(A^*) = \int_{A^*} \frac{d\nu}{d\mu} \, d\mu > \int_{A^*} \frac{d\pi_{i^*}}{d\mu} \, d\mu = \pi_{i^*}(A^*) \,,$$

which contradicts the fact that $\nu \leq \pi_{i^*}$. Following from (41), we have

$$\frac{d\nu}{d\mu} \leq \tilde{\bigwedge_{i \in \mathcal{I}}} \frac{d\pi_i}{d\mu} \quad (\mu\text{-a.e.})$$

by Definition 2, Part 2, and therefore $\nu \leq \nu^*$ by (39). Since ν was arbitrary, it follows that for each $A \in \mathcal{F}_\mathcal{Y}$,

$$\nu^*(A) \geq \sup \{ \nu(A) : \forall i \in \mathcal{I}, \, \nu \leq \pi_i \} \,. \tag{42}$$

Proceeding onwards, for each $A \in \mathcal{F}_\mathcal{Y}$, we have

$$\int_A \left(\tilde{\bigwedge_{i \in \mathcal{I}}} \frac{d\pi_i}{d\mu} \right) d\mu \overset{(a)}{=} \sup \{ \nu(A) : \forall i \in \mathcal{I}, \, \nu \leq \pi_i \} \overset{(b)}{=} \bigwedge_{i \in \mathcal{I}} \pi_i(A) \,,$$

where (a) holds by combining (39), (40), and (42) and (b) holds by (3). Hence, by definition of Radon-Nikodym derivative, $\frac{d}{d\mu} \bigwedge_{i \in \mathcal{I}} \pi_i = \tilde{\bigwedge}_{i \in \mathcal{I}} \frac{d\pi_i}{d\mu}$ as desired. □

Finally, we prove Theorem 1.

Proof of Theorem 1. Fix an absolutely continuous Markov kernel $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to [0, 1]$ with common dominating measure $\mu : \mathcal{F}_\mathcal{Y} \to \mathbb{R}_+$. We have

$$\tau(K) \overset{(a)}{=} \bigwedge_{x \in \mathcal{X}} K(\mathcal{Y} \mid x) \overset{(b)}{=} \int_\mathcal{Y} \left(\frac{d}{d\mu} \bigwedge_{x \in \mathcal{X}} K(\cdot \mid x) \right) d\mu \overset{(c)}{=} \int_\mathcal{Y} \left(\tilde{\bigwedge_{x \in \mathcal{X}}} \frac{dK}{d\mu}(\cdot \mid x) \right) d\mu \,, \tag{43}$$

where (a) holds by the greatest common component characterization of Doeblin coefficient (5), (b) holds by definition of Radon-Nikodym derivative, and (c) holds by Lemma 4, Part 3.

Next, we compute the lattice infimum $\tilde{\bigwedge}_{x \in \mathcal{X}} \frac{dK}{d\mu}(\cdot \mid x)$. For notational convenience, let

$$① \triangleq \inf_{n \in \mathbb{N}} \inf_{x_1, \ldots, x_n \in \mathcal{X}} \int_\mathcal{Y} \left(\min_{i \in [n]} \frac{dK}{d\mu}(\cdot \mid x_i) \right) d\mu \,. \tag{44}$$

For each $k \in \mathbb{N}$, let $s_k \triangleq \{\hat{x}_{k,1}, \ldots, \hat{x}_{k,n_k}\} \subseteq \mathcal{X}$ be a finite sequence such that

$$\int_\mathcal{Y} \left(\min_{i \in [n_k]} \frac{dK}{d\mu}(\cdot \mid \hat{x}_{k,i}) \right) d\mu \leq ① + \frac{1}{k} \,,$$

where the existence of such a sequence is guaranteed by the definition of ① as an infimum in (44). Construct an infinite sequence $\{\tilde{x}_1, \tilde{x}_2, \ldots\} \subseteq \mathcal{X}$ by concatenating the finite sequences s_k in ascending order of k. Define a sequence of functions $g_n : \mathcal{Y} \to \mathbb{R}$ as

$$\forall n \in \mathbb{N}, \quad g_n(y) \triangleq \min_{i \in [n]} \frac{dK}{d\mu}(y \mid \tilde{x}_i)\,. \tag{45}$$

By construction, it holds that

$$\lim_{n\to\infty} \int_{\mathcal{Y}} g_n \, d\mu = ①\,. \tag{46}$$

Moreover, each g_n is non-negative, by the non-negativity of the Radon-Nikodym derivative; and $\{g_n\}_{n\in\mathbb{N}}$ is a non-increasing sequence, since each successive g_n is the minimum over a larger set of i. Hence, $\{g_n\}_{n\in\mathbb{N}}$ converges μ-a.e. to a measurable function $g : \mathcal{Y} \to \mathbb{R}$, i.e.,

$$\lim_{n\to\infty} g_n = g \quad (\mu\text{-a.e.}) \tag{47}$$

The limiting function g satisfies

$$\int_{\mathcal{Y}} g \, d\mu \overset{(a)}{=} \lim_{n\to\infty} \int_{\mathcal{Y}} g_n \, d\mu \overset{(b)}{=} ①\,, \tag{48}$$

where (a) holds by the dominated convergence theorem because $g_n \le g_1$ for all $n \in \mathbb{N}$ and $\int_{\mathcal{Y}} g_1 \, d\mu = K(\mathcal{Y} \mid \tilde{x}_1) = 1 < \infty$, and (b) holds by (46).

Observe that for every $x \in \mathcal{X}$, we have

$$g \le \frac{dK}{d\mu}(\cdot \mid x) \qquad (\mu\text{-a.e.})$$

To see this, suppose the contrary: Assume that $g > \frac{dK}{d\mu}(\cdot \mid x^*)$ for some $x^* \in \mathcal{X}$, on some set $A^* \in \mathcal{F}_{\mathcal{Y}}$ with positive measure $\mu(A^*) > 0$. Then, for any arbitrary $\epsilon > 0$, it follows that

$$\begin{aligned}
① &\overset{(a)}{=} \int_{A^*} g \, d\mu + \int_{A^{*\mathrm{c}}} g \, d\mu \overset{(b)}{>} \underbrace{\int_{A^*} \frac{dK}{d\mu}(\cdot \mid x^*) \, d\mu + \int_{A^{*\mathrm{c}}} g \, d\mu}_{②} \overset{(c)}{=} \int_{A^*} \frac{dK}{d\mu}(\cdot \mid x^*) \, d\mu + \int_{A^{*\mathrm{c}}} \left(\lim_{n\to\infty} \min_{i\in[n]} \frac{dK}{d\mu}(\cdot \mid \tilde{x}_i) \right) d\mu \\
&\overset{(d)}{\ge} \int_{\mathcal{Y}} \left(\lim_{n\to\infty} \min_{i\in[n]} \frac{dK}{d\mu}(\cdot \mid x_i^*) \right) d\mu \overset{(e)}{=} \lim_{n\to\infty} \int_{\mathcal{Y}} \left(\min_{i\in[n]} \frac{dK}{d\mu}(\cdot \mid x_i^*) \right) d\mu \overset{(f)}{\ge} \int_{\mathcal{Y}} \left(\min_{i\in[n^*]} \frac{dK}{d\mu}(\cdot \mid x_i^*) \right) d\mu - \epsilon \overset{(g)}{\ge} ① - \epsilon
\end{aligned}$$

and we obtain the contradiction ① > ② ≥ ①, where (a) holds by (48) and because $\mathcal{Y} = A^* \cup A^{*\mathrm{c}}$, (b) holds by the supposition, (c) holds by (45) and (47), (d) holds for the sequence x_i^* given by $x_1^* = x^*$ and $x_i^* = \tilde{x}_{i-1}$ for all $i \ge 2$, (e) holds by the dominated convergence theorem, (f) holds for some n^* (possibly depending on ϵ) by definition of limit, and (g) holds by lower-bounding the value for the specific sequence $\{x_1^*, \ldots, x_{n^*}^*\}$ with the infimum over all finite sequences.

Similarly, observe that any function $h : \mathcal{Y} \to \mathbb{R}$ where $h \le \frac{dK}{d\mu}(\cdot \mid x)$ μ-a.e. for every $x \in \mathcal{X}$ satisfies $h \le g$ μ-a.e. To see this, suppose the contrary: Assume that $h \le \frac{dK}{d\mu}(\cdot \mid x)$ μ-a.e. for every $x \in \mathcal{X}$, and $h > g$ on some set $A^* \in \mathcal{F}_{\mathcal{Y}}$ with positive measure $\mu(A^*) > 0$. Then, for any arbitrary $\epsilon > 0$, it follows that

$$\begin{aligned}
① &\overset{(a)}{=} \int_{A^*} g \, d\mu + \int_{A^{*\mathrm{c}}} g \, d\mu \overset{(b)}{<} \underbrace{\int_{A^*} h \, d\mu + \int_{A^{*\mathrm{c}}} g \, d\mu}_{③} \overset{(c)}{\le} \int_{A^*} \left(\min_{i\in[n^*]} \frac{dK}{d\mu}(\cdot \mid \tilde{x}_i) \right) d\mu + \int_{A^{*\mathrm{c}}} g \, d\mu \\
&\overset{(d)}{\le} \lim_{n\to\infty} \int_{A^*} \left(\min_{i\in[n]} \frac{dK}{d\mu}(\cdot \mid \tilde{x}_i) \right) d\mu + \epsilon + \int_{A^{*\mathrm{c}}} g \, d\mu \overset{(e)}{=} \int_{A^*} \left(\lim_{n\to\infty} \min_{i\in[n]} \frac{dK}{d\mu}(\cdot \mid \tilde{x}_i) \right) d\mu + \epsilon + \int_{A^{*\mathrm{c}}} g \, d\mu \\
&\overset{(f)}{=} \int_{A^*} g \, d\mu + \epsilon + \int_{A^{*\mathrm{c}}} g \, d\mu \overset{(g)}{=} ① + \epsilon
\end{aligned}$$

and we obtain the contradiction ① < ③ ≤ ①, where (a) holds by (48) and because $\mathcal{Y} = A^* \cup A^{*\mathrm{c}}$, (b) and (c) hold by the supposition, (c) holds for *any* finite $n^* \in \mathbb{N}$ because the finite union of measure-zero sets has measure zero, there always exists some n^* (possibly depending on ϵ) to make (d) hold by definition of limit, (e) holds by the dominated convergence theorem, (f) holds by (45) and (47), and (g) holds by (48) and because $\mathcal{Y} = A^* \cup A^{*\mathrm{c}}$.

Hence, by definition of lattice infimum (Definition 2), $g = \tilde{\bigwedge}_{x\in\mathcal{X}} \frac{dK}{d\mu}(\cdot \mid x)$. Proceeding from (43), we have

$$\tau(K) = \int_{\mathcal{Y}} g \, d\mu \overset{(a)}{=} ① \overset{(b)}{=} \inf_{n\in\mathbb{N}} \inf_{x_1,\ldots,x_n \in \mathcal{X}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} K(A_i \mid x_i)$$

as desired, where (a) holds by (48) and (b) holds by (44) and Proposition 2. □

B. Proofs of Doeblin Curve Properties

In this subsection, we first prove Proposition 3. We begin by establishing two technical lemmata used in the proofs of Proposition 3 and subsequent results.

Lemma 5 (Doeblin Coefficients Under Affine Transformation)**.** *Let $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ and $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ be Markov kernels. For any probability measures $\pi, \mu : \mathcal{F}_{\mathcal{X}} \to [0,1]$ and scalars $\lambda, \alpha \geq 0$ such that $W - \alpha\pi \geq 0$ and $1 - \lambda\bar{\alpha} \geq 0$ (where we define $\bar{\alpha} \triangleq 1 - \alpha$ for convenience), consider the Markov kernel $\hat{W} : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ given by $\hat{W} = \lambda\,(W - \alpha\pi) + (1 - \lambda\bar{\alpha})\,\mu$. Then, the complementary Doeblin coefficients of $\hat{W}$ and $\hat{W}K$ are $\rho(\hat{W}) = \lambda\,\rho(W)$ and $\rho(\hat{W}K) = \lambda\,\rho(WK)$.*

Proof of Lemma 5. Fix Markov kernels W and K, probability measures π and μ, and scalars λ and α satisfying the conditions in Lemma 5. The complementary Doeblin coefficient of $\hat{W}$ is

$$\begin{aligned}\rho(\hat{W}) &\overset{(a)}{=} 1 - \bigwedge_{u \in \mathcal{U}} \hat{W}(\mathcal{X} \mid u) \overset{(b)}{=} 1 - \left(\lambda \left(\bigwedge_{u \in \mathcal{U}} W(\mathcal{X} \mid u) - \alpha\pi(\mathcal{X})\right) + (1 - \lambda\bar{\alpha})\,\mu(\mathcal{X})\right) \\ &\overset{(c)}{=} 1 - \left(\lambda\,(\tau(W) - \alpha\pi(\mathcal{X})) + (1 - \lambda\bar{\alpha})\,\mu(\mathcal{X})\right) \overset{(d)}{=} 1 - \left(\lambda\,(\tau(W) - \alpha) + (1 - \lambda\bar{\alpha})\right) = \lambda\,\rho(W)\end{aligned}$$

as desired, where (a) and (c) hold by the greatest common component characterization of Doeblin coefficient (5), (b) holds by Lemma 8, and (d) holds because π and μ are probability measures. The composition of $\hat{W}$ with K is

$$\forall u \in \mathcal{U},\;\; \hat{W}K(\cdot \mid u) = \lambda \left(WK(\cdot \mid u) - \alpha\pi K(\cdot)\right) + (1 - \lambda\bar{\alpha})\,\mu K(\cdot)\,, \tag{49}$$

so the complementary Doeblin coefficient of $\hat{W}K$ is

$$\begin{aligned}\rho(\hat{W}K) &\overset{(a)}{=} 1 - \bigwedge_{u \in \mathcal{U}} \hat{W}K(\mathcal{Y} \mid u) \overset{(b)}{=} 1 - \left(\lambda \left(\bigwedge_{u \in \mathcal{U}} WK(\mathcal{Y} \mid u) - \alpha\pi K(\mathcal{Y})\right) + (1 - \lambda\bar{\alpha})\,\mu K(\mathcal{Y})\right) \\ &\overset{(c)}{=} 1 - \left(\lambda\,(\tau(WK) - \alpha\pi K(\mathcal{Y})) + (1 - \lambda\bar{\alpha})\,\mu K(\mathcal{Y})\right) \overset{(d)}{=} 1 - \left(\lambda\,(\tau(WK) - \alpha) + (1 - \lambda\bar{\alpha})\right) = \lambda\,\rho(WK)\end{aligned}$$

as desired, where (a) and (c) hold by the greatest common component characterization of Doeblin coefficient (5), (b) holds by Lemma 8 and (49), and (d) holds because πK and μK are probability measures. □

Lemma 6 (Properties of Identity Kernel)**.** *Let $(\mathcal{U}, \mathcal{F}_{\mathcal{U}})$ and $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ be Polish spaces with $\mathcal{U} \subseteq \mathcal{X}$. Let $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ be a Markov kernel. The identity kernel from $\mathcal{U}$ to $\mathcal{X}$, i.e., $\Xi : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ where $\Xi(A \mid u) = \delta_u(A)$ for all $u \in \mathcal{U}$ and $A \in \mathcal{F}_{\mathcal{X}}$, satisfies $\rho(\Xi) = 1$ and $\rho(\Xi K) = \rho_{\mathcal{U}}(K)$.*

Proof of Lemma 6. By inspection, the greatest common component of Ξ is $\bigwedge_{u \in \mathcal{U}} \Xi(\cdot \mid u) = 0$, and so the complementary Doeblin coefficient of Ξ, by (5), is $\rho(\Xi) = 1 - \bigwedge_{u \in \mathcal{U}} \Xi(\mathcal{X} \mid u) = 1$ as desired. By inspection, the composition of Ξ with K is $\Xi K(A \mid u) = K(A \mid u)$ for all $u \in \mathcal{U}$ and $A \in \mathcal{F}_{\mathcal{Y}}$, and so the complementary Doeblin coefficient of ΞK is $\rho(\Xi K) = 1 - \bigwedge_{u \in \mathcal{U}} K(\mathcal{Y} \mid u) = \rho_{\mathcal{U}}(K)$ as desired. □

Now, we are ready to prove Proposition 3.

Proof of Proposition 3.

Part 1: Fix Markov kernels K_1 and K_2 and constraint sets $\mathcal{G}_1$ and $\mathcal{G}_2$. We have

$$\mathrm{F}_{K_1K_2}(t; \mathcal{G}) \overset{(a)}{=} \sup\{\rho(WK_1K_2) \colon \rho(W) \leq t,\, W \in \mathcal{G}\} \overset{(b)}{\leq} \sup\{\rho(VK_2) \colon \rho(V) \leq \mathrm{F}_{K_1}(t; \mathcal{G}_1), V \in \mathcal{G}_2\} \overset{(c)}{=} \mathrm{F}_{K_2}(\mathrm{F}_{K_1}(t; \mathcal{G}_1); \mathcal{G}_2)$$

as desired, where (a) and (c) hold by definition of Doeblin curve (9), and (b) holds because any kernel $W \in \mathcal{G}$ with $\rho(W) \leq t$ satisfies $\rho(WK_1) \leq \sup\{\rho(WK_1) : \rho(W) \leq t,\, W \in \mathcal{G}_1\} = \mathrm{F}_{K_1}(t; \mathcal{G}_1)$ and $WK_1 \in \mathcal{G}_2$.

Part 2: Fix a Markov kernel K and a convex constraint set $\mathcal{G}$ containing the identity kernel $\Xi(\cdot \mid u) = \delta_u(\cdot)$ and a constant kernel $\bar{W}(\cdot \mid u) = \pi(\cdot)$. For $t = 0$, we trivially have $\mathrm{F}_K(t; \mathcal{G}) = 0 = \rho(K)\,t$, so fix $t \in (0,1]$ for the remainder of this part. Let $\hat{W} : \mathcal{X} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ be the Markov kernel given by $\hat{W}(\cdot \mid u) \triangleq t\,\Xi(\cdot \mid u) + (1 - t)\,\bar{W}(\cdot \mid u) = t\,\Xi(\cdot \mid u) + (1 - t)\,\pi(\cdot)$ for all $u \in \mathcal{X}$. The complementary Doeblin coefficients of $\hat{W}$ and $\hat{W}K$ are

$$\begin{aligned}\rho(\hat{W}) &\overset{(a)}{=} t\,\rho(\Xi) \overset{(b)}{=} t\,, \\ \rho(\hat{W}K) &\overset{(a)}{=} t\,\rho(\Xi K) \overset{(b)}{=} t\,\rho_{\mathcal{X}}(K) = t\,\rho(K)\,,\end{aligned}$$

where in each line (a) holds by Lemma 5 and (b) holds by Lemma 6. By convexity of $\mathcal{G}$, we have $\hat{W} \in \mathcal{G}$. Hence,

$$\mathrm{F}_K(t; \mathcal{G}) \overset{(a)}{=} \sup\{\rho(WK) : \rho(W) \leq t,\, W \in \mathcal{G}\} \geq \rho(\hat{W}K) = t\,\rho(K)\,,$$

where (a) holds by definition of Doeblin curve (9). Combining this lower bound with (10) completes the proof.

Part 3: Fix a Markov kernel K, a convex constraint set $\mathcal{G}$ containing a constant kernel $\bar{W}(\cdot \mid u) = \pi(\cdot)$, and scalars $\lambda \in [0,1]$ and $t \in [0,1]$. Fix an arbitrary $\epsilon > 0$. By definition of Doeblin curve as a supremum (9), there exists a Markov kernel $W^* \in \mathcal{G}$ such that $\rho(W^*) \le t$ and $\rho(W^*K) \ge \mathrm{F}_K(t;\mathcal{G}) - \epsilon$. Define another Markov kernel $\hat{W}$ as $\hat{W}(\cdot \mid u) \triangleq \lambda W^*(\cdot \mid u) + (1-\lambda)\bar{W}(\cdot \mid u) = \lambda W^*(\cdot \mid u) + (1-\lambda)\pi(\cdot)$ for all $u \in \mathcal{U}$. By Lemma 5, we have $\rho(\hat{W}) = \lambda\rho(W^*) \le \lambda t$ and $\rho(\hat{W}K) = \lambda\rho(W^*K) \ge \lambda(\mathrm{F}_K(t;\mathcal{G}) - \epsilon)$. By convexity of $\mathcal{G}$, we have $\hat{W} \in \mathcal{G}$. Hence,

$$\mathrm{F}_K(\lambda t;\mathcal{G}) \overset{(a)}{=} \sup\{\rho(WK) : \rho(W) \le \lambda t,\, W \in \mathcal{G}\} \ge \rho(\hat{W}K) \ge \lambda(\mathrm{F}_K(t;\mathcal{G}) - \epsilon)\,, \tag{50}$$

where (a) holds by definition of Doeblin curve (9). Since $\epsilon > 0$ was arbitrary, we have $\mathrm{F}_K(\lambda t;\mathcal{G}) \ge \lambda \mathrm{F}_K(t;\mathcal{G})$ as desired.

Finally, to show the equivalent characterization that $t \mapsto \mathrm{F}_K(t;\mathcal{G})/t$ is non-increasing, fix $s,t \in (0,1]$ such that $s < t$ and observe that

$$\frac{\mathrm{F}_K(s;\mathcal{G})}{s} \overset{(a)}{\ge} \left(\frac{s}{t}\right)\frac{\mathrm{F}_K(t;\mathcal{G})}{s} = \frac{\mathrm{F}_K(t;\mathcal{G})}{t}\,,$$

where (a) holds by applying (50) with $\lambda \triangleq s/t < 1$.

Part 4: By definition of Lipschitz continuity, we want to show that $|\mathrm{F}_K(s;\mathcal{G}) - \mathrm{F}_K(t;\mathcal{G})| \le |s-t|$ for all $s,t \in [0,1]$. Fix $s,t \in [0,1]$ such that $s < t$. If $s = 0$, the result trivially follows from (11). Otherwise, we have

$$|\mathrm{F}_K(s;\mathcal{G}) - \mathrm{F}_K(t;\mathcal{G})| \overset{(a)}{=} \frac{t}{s}\left(\frac{s}{t}\right)\mathrm{F}_K(t;\mathcal{G}) - \mathrm{F}_K(s;\mathcal{G}) \overset{(b)}{\le} \frac{t}{s}\mathrm{F}_K(s;\mathcal{G}) - \mathrm{F}_K(s;\mathcal{G}) = (t-s)\frac{\mathrm{F}_K(s;\mathcal{G})}{s} \overset{(c)}{\le} t - s \overset{(d)}{=} |s-t|$$

as desired, where (a) holds because $\mathrm{F}_K(s;\mathcal{G}) \le \mathrm{F}_K(t;\mathcal{G})$ by monotonicity of Doeblin curves, (b) holds by applying Part 3 with $\lambda \triangleq s/t < 1$, (c) holds because $\mathrm{F}_K(s) \le s$ by (11), and (d) holds because $s < t$. □

Next, we prove Proposition 4.

Proof of Proposition 4. Fix row stochastic matrices $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{2\times n}$ and $\mathbf{W} \in \mathbb{R}_{\mathsf{sto}}^{m\times 2}$. By definition of joint range, it suffices to show that $\rho(\mathbf{WK}) = \rho(\mathbf{W})\rho(\mathbf{K})$. We have

$$\begin{aligned}\rho(\mathbf{WK}) &\overset{(a)}{=} 1 - \sum_{j=1}^{n}\min_{i\in[m]}[\mathbf{WK}]_{i,j} \overset{(b)}{=} 1 - \sum_{j=1}^{n}\min_{i\in[m]}\Big\{[\mathbf{W}]_{i,1}[\mathbf{K}]_{1,j} + \Big(1 - [\mathbf{W}]_{i,1}\Big)[\mathbf{K}]_{2,j}\Big\}\\ &= 1 - \sum_{j=1}^{n}\min_{i\in[m]}\Big\{[\mathbf{W}]_{i,1}\Big([\mathbf{K}]_{1,j} - [\mathbf{K}]_{2,j}\Big)\Big\} - \sum_{j=1}^{n}[\mathbf{K}]_{2,j} \overset{(c)}{=} \sum_{j=1}^{n}\max_{i\in[m]}\Big\{[\mathbf{W}]_{i,1}\Big([\mathbf{K}]_{2,j} - [\mathbf{K}]_{1,j}\Big)\Big\}\,,\end{aligned} \tag{51}$$

where (a) holds by definition of complementary Doeblin coefficient (6), (b) holds because $\mathbf{W}$ is row stochastic, and (c) holds because $\mathbf{K}$ is row stochastic. Next, let $\mathcal{S} = \{j \in [n] : [\mathbf{K}]_{2,j} \ge [\mathbf{K}]_{1,j}\}$ akin to [61, Proposition 4.2]. Proceeding from (51), we have

$$\rho(\mathbf{WK}) = \sum_{j\in\mathcal{S}}\Big([\mathbf{K}]_{2,j} - [\mathbf{K}]_{1,j}\Big)\max_{i\in[m]}[\mathbf{W}]_{i,1} + \sum_{j\in\mathcal{S}^c}\Big([\mathbf{K}]_{2,j} - [\mathbf{K}]_{1,j}\Big)\min_{i\in[m]}[\mathbf{W}]_{i,1}\,. \tag{52}$$

Since $\mathbf{K}$ is row stochastic, it holds that

$$\sum_{j=1}^{n}[\mathbf{K}]_{1,j} \overset{(a)}{=} \sum_{j=1}^{n}[\mathbf{K}]_{2,j} = 1\,.$$

Rearranging (a), we obtain

$$\sum_{j\in\mathcal{S}}\Big([\mathbf{K}]_{2,j} - [\mathbf{K}]_{1,j}\Big) = \sum_{j\in\mathcal{S}^c}\Big([\mathbf{K}]_{1,j} - [\mathbf{K}]_{2,j}\Big)\,. \tag{53}$$

Combining (52) and (53),

$$\rho(\mathbf{WK}) = \underbrace{\left(\sum_{j\in\mathcal{S}}\Big([\mathbf{K}]_{2,j} - [\mathbf{K}]_{1,j}\Big)\right)}_{①}\underbrace{\left(\max_{i\in[m]}[\mathbf{W}]_{i,1} - \min_{i\in[m]}[\mathbf{W}]_{i,1}\right)}_{②}.$$

Next, we evaluate ①. We have

$$\begin{aligned}① &\overset{(a)}{=} \frac{1}{2}\left(\sum_{j\in\mathcal{S}}\Big([\mathbf{K}]_{2,j} - [\mathbf{K}]_{1,j}\Big) + \sum_{j\in\mathcal{S}^c}\Big([\mathbf{K}]_{1,j} - [\mathbf{K}]_{2,j}\Big)\right) \overset{(b)}{=} \frac{1}{2}\sum_{j=1}^{n}\left(\max_{\ell\in[2]}[\mathbf{K}]_{\ell,j} - \min_{\ell\in[2]}[\mathbf{K}]_{\ell,j}\right)\\ &= \frac{1}{2}\sum_{j=1}^{n}\left(\sum_{\ell=1}^{2}[\mathbf{K}]_{\ell,j} - 2\min_{\ell\in[2]}[\mathbf{K}]_{\ell,j}\right) \overset{(c)}{=} 1 - \sum_{j=1}^{n}\min_{\ell\in[2]}[\mathbf{K}]_{\ell,j} \overset{(d)}{=} \rho(\mathbf{K})\,,\end{aligned}$$

where (a) holds by (53), (b) holds by definition of $\mathcal{S}$, (c) holds because $\mathbf{K}$ is row stochastic, and (d) holds by definition of complementary Doeblin coefficient (6). Next, we evaluate ②. We have

$$② \overset{(a)}{=} \left(1 - \min_{i\in[m]} [\mathbf{W}]_{i,2}\right) - \min_{i\in[m]} [\mathbf{W}]_{i,1} = 1 - \sum_{\ell=1}^{2} \min_{i\in[m]} [\mathbf{W}]_{i,\ell} \overset{(b)}{=} \rho(\mathbf{W}),$$

where (a) holds because $\mathbf{W}$ is row stochastic and (b) holds by definition of complementary Doeblin coefficient (6). Hence, $\rho(\mathbf{WK}) = \rho(\mathbf{W})\rho(\mathbf{K})$ as desired. □

Next, we prove Proposition 5. We begin by establishing an additional technical lemma used in the proof of Proposition 5.

Lemma 7 (Power Levels Under Affine Transformation)**.** *Let $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ be a Markov kernel. Let $\mathbf{0} \in \mathcal{X}$ denote the element in the Banach space $\mathcal{X}$ such that $\|\mathbf{0}\| = 0$. For any probability measure $\pi : \mathcal{F}_{\mathcal{X}} \to [0,1]$ and scalars $\lambda, \alpha \geq 0$ such that $W - \alpha\pi \geq 0$ and $1 - \lambda\bar{\alpha} \geq 0$ (where we define $\bar{\alpha} \triangleq 1 - \alpha$ for convenience), the Markov kernel*

$$\hat{W} = \lambda\,(W - \alpha\pi) + (1 - \lambda\bar{\alpha})\,\delta_{\mathbf{0}}$$

has power levels

$$|||\hat{W}|||_{\mathsf{UA}} \leq \lambda\,|||W|||_{\mathsf{UA}}\,, \tag{54}$$

$$|||\hat{W}|||_{\mathsf{AE}} \leq \lambda\,|||W|||_{\mathsf{AE}}\,, \tag{55}$$

where (55) holds if average extremal power is well-defined for W. Furthermore, if $\alpha = 0$, the bounds in (54) and (55) hold with equality.

Proof of Lemma 7. Fix a Markov kernel W, a probability measure π, and scalars λ and α satisfying the conditions in Lemma 7.

Firstly, the uniform average power of $\hat{W}$ is

$$\begin{aligned}|||\hat{W}|||_{\mathsf{UA}} &\overset{(a)}{=} \lambda \sup_{u\in\mathcal{U}} \int_{\mathcal{X}} M(\|x\|)\left(W(dx \mid u) - \alpha\pi(dx)\right) + (1-\lambda\bar{\alpha}) \int_{\mathcal{X}} M(\|x\|)\,\delta_{\mathbf{0}}(dx) \\ &\overset{(b)}{=} \lambda \sup_{u\in\mathcal{U}} \int_{\mathcal{X}} M(\|x\|)\left(W(dx \mid u) - \alpha\pi(dx)\right) \leq \lambda \sup_{u\in\mathcal{U}} \int_{\mathcal{X}} M(\|x\|)\,W(dx \mid u) \overset{(c)}{=} \lambda\,|||W|||_{\mathsf{UA}}\end{aligned} \tag{56}$$

as desired, where (a) holds by definition of uniform average power (12) and linearity of Lebesgue integration with respect to the measure, (b) holds because $M(\|\mathbf{0}\|) = M(0) = 0$, and (c) holds by definition of uniform average power (12).

Secondly, to compute the average extremal power of W, we consider the maximal coupling of random variables $\{X_u\}_{u\in\mathcal{U}}$ with $X_u \sim W(\cdot \mid u)$ for each $u \in \mathcal{U}$, defined in (34) as

$$\forall u \in \mathcal{U},\;\; X_u \triangleq \begin{cases} X^*, & \text{if } I = 1\,, \\ \tilde{X}_u\,, & \text{if } I = 0\,, \end{cases} \tag{57}$$

where the random variables I, X^*, and $\{\tilde{X}_u\}_{u\in\mathcal{U}}$ are sampled independently from the probability measures

$$I \sim \mathsf{Bernoulli}(\tau(W))\,, \tag{58}$$

$$X^* \sim \frac{\bigwedge_{u\in\mathcal{U}} W(\cdot \mid u)}{\tau(W)}\,, \tag{59}$$

$$\forall u \in \mathcal{U},\;\; \tilde{X}_u \sim \frac{W(\cdot \mid u) - \bigwedge_{u'\in\mathcal{U}} W(\cdot \mid u')}{\rho(W)}\,. \tag{60}$$

For notational convenience, let ν be the product measure

$$\nu \triangleq \bigotimes_{u\in\mathcal{U}} \frac{W(\cdot \mid u) - \bigwedge_{u'\in\mathcal{U}} W(\cdot \mid u')}{\rho(W)}\,.$$

The average extremal power of W is

$$\begin{aligned}|||W|||_{\mathsf{AE}} &\overset{(a)}{=} \mathbb{P}(I=1)\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, I = 1\right] + \mathbb{P}(I=0)\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, I = 0\right] \\ &\overset{(b)}{=} \tau(W)\,\mathbb{E}[M(\|X^*\|)] + \rho(W)\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} M\left(\|\tilde{X}_u\|\right)\right] \\ &\overset{(c)}{=} \tau(W) \int_{\mathcal{X}} M(\|x^*\|)\,\frac{\bigwedge_{u\in\mathcal{U}} W(dx^* \mid u)}{\tau(W)} + \rho(W) \int_{\mathcal{X}^{\mathcal{U}}} \left(\sup_{u\in\mathcal{U}} M(\|\tilde{x}_u\|)\right) d\nu\left(\{\tilde{x}_u\}_{u\in\mathcal{U}}\right) \\ &= \int_{\mathcal{X}} M(\|x^*\|) \bigwedge_{u\in\mathcal{U}} W(dx^* \mid u) + \rho(W) \int_{\mathcal{X}^{\mathcal{U}}} \left(\sup_{u\in\mathcal{U}} M(\|\tilde{x}_u\|)\right) d\nu\left(\{\tilde{x}_u\}_{u\in\mathcal{U}}\right),\end{aligned} \tag{61}$$

where all expectations are taken with respect to the maximal coupling (57), (a) holds by definition of average extremal power (13) and the law of total expectation, (b) holds by (57) and (58), and (c) holds by (59) and (60).

Similarly, to compute the average extremal power of $\hat{W}$, we consider the maximal coupling of random variables $\{X_u\}_{u\in\mathcal{U}}$ with $X_u \sim \hat{W}(\cdot \mid u)$ for each $u \in \mathcal{U}$, defined in (34) as

$$\forall u \in \mathcal{U},\ \ X_u \triangleq \begin{cases} X^*\,, & \text{if } I = 1\,, \\ \tilde{X}_u\,, & \text{if } I = 0\,, \end{cases} \tag{62}$$

where the random variables I, X^*, and $\{\tilde{X}_u\}_{u\in\mathcal{U}}$ are sampled independently from the probability measures

$$I \sim \mathsf{Bernoulli}\Big(\tau(\hat{W})\Big)\,, \tag{63}$$

$$X^* \sim \frac{\bigwedge_{u\in\mathcal{U}} \hat{W}(\cdot \mid u)}{\tau(\hat{W})}\,, \tag{64}$$

$$\forall u \in \mathcal{U},\ \ \tilde{X}_u \sim \frac{\hat{W}(\cdot \mid u) - \bigwedge_{u'\in\mathcal{U}} \hat{W}(\cdot \mid u')}{\rho(\hat{W})}\,. \tag{65}$$

By Lemmas 5 and 8 respectively, we have

$$\rho(\hat{W}) = \lambda\, \rho(W)\,, \tag{66}$$

$$\bigwedge_{u\in\mathcal{U}} \hat{W}(\cdot \mid u) = \lambda \Big(\bigwedge_{u\in\mathcal{U}} W(\cdot \mid u) - \alpha\pi(\cdot) \Big) + (1 - \lambda\bar{\alpha})\, \delta_{\mathbf{0}}(\cdot)\,, \tag{67}$$

and so,

$$\bigotimes_{u\in\mathcal{U}} \frac{\hat{W}(\cdot \mid u) - \bigwedge_{u'\in\mathcal{U}} \hat{W}(\cdot \mid u')}{\rho(\hat{W})} = \bigotimes_{u\in\mathcal{U}} \frac{\lambda \left(W(\cdot \mid u) - \bigwedge_{u'\in\mathcal{U}} W(\cdot \mid u')\right)}{\lambda\, \rho(W)} = \nu\,.$$

Thus, the average extremal power of $\hat{W}$ is

$$\begin{aligned} |||\hat{W}|||_{\mathsf{AE}} &\overset{(a)}{=} \mathbb{P}(I=1)\, \mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, I = 1 \right] + \mathbb{P}(I=0)\, \mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, I = 0 \right] \\ &\overset{(b)}{=} \tau(\hat{W})\, \mathbb{E}[M(\|X^*\|)] + \rho(\hat{W})\, \mathbb{E}\left[\sup_{u\in\mathcal{U}} M\Big(\|\tilde{X}_u\|\Big) \right] \\ &\overset{(c)}{=} \underbrace{\tau(\hat{W}) \int_{\mathcal{X}} M(\|x^*\|)\, \frac{\bigwedge_{u\in\mathcal{U}} \hat{W}(dx^* \mid u)}{\tau(\hat{W})}}_{①} + \underbrace{\rho(\hat{W}) \int_{\mathcal{X}^{\mathcal{U}}} \left(\sup_{u\in\mathcal{U}} M(\|\tilde{x}_u\|) \right) d\nu\big(\{\tilde{x}_u\}_{u\in\mathcal{U}}\big)}_{②}\,, \end{aligned} \tag{68}$$

where all expectations are taken with respect to the maximal coupling (62), (a) holds by definition of average extremal power (13) and the law of total expectation, (b) holds by (62) and (63), and (c) holds by (64) and (65). Next, we evaluate ①. We have

$$\begin{aligned} ① &\overset{(a)}{=} \lambda \int_{\mathcal{X}} M(\|x^*\|) \Big(\bigwedge_{u\in\mathcal{U}} W(dx^* \mid u) - \alpha\pi(dx^*) \Big) + (1 - \lambda\bar{\alpha}) \int_{\mathcal{X}} M(\|x^*\|)\, \delta_{\mathbf{0}}(dx^*) \\ &\overset{(b)}{=} \lambda \int_{\mathcal{X}} M(\|x^*\|) \Big(\bigwedge_{u\in\mathcal{U}} W(dx^* \mid u) - \alpha\pi(dx^*) \Big) \le \lambda \int_{\mathcal{X}} M(\|x^*\|) \bigwedge_{u\in\mathcal{U}} W(dx^* \mid u)\,, \end{aligned} \tag{69}$$

where (a) holds by (67) and linearity of Lebesgue integration with respect to the measure, and (b) holds because $M(\|\mathbf{0}\|) = M(0) = 0$. Next, we evaluate ②. By (66), we have

$$② = \lambda\, \rho(W) \int_{\mathcal{X}^{\mathcal{U}}} \left(\sup_{u\in\mathcal{U}} M(\|\tilde{x}_u\|) \right) d\nu\big(\{\tilde{x}_u\}_{u\in\mathcal{U}}\big)\,. \tag{70}$$

Combining (61) and (68) to (70) yields $|||\hat{W}|||_{\mathsf{AE}} \le \lambda\, |||W|||_{\mathsf{AE}}$ as desired.

Lastly, if $\alpha = 0$, the bounds in (56) and (69) hold with equality, as desired. □

Now, we are ready to prove Proposition 5.

Proof of Proposition 5. Fix a Markov kernel $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$. Fix scalars $t \in [0,1]$, $p \in [0,\infty)$, and $\lambda > 0$ such that $\lambda t \le 1$.[9] Throughout this proof, let F_K and $|||\cdot|||$ denote either $\mathrm{F}_K^{\mathsf{UA}}$ and $|||\cdot|||_{\mathsf{UA}}$, or $\mathrm{F}_K^{\mathsf{AE}}$ and $|||\cdot|||_{\mathsf{AE}}$, in general. We will consider two cases.

Case 1: $\lambda \le 1$. We follow the argument from [25, Proposition 2]. First, we will show that

$$\mathrm{F}_K(\lambda t; \lambda p) \ge \lambda \mathrm{F}_K(t; p)\,. \tag{71}$$

[9]The proposition trivially holds for $\lambda = 0$, so we assume $\lambda > 0$ throughout this proof.

Fix any arbitrary $\epsilon > 0$. By definition of Doeblin curve, there exists a Markov kernel $W^* : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ such that $\rho(W^*) \le t$, $|||W^*||| \le p$, and $\rho(W^*K) \ge \mathrm{F}_K(t;p) - \epsilon$. Since $\mathcal{X}$ is a Banach space, there exists an element $\mathbf{0} \in \mathcal{X}$ such that $\|\mathbf{0}\| = 0$; we denote this element in bold font to distinguish it from the scalar $0 \in \mathbb{R}$. Define another Markov kernel $\hat{W} : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ as $\hat{W}(\cdot \mid u) = \lambda W^*(\cdot \mid u) + (1-\lambda)\,\delta_{\mathbf{0}}(\cdot)$ for all $u \in \mathcal{U}$. By Lemmas 5 and 7, we have

$$\rho(\hat{W}) = \lambda\,\rho(W^*) \le \lambda t\,, \tag{72}$$

$$\rho(\hat{W}K) = \lambda\,\rho(W^*K) \ge \lambda\,(\mathrm{F}_K(t;p) - \epsilon)\,, \tag{73}$$

$$|||\hat{W}||| = \lambda\,|||W^*||| \le \lambda p\,. \tag{74}$$

Hence,

$$\mathrm{F}_K(\lambda t;\lambda p) \overset{(a)}{=} \sup\{\rho(WK) : \rho(W) \le \lambda t,\ |||W||| \le \lambda p\} \overset{(b)}{\ge} \rho(\hat{W}K) \overset{(c)}{\ge} \lambda\,(\mathrm{F}_K(t;p) - \epsilon)\,,$$

where (a) holds by definition of Doeblin curve, (b) holds by (72) and (74), and (c) holds by (73). Since $\epsilon > 0$ was arbitrary, this shows (71) as desired.

Next, we will show that

$$\lambda \mathrm{F}_K(t;p) \ge \mathrm{F}_K(\lambda t;\lambda p)\,. \tag{75}$$

Fix any arbitrary $\epsilon > 0$. By definition of Doeblin curve, there exists a Markov kernel $W^* : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ such that $\rho(W^*) \le \lambda t$, $|||W^*||| \le \lambda p$, and $\rho(W^*K) \ge \mathrm{F}_K(\lambda t;\lambda p) - \epsilon$. Let α be any scalar such that $1-\lambda \le \alpha \le \tau(W^*)$.[10] For notational convenience, define $\bar{\alpha} \triangleq 1-\alpha$. By definition of Doeblin coefficient, there exists a probability measure π^* such that $W^* \ge \alpha\pi^*$.[11] Define another Markov kernel $\hat{W} : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ as $\hat{W}(\cdot \mid u) = \frac{1}{\lambda}\big(W^*(\cdot \mid u) - \alpha\pi^*(\cdot)\big) + \big(1 - \frac{\bar{\alpha}}{\lambda}\big)\delta_{\mathbf{0}}(\cdot)$ for all $u \in \mathcal{U}$. This is a valid Markov kernel because $W^* \ge \alpha\pi^*$, non-negative[12] linear combinations of measures are measures, and for any $u \in \mathcal{U}$, $\hat{W}(\mathcal{X} \mid u) = \frac{1}{\lambda}\big(W^*(\mathcal{X} \mid u) - \alpha\pi^*(\mathcal{X})\big) + \big(1 - \frac{\bar{\alpha}}{\lambda}\big)\delta_{\mathbf{0}}(\mathcal{X}) = \frac{1}{\lambda}\big(1-\alpha\big) + \big(1 - \frac{\bar{\alpha}}{\lambda}\big) = 1$. By Lemmas 5 and 7, we have

$$\rho(\hat{W}) = \frac{\rho(W^*)}{\lambda} \le t\,, \tag{76}$$

$$\rho(\hat{W}K) = \frac{\rho(W^*K)}{\lambda} \ge \frac{\mathrm{F}_K(\lambda t;\lambda p) - \epsilon}{\lambda}\,, \tag{77}$$

$$|||\hat{W}||| = \frac{|||W^*|||}{\lambda} \le p\,. \tag{78}$$

Hence,

$$\lambda\mathrm{F}_K(t;p) \overset{(a)}{=} \lambda \sup\{\rho(WK) : \rho(W) \le t,\ |||W||| \le p\} \overset{(b)}{\ge} \lambda\,\rho(\hat{W}K) \overset{(c)}{\ge} \mathrm{F}_K(\lambda t;\lambda p) - \epsilon\,,$$

where (a) holds by definition of Doeblin curve, (b) holds by (76) and (78), and (c) holds by (77). Since $\epsilon > 0$ was arbitrary, this shows (75) as desired.

Finally, combining (71) and (75), we have $\mathrm{F}_K(\lambda t;\lambda p) = \lambda \mathrm{F}_K(t;p)$ as desired.

Case 2: $\lambda > 1$. We have

$$\mathrm{F}_K(\lambda t;\lambda p) = \lambda\left(\frac{1}{\lambda}\right)\mathrm{F}_K(\lambda t;\lambda p) \overset{(a)}{=} \lambda \mathrm{F}_K\left(\left(\frac{1}{\lambda}\right)\lambda t;\left(\frac{1}{\lambda}\right)\lambda p\right) = \lambda \mathrm{F}_K(t;p)$$

as desired, where (a) holds by applying Case 1 with $1/\lambda < 1$. □

Next, we prove Proposition 6.

Proof of Proposition 6. First, we define scalar multiplication of sets as $\lambda A \triangleq \{\lambda a : a \in A\}$ for any $\lambda \in \mathbb{R}$ and set A. Note that for all $\lambda > 0$, $A \in \mathcal{F}_{\mathbb{R}^d} \iff \lambda A \in \mathcal{F}_{\mathbb{R}^d}$. Next, note that for all $A \in \mathcal{F}_{\mathcal{Y}}$ and $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$,

$$K_\sigma(\sigma A \mid \sigma\mathbf{x}) = \int_{\sigma A} \frac{1}{\sigma^d} f\left(\frac{\mathbf{y} - \sigma\mathbf{x}}{\sigma}\right) d\mathbf{y} = \int_A f(\mathbf{y}' - \mathbf{x}) d\mathbf{y}' = K_1(A \mid \mathbf{x}). \tag{79}$$

Given any $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$, define W_σ such that $W_\sigma(A \mid u) \triangleq W(A/\sigma \mid u)$ (i.e., $W_\sigma(\cdot \mid u)$ is the pushforward measure of $W(\cdot \mid u)$ by $\mathbf{g}(\mathbf{x}) = \sigma\mathbf{x}$). Then,

$$|||W_\sigma|||_{\mathsf{UA}} = \sup_{u\in\mathcal{U}} \int_{\mathcal{X}} \|\mathbf{x}\|^2\, W_\sigma(d\mathbf{x} \mid u) = \sup_{u\in\mathcal{U}} \int_{\mathcal{X}} \|\mathbf{x}\|^2\, W(d\mathbf{x}/\sigma \mid u) = \sup_{u\in\mathcal{U}} \int_{\mathcal{X}} \|\sigma\mathbf{x}'\|^2\, W(d\mathbf{x}' \mid u) = \sigma^2 |||W|||_{\mathsf{UA}}.$$

In addition,

$$\rho(W_\sigma) = 1 - \sup\{\alpha \in \mathbb{R} : \exists \pi \in \mathscr{P}, \forall u \in \mathcal{U}, \forall A \in \mathcal{F}_{\mathcal{X}}, W_\sigma(A \mid u) \ge \alpha\pi(A)\}$$

[10] Such an α exists because $t \le 1$ and $\lambda > 0$, and so $1-\lambda \le 1-\lambda t \le 1 - \rho(W^*) = \tau(W^*)$.

[11] This is true even for $\alpha = \tau(W^*)$ because the supremum in (4) is always achieved, as per the discussion following (4).

[12] We have $1 - \bar{\alpha}/\lambda \ge 0$ because $1 - \lambda \le \alpha$.

$$= 1 - \sup\{\alpha \in \mathbb{R} : \exists \pi \in \mathscr{P}, \forall u \in \mathcal{U}, \forall A' \in \mathcal{F}_{\mathcal{X}}, W(A' \mid u) \geq \alpha \pi(\sigma A')\}$$
$$= 1 - \sup\{\alpha \in \mathbb{R} : \exists \pi' \in \mathscr{P}, \forall u \in \mathcal{U}, \forall A' \in \mathcal{F}_{\mathcal{X}}, W(A' \mid u) \geq \alpha \pi'(A')\} = \rho(W),$$

and similarly $\rho(K_\sigma W_\sigma) = \rho(K_W)$ since $(K_\sigma W_\sigma)(\sigma A \mid u) = (K_1 W)(A \mid u)$ due to (79) and the definition of W_σ. Thus, $\mathrm{F}_{K_\sigma}^{\mathsf{UA}}(t; \sigma^2 p) = \mathrm{F}_{K_1}^{\mathsf{UA}}(t; p)$ for all $p \geq 0$, completing the proof. □

C. Proofs of Bounds on Power-Constrained Doeblin Curves

First, we prove Theorem 2.

Proof of Theorem 2. Fix an absolutely continuous Markov kernel $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ with common dominating measure $\mu : \mathcal{F}_{\mathcal{Y}} \to \mathbb{R}_+$.

Upper bound: Fix $t \in (0,1]$ and $p \in [0, \infty)$. By definition of Doeblin curve, we want to show that for any Markov kernel $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ satisfying $\rho(W) \leq t$ and $|||W|||_{\mathsf{AE}} \leq p$, we have

$$\rho(WK) \leq t\, \hat{\Theta}\Big(2\gamma\, M^{-1}\Big(\frac{p}{t}\Big)\Big)\,. \tag{80}$$

Fix a Polish space $(\mathcal{U}, \mathcal{F}_{\mathcal{U}})$ and such a Markov kernel W. We have

$$\rho(WK) \overset{(a)}{=} 1 - \inf_{n \in \mathbb{N}} \inf_{u_1,\ldots,u_n \in \mathcal{U}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^n WK(A_i \mid u_i) \overset{(b)}{=} 1 - \inf_{n \in \mathbb{N}} \inf_{u_1,\ldots,u_n \in \mathcal{U}} \underbrace{\inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^n \int_{\mathcal{X}} K(A_i \mid x)\, W(dx \mid u_i)}_{①}\,, \tag{81}$$

where (a) holds by Theorem 1 and (b) holds by definition of composition of two kernels (8). Next, we lower-bound ① for any $n \in \mathbb{N}$ and $u_1, \ldots, u_n \in \mathcal{U}$. Let $\mathbb{P}$ be the maximal coupling of random variables $\{X_u\}_{u \in \mathcal{U}}$ with $X_u \sim W(\cdot \mid u)$ for each $u \in \mathcal{U}$, as defined in (34). We have

$$\begin{aligned} ① &\overset{(a)}{=} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^n \int_{\mathcal{X}} K(A_i \mid x)\, \mathbb{P}_{X_{u_i}}(dx) \overset{(b)}{=} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^n \mathbb{E}[K(A_i \mid X_{u_i})] \\ &\overset{(c)}{=} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \mathbb{E}\left[\sum_{i=1}^n K(A_i \mid X_{u_i})\right] \geq \mathbb{E}\left[\inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^n K(A_i \mid X_{u_i})\right] \overset{(d)}{=} \mathbb{E}\left[\int_{\mathcal{Y}} \left(\min_{i \in [n]} \frac{dK}{d\mu}(\cdot \mid X_{u_i})\right) d\mu\right], \end{aligned} \tag{82}$$

where (a) holds by definition of a coupling, the expectations from (b) onwards are taken with respect to $\{X_u\}_{u \in \mathcal{U}} \sim \mathbb{P}$, (c) holds by linearity of expectation, and (d) holds by Proposition 2. Combining (81) and (82), we have

$$\rho(WK) \leq 1 - \underbrace{\inf_{n \in \mathbb{N}} \inf_{u_1,\ldots,u_n \in \mathcal{U}} \mathbb{E}\left[\int_{\mathcal{Y}} \left(\min_{i \in [n]} \frac{dK}{d\mu}(\cdot \mid X_{u_i})\right) d\mu\right]}_{②}\,. \tag{83}$$

Next, we lower-bound ②. For each $k \in \mathbb{N}$, let $s_k \triangleq \{\hat{u}_{k,1}, \ldots, \hat{u}_{k,n_k}\} \subseteq \mathcal{U}$ be a finite sequence such that

$$\mathbb{E}\left[\int_{\mathcal{Y}} \left(\min_{i \in [n_k]} \frac{dK}{d\mu}(\cdot \mid X_{\hat{u}_{k,i}})\right) d\mu\right] \leq ② + \frac{1}{k}\,,$$

where the existence of such a sequence is guaranteed by the definition of ② as an infimum in (83). Construct an infinite sequence $\{\tilde{u}_1, \tilde{u}_2, \ldots\} \subseteq \mathcal{U}$ by concatenating the finite sequences s_k in ascending order of k. Define a sequence of functions $g_n : \mathcal{X}^{\mathcal{U}} \to \mathbb{R}$ (with respect to the cylinder σ-algebra on $\mathcal{X}^{\mathcal{U}}$) as

$$\forall n \in \mathbb{N},\ \ g_n\big(\{x_u\}_{u \in \mathcal{U}}\big) \triangleq \int_{\mathcal{Y}} \left(\min_{i \in [n]} \frac{dK}{d\mu}(\cdot \mid x_{\tilde{u}_i})\right) d\mu\,.$$

By construction, it holds that

$$\lim_{n \to \infty} \mathbb{E}\big[g_n\big(\{X_u\}_{u \in \mathcal{U}}\big)\big] = ②\,. \tag{84}$$

Each g_n is non-negative, by the non-negativity of the Radon-Nikodym derivative. Hence, we may define $h : \mathcal{X}^{\mathcal{U}} \to \mathbb{R}_+$ as the infimum $h \triangleq \inf_{n \in \mathbb{N}} g_n$, i.e.,

$$h\big(\{x_u\}_{u \in \mathcal{U}}\big) \triangleq \inf_{n \in \mathbb{N}} \int_{\mathcal{Y}} \left(\min_{i \in [n]} \frac{dK}{d\mu}(\cdot \mid x_{\tilde{u}_i})\right) d\mu\,. \tag{85}$$

By construction, $g_n \geq h$ for all $n \in \mathbb{N}$. Proceeding from (84), we thus have

$$② \geq \mathbb{E}\big[h\big(\{X_u\}_{u \in \mathcal{U}}\big)\big]\,. \tag{86}$$

Define the event $\mathcal{E} = \{\forall u, v \in \mathcal{U},\ X_u = X_v\}$, which is measurable since $\mathcal{U}$ is Polish. Combining (83) and (86), we have

$$\begin{aligned}\rho(WK) &\le 1 - \mathbb{E}\big[h\big(\{X_u\}_{u\in\mathcal{U}}\big)\big] \overset{(a)}{=} 1 - \mathbb{P}(\mathcal{E})\,\mathbb{E}\big[h\big(\{X_u\}_{u\in\mathcal{U}}\big) \mid \mathcal{E}\big] - \mathbb{P}(\mathcal{E}^{\mathrm{c}})\,\mathbb{E}\big[h\big(\{X_u\}_{u\in\mathcal{U}}\big) \mid \mathcal{E}^{\mathrm{c}}\big] \\ &\overset{(b)}{=} 1 - \mathbb{P}(\mathcal{E}) - \mathbb{P}(\mathcal{E}^{\mathrm{c}})\,\mathbb{E}\big[h\big(\{X_u\}_{u\in\mathcal{U}}\big) \mid \mathcal{E}^{\mathrm{c}}\big] \overset{(c)}{=} \mathbb{P}(\mathcal{E}^{\mathrm{c}})\,\underbrace{\mathbb{E}\big[1 - h\big(\{X_u\}_{u\in\mathcal{U}}\big) \mid \mathcal{E}^{\mathrm{c}}\big]}_{③},\end{aligned} \tag{87}$$

where (a) holds by the law of total expectation; (b) holds because, under the event $\mathcal{E}$, there exists some $x^* \in \mathcal{X}$ such that $X_u = x^*$ for all $u \in \mathcal{U}$ and so

$$h\big(\{X_u\}_{u\in\mathcal{U}}\big) = \int_{\mathcal{Y}} \left(\frac{dK}{d\mu}(\cdot \mid x^*)\right) d\mu = K(\mathcal{Y} \mid x^*) = 1$$

given $\mathcal{E}$; and (c) holds by the complement rule of probability and linearity of expectation. If $\mathbb{P}(\mathcal{E}^{\mathrm{c}}) = 0$, we have $\rho(WK) = 0$ and the desired upper bound (80) is trivially satisfied. Hence, assume $\mathbb{P}(\mathcal{E}^{\mathrm{c}}) > 0$ for the remainder of this argument.

Next, we upper-bound ③. Observe that for any bounded (multi)set $\mathcal{S} = \{x_u\}_{u\in\mathcal{U}} \in \mathcal{X}^{\mathcal{U}}$ (i.e., $\|\!|\mathcal{S}|\!\|_\infty < \infty$) and any $\epsilon > 0$,

$$\begin{aligned}1 - h(\mathcal{S}) &\overset{(a)}{=} 1 - \inf_{n\in\mathbb{N}} \int_{\mathcal{Y}} \left(\min_{i\in[n]} \frac{dK}{d\mu}(\cdot \mid x_{\tilde{u}_i})\right) d\mu \overset{(b)}{\le} 1 - \inf_{n\in\mathbb{N}} \inf_{u_1,\dots,u_n\in\mathcal{U}} \int_{\mathcal{Y}} \left(\min_{i\in[n]} \frac{dK}{d\mu}(\cdot \mid x_{u_i})\right) d\mu \\ &\overset{(c)}{=} 1 - \inf_{n\in\mathbb{N}} \inf_{u_1,\dots,u_n\in\mathcal{U}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\dots,A_n}} \sum_{i=1}^{n} K(A_i \mid x_{u_i}) \overset{(d)}{=} \rho_{\mathcal{S}}(K) \overset{(e)}{\le} \rho_{\mathcal{B}(a^*,\mathrm{rad}(\mathcal{S})+\epsilon)}(K) \\ &\overset{(f)}{\le} \sup_{a\in\mathcal{X}} \rho_{\mathcal{B}(a,\mathrm{rad}(\mathcal{S})+\epsilon)}(K) \overset{(g)}{\le} \sup_{a\in\mathcal{X}} \rho_{\mathcal{B}(a,\gamma\|\!|\mathcal{S}|\!\|_\infty+\epsilon)}(K) \overset{(h)}{=} \Theta(\gamma\,\|\!|\mathcal{S}|\!\|_\infty + \epsilon) \overset{(i)}{\le} \hat{\Theta}(\gamma\,\|\!|\mathcal{S}|\!\|_\infty + \epsilon)\,,\end{aligned}$$

where (a) holds by definition of h (85); (b) holds by lower-bounding the integrand for the specific sequence $\{\tilde{u}_1, \dots, \tilde{u}_n\}$ with the infimum over all length-n sequences; (c) holds by Proposition 2; (d) holds by Theorem 1 and the definition of complementary Doeblin coefficient; (e) holds for *some* center $a^* \in \mathcal{X}$ which satisfies $\sup_{x\in\mathcal{S}} \|x - a^*\| \le \mathrm{rad}(\mathcal{S}) + \epsilon$, because $\mathcal{S}' \mapsto \rho_{\mathcal{S}'}(K)$ is monotonically non-decreasing in its input set $\mathcal{S}'$; (f) holds by upper-bounding the value for the specific center a^* with the supremum over *all* centers $a \in \mathcal{X}$; (g) holds because $\mathrm{rad}(\mathcal{S}) \le \gamma\,\|\!|S|\!\|_\infty$ by rearranging (14); (h) holds by definition of Θ; and (i) holds because $\hat{\Theta}$ is the upper concave envelope (and hence an upper bound) of Θ. Therefore, if $\mathcal{S}$ has strictly positive diameter $\|\!|\mathcal{S}|\!\|_\infty > 0$,

$$\begin{aligned}1 - h(\mathcal{S}) &\overset{(a)}{\le} \hat{\Theta}(\gamma\,\|\!|\mathcal{S}|\!\|_\infty) \overset{(b)}{=} \hat{\Theta}\left(\gamma \sup_{u,v\in\mathcal{U}} \|x_u - x_v\|\right) \overset{(c)}{\le} \hat{\Theta}\left(2\gamma \sup_{u\in\mathcal{U}} \|x_u\|\right) \\ &\overset{(d)}{=} \hat{\Theta}\left(2\gamma\, M^{-1} \circ M\left(\sup_{u\in\mathcal{U}} \|x_u\|\right)\right) \overset{(e)}{=} \hat{\Theta}\left(2\gamma\, M^{-1}\left(\sup_{u\in\mathcal{U}} M(\|x_u\|)\right)\right),\end{aligned} \tag{88}$$

where (a) holds because $\hat{\Theta}$ is concave and thus continuous on its interior which contains $\gamma\,\|\!|\mathcal{S}|\!\|_\infty > 0$, and because $\epsilon > 0$ was arbitrary; (b) holds by definition of diameter (1); (c) holds by the triangle inequality, and because $\hat{\Theta}$ is non-decreasing by Lemma 9, Part 2; (d) holds because M is strictly increasing and hence invertible; and (e) holds because M is increasing. For notational convenience, define a function $G : \mathbb{R}_+ \to [0, 1]$ as $G(r) \triangleq \hat{\Theta}(2\gamma\, M^{-1}(r))$. Since $\hat{\Theta}$ is non-decreasing and concave, and since M^{-1} is increasing and concave (being the inverse of an increasing convex function), it follows that G is non-decreasing and concave. Combining (87) and (88), we have, akin to the argument of [25, Theorem 4],[13]

$$③ \le \mathbb{E}\left[G\left(\sup_{u\in\mathcal{U}} M(\|X_u\|)\right) \,\middle|\, \mathcal{E}^{\mathrm{c}}\right] \overset{(a)}{\le} G\left(\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, \mathcal{E}^{\mathrm{c}}\right]\right) \overset{(b)}{\le} G\left(\frac{1}{\mathbb{P}(\mathcal{E}^{\mathrm{c}})}\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|)\right]\right) \overset{(c)}{=} G\left(\frac{\|\!|W|\!\|_{\mathsf{AE}}}{\mathbb{P}(\mathcal{E}^{\mathrm{c}})}\right), \tag{89}$$

where (a) holds by Jensen's inequality and the concavity of G, (b) holds by the law of total expectation because

$$\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|)\right] = \mathbb{P}(\mathcal{E})\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, \mathcal{E}\right] + \mathbb{P}(\mathcal{E}^{\mathrm{c}})\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, \mathcal{E}^{\mathrm{c}}\right] \overset{(d)}{\ge} \mathbb{P}(\mathcal{E}^{\mathrm{c}})\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, \mathcal{E}^{\mathrm{c}}\right],$$

the division in (b) is well-defined because we assumed $\mathbb{P}(\mathcal{E}^{\mathrm{c}}) > 0$, (c) holds by definition of average extremal power (13) because $\mathbb{P}$ is the maximal coupling (34), and (d) holds because M is non-negative. Combining (87) and (89), we thus have

$$\rho(WK) \le \mathbb{P}(\mathcal{E}^{\mathrm{c}})\, G\left(\frac{\|\!|W|\!\|_{\mathsf{AE}}}{\mathbb{P}(\mathcal{E}^{\mathrm{c}})}\right). \tag{90}$$

Observe that the *perspective function* $H : (0, 1] \times \mathbb{R}_+ \to [0, 1]$ given by

$$H(s; a) \triangleq s\, G\left(\frac{a}{s}\right)$$

[13] It suffices to define G over $\mathbb{R}_+$ instead of $\mathbb{R}_+ \cup \{\infty\}$, because $\|\!|W|\!\|_{\mathsf{AE}} < \infty$ and so $\sup_{u\in\mathcal{U}} M(\|X_u\|) < \infty$ almost surely. Hence, the expectation $\mathbb{E}[G(\sup_{u\in\mathcal{U}} M(\|X_u\|))]$ is well-defined even if G is undefined at ∞.

is non-decreasing in s for any a, because for any $s,t \in (0,1]$, $s<t$ and $a>0$,[14] we have

$$H(s;a) \overset{(a)}{=} a\,\frac{G(a/s)-G(0)}{a/s-0} \overset{(b)}{\le} a\,\frac{G(a/t)-G(0)}{a/t-0} \overset{(c)}{=} H(t;a)\,,$$

where (a) and (c) hold because $G(0)=\hat{\Theta}(0)=\Theta(0)=0$, where the second equality follows from Lemma 9, Part 1; and (b) holds because G is concave and so the *difference quotient* $(G(r)-G(q))/(r-q)$ is monotonically non-decreasing in $r>q$ for any fixed q by the first inequality in [70, Exercise 3.1.b]. (Alternatively, if G is differentiable, we may use the first derivative test: for any $s\in(0,1]$ and $a\in\mathbb{R}_+$, we have

$$\frac{\partial H}{\partial s} = G\Big(\frac{a}{s}\Big) - \Big(\frac{a}{s}\Big)\,G'\Big(\frac{a}{s}\Big) \overset{(a)}{\ge} G\Big(\frac{a}{s}\Big) - \Big(\frac{a}{s}\Big)\,G'(\xi) \overset{(b)}{=} G\Big(\frac{a}{s}\Big) - \Big(G\Big(\frac{a}{s}\Big)-G(0)\Big) \overset{(c)}{=} 0\,,$$

where (a) holds for *any* $\xi\in[0,a/s]$ by concavity of G; (b) holds for *some* $\xi\in[0,a/s]$ by the mean value theorem; and (c) holds because $G(0)=\hat{\Theta}(0)=\Theta(0)=0$, where the second equality follows from Lemma 9, Part 1.) Next, observe that

$$\mathbb{P}(\mathcal{E}^{\mathrm{c}}) \overset{(a)}{=} \rho(W) \le t\,, \tag{91}$$

where (a) holds by Proposition 1 because $\mathbb{P}$ is the maximal coupling (34). Proceeding from (90), we have

$$\rho(WK) \overset{(a)}{=} H(\mathbb{P}(\mathcal{E}^{\mathrm{c}});|||W|||_{\mathsf{AE}}) \overset{(b)}{\le} H(t;|||W|||_{\mathsf{AE}}) \overset{(c)}{=} t\,\hat{\Theta}\Big(2\gamma\,M^{-1}\Big(\frac{|||W|||_{\mathsf{AE}}}{t}\Big)\Big) \overset{(d)}{\le} t\,\hat{\Theta}\Big(2\gamma\,M^{-1}\Big(\frac{p}{t}\Big)\Big) \tag{92}$$

as desired, where (a) holds by definition of H, (b) holds by (91) and because H is non-decreasing in s, (c) holds by definition of H and G, and (d) holds by the power constraint $|||W|||_{\mathsf{AE}}\le p$ and because $\hat{\Theta}$ and M^{-1} are non-decreasing.

Lower bound: Fix $t\in(0,1]$ and $p\in[0,\infty)$. By definition of Doeblin curve, we want to show that there exists a Markov kernel $W^*:\mathcal{U}\times\mathcal{F}_{\mathcal{X}}\to[0,1]$ satisfying $\rho(W^*)\le t$ and $|||W^*|||_{\mathsf{AE}}\le p$ such that $\rho(W^*K)\ge t\,\theta\big(\mathbf{0},M^{-1}\big(\frac{p}{t}\big)\big)$. Consider $\mathcal{U}\triangleq\mathcal{B}(\mathbf{0},M^{-1}(p/t))\subset\mathcal{X}$, equipped with the Borel σ-algebra induced on $\mathcal{U}$ by $\mathcal{X}$. Let $W:\mathcal{U}\times\mathcal{F}_{\mathcal{X}}\to[0,1]$ be the identity kernel $W(\cdot\mid u)\triangleq\delta_u(\cdot)$ for all $u\in\mathcal{U}$, and choose $W^*:\mathcal{U}\times\mathcal{F}_{\mathcal{X}}\to[0,1]$ to be the Markov kernel $W^*(\cdot\mid u)\triangleq t\,W(\cdot\mid u)+(1-t)\,\delta_{\mathbf{0}}(\cdot)$ for all $u\in\mathcal{U}$.

By inspection, the greatest common component of W is $\bigwedge_{u\in\mathcal{U}}W(\cdot\mid u)=0$. Thus, the complementary Doeblin coefficient of W, by (5), is $\rho(W)=1-\bigwedge_{u\in\mathcal{U}}W(\mathcal{X}\mid u)=1$, and the maximal coupling of random variables $\{X_u\}_{u\in\mathcal{U}}$ with $X_u\sim W(\cdot\mid u)$ for each $u\in\mathcal{U}$, as defined in (34), reduces to

$$\forall u\in\mathcal{U},\;\; X_u\sim\delta_u(\cdot)\,, \tag{93}$$

where $\{X_u\}_{u\in\mathcal{U}}$ are independent. Hence, the average extremal power of W is

$$|||W|||_{\mathsf{AE}} \overset{(a)}{=} \mathbb{E}\Big[\sup_{u\in\mathcal{U}}M(\|X_u\|)\Big] \overset{(b)}{=} \sup_{u\in\mathcal{U}}M(\|u\|) \overset{(c)}{=} M\circ M^{-1}\Big(\frac{p}{t}\Big) = \frac{p}{t}\,,$$

where (a) holds by definition of average extremal power (13), (b) holds by (93), and (c) holds because $\mathcal{U}=\mathcal{B}(\mathbf{0},M^{-1}(p/t))$ and M is increasing. By inspection, the composition of W with K is $WK(\cdot\mid u)=K(\cdot\mid u)$ for all $u\in\mathcal{U}$, and so the complementary Doeblin coefficient of WK is

$$\rho(WK)=\rho_{\mathcal{U}}(K) \overset{(a)}{=} \theta\Big(\mathbf{0},M^{-1}\Big(\frac{p}{t}\Big)\Big)\,,$$

where (a) holds by definition of θ and because $\mathcal{U}=\mathcal{B}(\mathbf{0},M^{-1}(p/t))$. Thus, by Lemmas 5 and 7, we have $\rho(W^*)=t\,\rho(W)=t$, $\rho(W^*K)=t\,\rho(WK)=t\,\theta\big(\mathbf{0},M^{-1}\big(\frac{p}{t}\big)\big)$, and $|||W^*|||_{\mathsf{AE}}=t\,|||W|||_{\mathsf{AE}}=p$ as desired. □

Next, we prove Corollary 2.

Proof of Corollary 2. Throughout this proof, let K denote one of the kernels K_1, K_2, or K_3 in general. For each kernel, consider the density function $g_K(z)=g_K(y-x)=\frac{dK}{dy}(y\mid x)$ with respect to the Lebesgue measure on $\mathbb{R}$:

$$g_{K_1}\big(z;\sigma^2\big)=\frac{1}{\sigma\sqrt{2\pi}}\exp\Big(-\frac{z^2}{2\sigma^2}\Big)\,,\;\; g_{K_2}(z;b)=\frac{1}{2b}\exp\Big(-\frac{|z|}{b}\Big)\,,\;\text{ and }\; g_{K_3}(z;\beta)=\frac{\sqrt{\beta}}{\pi}\Big(\frac{1}{1+\beta z^2}\Big)\,,$$

where g_K depends only on the difference $z=y-x$ because K is a convolution kernel. For any $r>0$, we have

$$\Theta(r) \overset{(a)}{=} \theta(0,r) \overset{(b)}{=} \rho_{[-r,r]}(K) \overset{(c)}{=} 1-\bigwedge_{x\in[-r,r]}K(\mathbb{R}\mid x) \overset{(d)}{=} 1-\int_{\mathbb{R}}\Big(\tilde{\bigwedge_{x\in[-r,r]}}\frac{dK}{dy}(y\mid x)\Big)\,dy = 1-\int_{\mathbb{R}}\Big(\min_{x\in[-r,r]}g_K(y-x)\Big)\,dy\,, \tag{94}$$

[14] The $a=0$ case is obvious by inspection, since $H(s;0)=s\,G(0)$ and $G(0)$ is non-negative by the range of G.

where (a) holds because K is a convolution kernel, (b) holds by definition of θ, (c) holds by the greatest common component characterization of Doeblin coefficient (5), and (d) holds by Lemma 4. By inspection, g_K is increasing on $(-\infty, 0]$ and decreasing on $[0, \infty)$. Therefore, the minimum of g_K over any interval $[a, b] \subset \mathbb{R}$ is achieved at one of the endpoints, i.e.,

$$\min_{z \in [a,b]} g_K(z) = \min\left\{g_K(a), g_K(b)\right\}. \tag{95}$$

Proceeding from (94), we thus have (cf. [25, Eq. 40])

$$\begin{aligned}
\Theta(r) &\overset{(a)}{=} 1 - \int_{\mathbb{R}} \min\left\{g_K(y-r), g_K(y+r)\right\} dy \\
&\overset{(b)}{=} 1 - \int_{-\infty}^{0} \min\left\{g_K(y-r), g_K(-|y+r|)\right\} dy - \int_{0}^{\infty} \min\left\{g_K(|y-r|), g_K(y+r)\right\} dy \\
&\overset{(c)}{=} 1 - \int_{-\infty}^{0} g_K(y-r)\, dy - \int_{0}^{\infty} g_K(y+r)\, dy \overset{(d)}{=} 1 - 2 f_K(-r),
\end{aligned} \tag{96}$$

where (a) holds by (95); (b) holds because g_K is symmetric about $z = 0$ by inspection; (c) holds because $y - r \le -|y+r| \le 0$ for $y \le 0$ and g_K is increasing on $z \le 0$, and $0 \le |y-r| \le y + r$ for $y \ge 0$ and g_K is decreasing on $z \ge 0$; f_K in (d) denotes the CDF $f_K(y) = \int_{-\infty}^{y} g_K(z)\, dz$; and (d) holds because g_K is symmetric about $z = 0$. Also, Θ is concave because

$$\Theta''(r) = -2 g_K'(-r) \overset{(a)}{<} 0 \tag{97}$$

for all $r > 0$, where (a) holds because g_K is increasing on $z < 0$. Hence, by Corollary 1 and (96),

$$\mathrm{F}_K^{\mathsf{AE}}(t; p) = t\left(1 - 2 f_K\left(-\sqrt{\frac{p}{t}}\right)\right). \tag{98}$$

It remains to compute the closed-form CDFs for the Gaussian, Laplace, and q-Gaussian kernels. For the Gaussian kernel K_1,

$$f_{K_1}(y) = \int_{-\infty}^{y} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{z^2}{2\sigma^2}\right) dz = \Phi\left(\frac{y}{\sigma}\right). \tag{99}$$

For the Laplace kernel K_2, for all $y \le 0$,

$$f_{K_2}(y) = \int_{-\infty}^{y} \frac{1}{2b} \exp\left(-\frac{|z|}{b}\right) dz = \frac{1}{2} \exp\left(\frac{y}{b}\right). \tag{100}$$

Lastly, for the q-Gaussian kernel K_3,

$$f_{K_3}(y) = \int_{-\infty}^{y} \frac{\sqrt{\beta}}{\pi}\left(\frac{1}{1+\beta z^2}\right) dz = \frac{1}{\pi} \arctan\left(\sqrt{\beta} y\right) + \frac{1}{2}. \tag{101}$$

Combining Equations (98) to (101) completes the proof. □

Next, we prove Proposition 7.

Proof of Proposition 7. Fix an absolutely continuous Markov kernel $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0, 1]$. By definition of Doeblin curve, we want to show that for any Markov kernel $W \in \mathcal{G}$ (i.e., satisfying the assumptions of Proposition 7) such that $\rho(W) \le t$, we have

$$\rho(WK) \le t\, \hat{\Theta}\left(2\gamma\, M^{-1}\left(\frac{p + \sigma\sqrt{2\log_e |\mathcal{U}|}}{t}\right)\right). \tag{102}$$

By following the proof of the upper bound in Theorem 2 up to (92), step (c), we have

$$\rho(WK) \le t\, \hat{\Theta}\left(2\gamma\, M^{-1}\left(\frac{|||W|||_{\mathsf{AE}}}{t}\right)\right). \tag{103}$$

Next, we upper-bound $|||W|||_{\mathsf{AE}}$. For notational convenience, let $Z_u = M(\|X_u\|)$ for each $u \in \mathcal{U}$. We have

$$\begin{aligned}
|||W|||_{\mathsf{AE}} &\overset{(a)}{=} \mathbb{E}\left[\max_{u \in \mathcal{U}} Z_u\right] \le \mathbb{E}\left[\max_{u \in \mathcal{U}} \mathbb{E}[Z_u] + \max_{u \in \mathcal{U}} \left\{Z_u - \mathbb{E}[Z_u]\right\}\right] \\
&= \max_{u \in \mathcal{U}} \mathbb{E}[Z_u] + \mathbb{E}\left[\max_{u \in \mathcal{U}} \left\{Z_u - \mathbb{E}[Z_u]\right\}\right] \overset{(b)}{\le} p + \underbrace{\mathbb{E}\left[\max_{u \in \mathcal{U}} \left\{Z_u - \mathbb{E}[Z_u]\right\}\right]}_{①},
\end{aligned} \tag{104}$$

where (a) holds by definition of average extremal power (13) and we use max instead of sup because $\mathcal{U}$ is finite, and (b) holds because any kernel $W \in \mathcal{G}$ satisfies $|||W|||_{\mathsf{UA}} \le p$. Next, we upper-bound ①. For any $\lambda > 0$, we have (cf. [71, Eq. 2])

$$① = \frac{1}{\lambda}\, \mathbb{E}\left[\log_e \exp\left(\max_{u \in \mathcal{U}} \lambda\left(Z_u - \mathbb{E}[Z_u]\right)\right)\right] \overset{(a)}{=} \frac{1}{\lambda}\, \mathbb{E}\left[\log_e \max_{u \in \mathcal{U}} e^{\lambda(Z_u - \mathbb{E}[Z_u])}\right] \overset{(b)}{\le} \frac{1}{\lambda} \log_e \mathbb{E}\left[\max_{u \in \mathcal{U}} e^{\lambda(Z_u - \mathbb{E}[Z_u])}\right]$$

$$\overset{(c)}{\leq} \frac{1}{\lambda} \log_e \mathbb{E}\left[\sum_{u\in\mathcal{U}} e^{\lambda(Z_u - \mathbb{E}[Z_u])}\right] \overset{(d)}{=} \frac{1}{\lambda} \log_e \sum_{u\in\mathcal{U}} \mathbb{E}\left[e^{\lambda(Z_u - \mathbb{E}[Z_u])}\right] \overset{(e)}{\leq} \frac{1}{\lambda} \log_e \sum_{u\in\mathcal{U}} \exp\left(\frac{\sigma^2\lambda^2}{2}\right) = \frac{\log_e |\mathcal{U}|}{\lambda} + \frac{\sigma^2\lambda}{2},$$

where (a) holds because exp is increasing, (b) holds by Jensen's inequality, (c) holds by upper-bounding the maximum of a collection of positive terms with their sum, (d) holds by linearity of expectation, and (e) holds by the sub-Gaussianity of W. Choosing the optimal value $\lambda = \sqrt{2\log_e |\mathcal{U}|}/\sigma$ to balance the summands, we have

$$① \leq \sigma\sqrt{2\log_e |\mathcal{U}|}\,. \tag{105}$$

Since $\hat{\Theta}$ and M^{-1} are non-decreasing (as established in the proof of Theorem 2), combining Equations (103) to (105) proves (102) as desired. □

Next, we prove Proposition 8.

Proof of Proposition 8. Fix an absolutely continuous Markov kernel $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to [0,1]$. By definition of Doeblin curve, we want to show that for any Markov kernel $W \in \mathcal{G}$ (i.e., satisfying the assumptions of Proposition 8) such that $\rho(W) \leq t$, we have

$$\rho(WK) \leq t\,\hat{\Theta}\left(2\gamma\, M^{-1}\left(\frac{p}{t} + \frac{32\sigma}{t}\int_0^{\|\mathcal{U}\|_\infty} \sqrt{\log_e N(\epsilon, \mathcal{U}, d_\mathcal{U})}\, d\epsilon\right)\right). \tag{106}$$

By following the proof of the upper bound in Theorem 2 up to (92), step (c), we have

$$\rho(WK) \leq t\,\hat{\Theta}\left(2\gamma\, M^{-1}\left(\frac{|||W|||_{\mathsf{AE}}}{t}\right)\right). \tag{107}$$

Next, we upper-bound $|||W|||_{\mathsf{AE}}$. For notational convenience, let $Z_u = M(\|X_u\|)$ for each $u \in \mathcal{U}$. Observe that

$$\sup_{u\in\mathcal{U}} Z_u \leq \sup_{u\in\mathcal{U}} \mathbb{E}[Z_u] + \sup_{u\in\mathcal{U}} \{Z_u - \mathbb{E}[Z_u]\} \overset{(a)}{=} |||W|||_{\mathsf{UA}} + \sup_{u\in\mathcal{U}} \{Z_u - \mathbb{E}[Z_u]\} \overset{(b)}{\leq} p + \sup_{u\in\mathcal{U}} \{Z_u - \mathbb{E}[Z_u]\}, \tag{108}$$

where (a) holds by definition of uniform average power (12) and (b) holds because any kernel $W \in \mathcal{G}$ satisfies $|||W|||_{\mathsf{UA}} \leq p$. Hence,

$$\begin{aligned} |||W|||_{\mathsf{AE}} &\overset{(a)}{=} \mathbb{E}\left[\sup_{u\in\mathcal{U}} Z_u\right] \overset{(b)}{\leq} p + \mathbb{E}\left[\sup_{u\in\mathcal{U}} \{Z_u - \mathbb{E}[Z_u]\}\right] \overset{(c)}{=} p + \mathbb{E}\left[\sup_{u\in\mathcal{U}} \{Z_u - \mathbb{E}[Z_u]\} - (Z_{v^*} - \mathbb{E}[Z_{v^*}])\right] \\ &\leq p + \mathbb{E}\left[\sup_{u,v\in\mathcal{U}} \{(Z_u - \mathbb{E}[Z_u]) - (Z_v - \mathbb{E}[Z_v])\}\right] \overset{(d)}{\leq} p + 32\sigma \int_0^{\|\mathcal{U}\|_\infty} \sqrt{\log_e N(\epsilon, \mathcal{U}, d_\mathcal{U})}\, d\epsilon, \end{aligned} \tag{109}$$

where all the expectations are taken with respect to the maximal coupling $\{X_u\}_{u\in\mathcal{U}} \sim \mathbb{P}$ as defined in (34), (a) holds by definition of average extremal power (13), (b) holds by (108), (c) holds for *any fixed* $v^* \in \mathcal{U}$ (i.e., v^* is chosen independently of $\{Z_u\}_{u\in\mathcal{U}}$), and (d) holds by *Dudley's entropy integral bound* [72, Theorem 5.22]. Since $\hat{\Theta}$ and M^{-1} are non-decreasing (as established in the proof of Theorem 2), combining (107) and (109) proves (106) as desired. □

Lastly, we prove Proposition 9.

Proof of Proposition 9. Fix an absolutely continuous Markov kernel $K : \mathcal{X} \times \mathcal{F}_\mathcal{Y} \to [0,1]$. Fix $t \in (0,1]$. By following the proof of the lower bound in Theorem 2, we obtain that the Markov kernel $W^* : \mathcal{U} \times \mathcal{F}_\mathcal{X} \to [0,1]$ from $\mathcal{U} \triangleq \mathcal{B}(\mathbf{0}, M^{-1}(p/t))$ to $\mathcal{X}$ given by $W^*(A \mid u) \triangleq (1-t)\,\delta_\mathbf{0}(A) + t\,\delta_u(A)$ for all $u \in \mathcal{U}$ and $A \in \mathcal{F}_\mathcal{X}$ satisfies $\rho(W^*) = t$, $|||W^*|||_{\mathsf{AE}} = p$, and $\rho(W^*K) = t\,\theta(\mathbf{0}, M^{-1}(p/t))$. By the discussion immediately following Definition 1, we have $|||W^*|||_{\mathsf{UA}} \leq |||W^*|||_{\mathsf{AE}} = p$. Hence, $\mathrm{F}_K^{\mathsf{UA}}(t;p) = \sup\{\rho(WK) : \rho(W) \leq t,\ |||W|||_{\mathsf{UA}} \leq p\} \geq \rho(W^*K) = t\,\theta\big(\mathbf{0}, M^{-1}\big(\frac{p}{t}\big)\big)$ as desired. □

V. Proofs of Main Results on Applications

In this section, we prove the main results presented in Section III, pertaining to applications of Doeblin curves.

A. Proofs for Generalization Error

First, we prove Lemma 2.

Proof of Lemma 2. For each $t \in [T]$, define the following random variables representing intermediate computations within the update rule:

$$U_t \triangleq W_{t-1} - \frac{\eta_t}{|\mathcal{B}_t|} \sum_{j\in\mathcal{B}_t} \nabla g(W_{t-1}, Z_j)\,, \quad V_t \triangleq U_t + m_t N_t\,, \quad W_t \triangleq \operatorname*{proj}_{\mathcal{W}}(V_t)\,.$$

Conditioned on the event that $i \in \mathcal{B}_t$ (i.e., Z_i is used in the tth iteration), the following Markov chain holds, because Z_i will not be used in any future iteration (cf. [41]):

$$Z_i \to \underline{U_t \to V_t \to W_t} \to \underline{U_{t+1} \to V_{t+1} \to W_{t+1}} \to \cdots \to \underline{U_{T-1} \to V_{T-1} \to W_{T-1}} \to \underline{U_T \to V_T \to W_T}\,.$$

It follows that

$$\begin{aligned} I_{\mathsf{TV}}(W_T; Z_i) &\overset{(a)}{\le} I_{\mathsf{TV}}(V_T; Z_i) \overset{(b)}{=} \|P_{V_T,Z_i} - P_{V_T} \otimes P_{Z_i}\|_{\mathsf{TV}} \overset{(c)}{=} \mathop{\mathbb{E}}_{z\sim P_{Z_i}}\left[\left\|P_{V_T|Z_i=z} - P_{V_T}\right\|_{\mathsf{TV}}\right] \\ &\overset{(d)}{=} \mathop{\mathbb{E}}_{z\sim P_{Z_i}}\left[\left\|P_{U_T|Z_i=z} * \mathsf{Normal}\big(0, m_T^2 I\big) - P_{U_T} * \mathsf{Normal}\big(0, m_T^2 I\big)\right\|_{\mathsf{TV}}\right], \end{aligned} \tag{110}$$

where (a) holds by the data processing inequality for f-information [46, Theorem 7.16], (b) holds by definition of TV-information (17), (c) holds because $P_{V_T,Z_i} = P_{Z_i} P_{V_T|Z_i}$ by the chain rule of probability, and (d) holds by definition of V_T. Next, observe that for any $z \in \mathcal{Z}$, the power of the distribution $P_{U_T|Z_i=z}$ is bounded as

$$\begin{aligned} \mathbb{E}\Big[\|U_T\|_2^2 \,\Big|\, Z_i = z\Big] &\overset{(a)}{=} \mathbb{E}\left[\left\|W_{T-1} - \frac{\eta_T}{|\mathcal{B}_T|}\sum_{j\in\mathcal{B}_T} \nabla g(W_{T-1}, Z_j)\right\|_2^2 \,\middle|\, Z_i = z\right] \\ &\overset{(b)}{\le} 2\mathbb{E}\Big[\|W_{T-1}\|_2^2 \,\Big|\, Z_i = z\Big] + 2\mathbb{E}\left[\left\|\frac{\eta_T}{|\mathcal{B}_T|}\sum_{j\in\mathcal{B}_T} \nabla g(W_{T-1}, Z_j)\right\|_2^2 \,\middle|\, Z_i = z\right] \\ &\overset{(c)}{\le} 2\mathbb{E}\Big[\|W_{T-1}\|_2^2 \,\Big|\, Z_i = z\Big] + 2\mathbb{E}\left[\left(\frac{\eta_T}{|\mathcal{B}_T|}\sum_{j\in\mathcal{B}_T} \|\nabla g(W_{T-1}, Z_j)\|_2\right)^2 \,\middle|\, Z_i = z\right] \\ &\overset{(d)}{\le} 2\mathbb{E}\Big[\|W_{T-1}\|_2^2 \,\Big|\, Z_i = z\Big] + 2\eta_T^2 L^2 \overset{(e)}{\le} m_T^2 p_T\,, \end{aligned} \tag{111}$$

where (a) holds by definition of U_T, (b) holds by the identity

$$\|x+y\|_2^2 \le \|x+y\|_2^2 + \|x-y\|_2^2 = 2\|x\|_2^2 + 2\|y\|_2^2$$

and linearity of expectation, (c) holds by the triangle inequality, (d) holds by the bound on the loss gradient (15), and (e) holds by definition of p_T (19). Similarly, the power of the distribution P_{U_T} is bounded as

$$\mathbb{E}\Big[\|U_T\|_2^2\Big] \overset{(a)}{=} \mathop{\mathbb{E}}_{z\sim P_{Z_i}}\Big[\mathbb{E}\Big[\|U_T\|_2^2 \,\Big|\, Z_i = z\Big]\Big] \overset{(b)}{\le} \mathop{\mathbb{E}}_{z\sim P_{Z_i}}\left[m_T^2 p_T\right] = m_T^2 p_T\,,$$

where (a) holds by the tower rule of expectation and (b) holds by (111). Continuing from (110), we have

$$\begin{aligned} I_{\mathsf{TV}}(W_T; Z_i) &\overset{(a)}{\le} \mathop{\mathbb{E}}_{z\sim P_{Z_i}}\left[\mathrm{F}_\Phi^{\mathsf{UA}}\left(\left\|P_{U_T|Z_i=z} - P_{U_T}\right\|_{\mathsf{TV}}; p_T\right)\right] \overset{(b)}{\le} \mathop{\mathbb{E}}_{z\sim P_{Z_i}}\left[\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\left(\left\|P_{U_T|Z_i=z} - P_{U_T}\right\|_{\mathsf{TV}}; p_T\right)\right] \\ &\overset{(c)}{\le} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\left(\mathop{\mathbb{E}}_{z\sim P_{Z_i}}\left[\left\|P_{U_T|Z_i=z} - P_{U_T}\right\|_{\mathsf{TV}}\right]; p_T\right) \overset{(d)}{=} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\left(\|P_{U_T,Z_i} - P_{U_T} \otimes P_{Z_i}\|_{\mathsf{TV}}; p_T\right) \\ &\overset{(e)}{=} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\left(I_{\mathsf{TV}}(U_T; Z_i); p_T\right) \overset{(f)}{\le} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\left(I_{\mathsf{TV}}(W_{T-1}; Z_i); p_T\right), \end{aligned} \tag{112}$$

where (a) holds by Proposition 6; (b) holds because $\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}$ is the upper concave envelope, and hence an upper bound, of $\mathrm{F}_\Phi^{\mathsf{UA}}$; (c) holds by Jensen's inequality; (d) holds by the chain rule of probability; (e) holds by definition of TV-information (17); and (f) holds by the data processing inequality, and because $\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}$ is non-decreasing by Lemma 9, Part 2.

Finally, by recursively applying the arguments in (110) and (112) above, we obtain

$$I_{\mathsf{TV}}(W_T; Z_i) \le \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Big(\cdots \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Big(\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Big(I_{\mathsf{TV}}(W_t; Z_i); p_{t+1}\Big); p_{t+2}\Big)\cdots; p_T\Big)$$

as desired, where the direction of the bound holds at each recursive step because each layer of $\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}$ is non-decreasing by Lemma 9, Part 2. □

Now, we are ready to prove Theorem 3.

Proof of Theorem 3. Using the high-level argument of [41], we first bound the expected generalization error in terms of the TV-information between model parameters and data samples. We have

$$|\mathbb{E}[G_\mu(W_T) - G_{\mathcal{S}}(W_T)]| \overset{(a)}{\le} \frac{A}{n}\sum_{i=1}^n I_{\mathsf{TV}}(W_T; Z_i) \overset{(b)}{=} \frac{A}{n}\sum_{t=1}^T \sum_{i\in\mathcal{B}_t} I_{\mathsf{TV}}(W_T; Z_i) + \frac{A}{n}\sum_{i\notin\cup_{t=1}^T \mathcal{B}_t} I_{\mathsf{TV}}(W_T; Z_i)$$

$$\overset{(c)}{=} \frac{A}{n}\sum_{t=1}^{T}\sum_{i\in\mathcal{B}_t} I_{\mathsf{TV}}(W_T;Z_i) \overset{(d)}{\le} \frac{A}{n}\sum_{t=1}^{T}\sum_{i\in\mathcal{B}_t} \hat{\mathrm{F}}_{\Phi}^{\mathsf{UA}}\Big(\cdots \hat{\mathrm{F}}_{\Phi}^{\mathsf{UA}}\Big(\underbrace{I_{\mathsf{TV}}(W_t;Z_i)}_{①};p_{t+1}\Big)\cdots;p_T\Big)\,, \tag{113}$$

where (a) holds by Lemma 1; (b) holds by grouping the data indices $i\in[n]$ by the iteration $t\in[T]$ in which they are used (if any), since the mini-batches $\mathcal{B}_1,\dots,\mathcal{B}_T$ are disjoint; (c) holds because data samples which are unused during training are statistically independent of the final parameters and thus have zero TV-information; and (d) holds by Lemma 2.

Next, we upper-bound ①. Define the following random variables representing intermediate steps in the update rule:

$$U_t \triangleq W_{t-1} - \frac{\eta_t}{|\mathcal{B}_t|}\sum_{j\in\mathcal{B}_t}\nabla g(W_{t-1},Z_j)\,,\quad V_t \triangleq U_t + m_t N_t\,,\quad W_t \triangleq \operatorname*{proj}_{\mathcal{W}}(V_t)\,,$$

which form the Markov chain $Z_i \to U_t \to V_t \to W_t$. Following the argument in [41, Lemma 9], we have

$$① \overset{(a)}{\le} I_{\mathsf{TV}}(V_t;Z_i) \overset{(b)}{=} \|P_{V_t,Z_i} - P_{V_t}\otimes P_{Z_i}\|_{\mathsf{TV}} \overset{(c)}{=} \mathop{\mathbb{E}}_{z\sim\mu}\big[\big\|P_{V_t|Z_i=z} - P_{V_t}\big\|_{\mathsf{TV}}\big] \overset{(d)}{=} \mathop{\mathbb{E}}_{z\sim\mu}\big[\underbrace{\big\|P_{U_t+m_tN_t|Z_i=z} - P_{U_t+m_tN_t}\big\|_{\mathsf{TV}}}_{②}\big]\,, \tag{114}$$

where (a) holds by the data processing inequality for f-information [46, Theorem 7.16], (b) holds by definition of TV-information (17), (c) holds because $P_{V_t,Z_i} = P_{Z_i}P_{V_t|Z_i}$ by the chain rule of probability and $P_{Z_i}=\mu$ since the data samples are drawn i.i.d., and (d) holds by definition of V_t.

Next, we upper-bound ② for any fixed $z\in\mathcal{Z}$. For any two distributions P and Q, define the *optimal transport cost*

$$\mathsf{W}(P,Q;m_t) \triangleq \frac{1}{2m_t}\inf_{\substack{P_{X,Y}:\\ P_X=P,\,P_Y=Q}} \mathop{\mathbb{E}}_{(X,Y)\sim P_{X,Y}}\big[\|X-Y\|_2\big]\,, \tag{115}$$

where the infimum is taken over all couplings $P_{X,Y}$ of the random variables X and Y with respective marginals $P_X=P$ and $P_Y=Q$. By [41, Lemma 6], it holds that

$$② \le \mathsf{W}\big(P_{U_t|Z_i=z}, P_{U_t}; m_t\big)\,. \tag{116}$$

Define two random variables

$$U^* \triangleq W_{t-1} - \frac{\eta_t}{|\mathcal{B}_t|}\left(\sum_{\substack{j\in\mathcal{B}_t:\\ j\neq i}}\nabla g(W_{t-1},Z_j) + \nabla g(W_{t-1},z)\right),\quad U^\dagger \triangleq W_{t-1} - \frac{\eta_t}{|\mathcal{B}_t|}\sum_{j\in\mathcal{B}_t}\nabla g(W_{t-1},Z_j)\,. \tag{117}$$

These random variables have marginals $U^*\sim P_{U_t|Z_i=z}$ and $U^\dagger\sim P_{U_t}$, by definition of U_t. Hence, continuing from (116),

$$② \overset{(a)}{\le} \frac{1}{2m_t}\mathbb{E}\big[\big\|U^*-U^\dagger\big\|_2\big] \overset{(b)}{=} \frac{\eta_t}{2m_t\,|\mathcal{B}_t|}\mathbb{E}\Big[\big\|\nabla g(W_{t-1},Z_i) - \nabla g(W_{t-1},z)\big\|_2\Big]\,, \tag{118}$$

where (a) holds by upper-bounding the infimum in (115) with the specific instance (117), and (b) holds by definition of U^* and $U^\dagger$.

Next, observe that W_{t-1} and Z_i are statistically independent, because $i\in\mathcal{B}_t$ by (113) and thus Z_i is not used in any iteration except t. Combining (114) and (118), we have

$$\begin{aligned}
① &\le \mathop{\mathbb{E}}_{z\sim\mu}\left[\frac{\eta_t}{2m_t\,|\mathcal{B}_t|}\mathop{\mathbb{E}}_{\substack{(W_{t-1},Z_i)\\ \sim P_{W_{t-1}}\otimes\mu}}\Big[\big\|\nabla g(W_{t-1},Z_i) - \nabla g(W_{t-1},z)\big\|_2\Big]\right]\\
&\overset{(a)}{=} \frac{\eta_t}{2m_t\,|\mathcal{B}_t|}\mathop{\mathbb{E}}_{\substack{(W_{t-1},Z_i,z)\\ \sim P_{W_{t-1}}\otimes\mu\otimes\mu}}\Big[\big\|\nabla g(W_{t-1},Z_i) - \nabla g(W_{t-1},z)\big\|_2\Big]\\
&\overset{(b)}{\le} \frac{\eta_t}{2m_t\,|\mathcal{B}_t|}\left(\mathop{\mathbb{E}}_{\substack{(W_{t-1},Z_i)\\ \sim P_{W_{t-1}}\otimes\mu}}\Big[\Big\|\nabla g(W_{t-1},Z_i) - \mathop{\mathbb{E}}_{\substack{(W_{t-1},Z)\\ \sim P_{W_{t-1}}\otimes\mu}}[\nabla g(W_{t-1},Z)]\Big\|_2\Big] + \mathop{\mathbb{E}}_{\substack{(W_{t-1},z)\\ \sim P_{W_{t-1}}\otimes\mu}}\Big[\Big\|\nabla g(W_{t-1},z) - \mathop{\mathbb{E}}_{\substack{(W_{t-1},Z)\\ \sim P_{W_{t-1}}\otimes\mu}}[\nabla g(W_{t-1},Z)]\Big\|_2\Big]\right)\\
&\overset{(c)}{\le} \frac{\eta_t\sigma_{t-1}}{m_t\,|\mathcal{B}_t|}\,,
\end{aligned} \tag{119}$$

where (a) holds by Tonelli's theorem and linearity of expectation, (b) holds by the triangle inequality and linearity of expectation, and (c) holds by definition of σ_{t-1}. Combining (113) and (119), we have

$$|\mathbb{E}[G_\mu(W_T) - G_{\mathcal{S}}(W_T)]| \le \frac{A}{n}\sum_{t=1}^{T}\sum_{i\in\mathcal{B}_t}\hat{\mathrm{F}}_{\Phi}^{\mathsf{UA}}\Big(\cdots\hat{\mathrm{F}}_{\Phi}^{\mathsf{UA}}\Big(\frac{\eta_t\sigma_{t-1}}{m_t\,|\mathcal{B}_t|};p_{t+1}\Big)\cdots;p_T\Big) = A\sum_{t=1}^{T}\frac{|\mathcal{B}_t|}{n}\,\hat{\mathrm{F}}_{\Phi}^{\mathsf{UA}}\Big(\cdots\hat{\mathrm{F}}_{\Phi}^{\mathsf{UA}}\Big(\frac{\eta_t\sigma_{t-1}}{m_t\,|\mathcal{B}_t|};p_{t+1}\Big)\cdots;p_T\Big)$$

as desired. □

B. Proofs for Reliable Computation

First, we prove Lemma 3.

Proof of Lemma 3. Let $\Psi : \mathcal{U}^q \times \mathcal{F}_{\mathcal{V}}^{\otimes q} \to [0,1]$ denote the Markov kernel from $\mathcal{U}^q$ to $\mathcal{V}^q$ such that for all $u^{(1:q)} \in \mathcal{U}^q$, $\Psi(\cdot \mid u^{(1:q)})$ is the maximal coupling of $\Phi(\cdot \mid u^{(1)}), \ldots, \Phi(\cdot \mid u^{(q)})$ defined in (34). Given π_U, construct the coupling $\pi_V : \mathcal{F}_{\mathcal{V}}^{\otimes q} \to [0,1]$ as

$$\forall B \in \mathcal{F}_{\mathcal{V}}^{\otimes q}, \quad \pi_V(B) \triangleq \int_{\mathcal{U}^q} \Psi\Big(B \,\Big|\, u^{(1:q)}\Big)\, d\pi_U\Big(u^{(1:q)}\Big) \tag{120}$$

$$= \mathop{\mathbb{E}}_{u^{(1:q)} \sim \pi_U}\Big[\Psi\Big(B \,\Big|\, u^{(1:q)}\Big)\Big]. \tag{121}$$

This defines a valid coupling of $P_V^{(1)}, \ldots, P_V^{(q)}$, because for any $\ell \in [q]$ and any $A \in \mathcal{F}_{\mathcal{V}}$,

$$\pi_V\big(\mathcal{V}^{\ell-1} \times A \times \mathcal{V}^{q-\ell}\big) \overset{(a)}{=} \int_{\mathcal{U}^q} \Psi\Big(\mathcal{V}^{\ell-1} \times A \times \mathcal{V}^{q-\ell} \,\Big|\, u^{(1:q)}\Big)\, d\pi_U\Big(u^{(1:q)}\Big)$$
$$\overset{(b)}{=} \int_{\mathcal{U}^q} \Phi\Big(A \,\Big|\, u^{(\ell)}\Big)\, d\pi_U\Big(u^{(1:q)}\Big) \overset{(c)}{=} \int_{\mathcal{U}} \Phi\Big(A \,\Big|\, u^{(\ell)}\Big)\, dP_U^{(\ell)}\Big(u^{(\ell)}\Big) \overset{(d)}{=} P_V^{(\ell)}(A)\,,$$

where (a) holds by the definition of π_V (120), (b) holds because $\Psi(\cdot \mid u^{(1:q)})$ is a coupling of $\Phi(\cdot \mid u^{(1)}), \ldots, \Phi(\cdot \mid u^{(q)})$, (c) holds because π_U is a coupling of $P_U^{(1)}, \ldots, P_U^{(q)}$, and (d) holds by (20). Also, π_V satisfies

$$\mathop{\mathbb{P}}_{V^{(1:q)} \sim \pi_V}\Big(\neg\Big(V^{(1)} = \cdots = V^{(q)}\Big)\Big)$$
$$\overset{(a)}{=} \mathop{\mathbb{E}}_{U^{(1:q)} \sim \pi_U}\left[\mathop{\mathbb{P}}_{V^{(1:q)} \sim \Psi(\cdot \mid U^{(1:q)})}\Big(\neg\Big(V^{(1)} = \cdots = V^{(q)}\Big)\Big)\right] \overset{(b)}{=} \mathop{\mathbb{E}}_{U^{(1:q)} \sim \pi_U}\Big[\rho\Big(\big[\Phi(\cdot \mid U^{(1)}), \ldots, \Phi(\cdot \mid U^{(q)})\big]\Big)\Big]$$
$$= \mathop{\mathbb{E}}_{U^{(1:q)} \sim \pi_U}\Big[\rho\Big(\big[\delta_{U^{(1)}}, \ldots, \delta_{U^{(q)}}\big]\Phi\Big)\Big] \overset{(c)}{\leq} \mathop{\mathbb{E}}_{U^{(1:q)} \sim \pi_U}\Big[\mathrm{F}_\Phi^{\mathsf{UA}}\Big(\rho\Big(\big[\delta_{U^{(1)}}, \ldots, \delta_{U^{(q)}}\big]\Big); p\Big)\Big]$$
$$= \mathop{\mathbb{E}}_{U^{(1:q)} \sim \pi_U}\Big[\mathrm{F}_\Phi^{\mathsf{UA}}\Big(\mathbb{1}\Big\{\neg\Big(U^{(1)} = \cdots = U^{(q)}\Big)\Big\}; p\Big)\Big] \overset{(d)}{\leq} \mathop{\mathbb{E}}_{U^{(1:q)} \sim \pi_U}\Big[\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Big(\mathbb{1}\Big\{\neg\Big(U^{(1)} = \cdots = U^{(q)}\Big)\Big\}; p\Big)\Big]$$
$$\overset{(e)}{\leq} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Big(\mathop{\mathbb{E}}_{U^{(1:q)} \sim \pi_U}\Big[\mathbb{1}\Big\{\neg\Big(U^{(1)} = \cdots = U^{(q)}\Big)\Big\}\Big]; p\Big) \overset{(f)}{=} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Big(\mathop{\mathbb{P}}_{U^{(1:q)} \sim \pi_U}\Big(\neg\Big(U^{(1)} = \cdots = U^{(q)}\Big)\Big); p\Big)$$

as desired, where (a) holds by (121), (b) holds by the maximal coupling characterization of Doeblin coefficients (Proposition 1), (c) holds by definition of Doeblin curve (9) because the kernel of Dirac measures has uniform average power

$$\big|\!\big|\!\big|\big[\delta_{U^{(1)}}, \ldots, \delta_{U^{(q)}}\big]\big|\!\big|\!\big|_{\mathsf{UA}} = \max_{\ell \in [q]} \int_{\mathcal{U}} M(\|u\|)\, d\delta_{U^{(\ell)}}(u) = \max_{\ell \in [q]} M\Big(\|U^{(\ell)}\|\Big) \leq \max_{u \in \mathcal{U}} M(\|u\|) = p\,,$$

(d) holds because $\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}$ is the upper concave envelope (and hence an upper bound) of $\mathrm{F}_\Phi^{\mathsf{UA}}$, (e) holds by Jensen's inequality, and (f) holds because the probability of an event is the expectation of its indicator. □

Now, we are ready to prove Theorem 4.

Proof of Theorem 4. Define a function $f : [0,1] \to [0,1]$ as $f(t) \triangleq \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}(\min\{1, bt\}; p)$. As a prelude to our main argument, we establish the following useful result concerning fixed point convergence of f. Observe that $[0,1]$ is a compact and totally ordered set so that any decreasing sequence $\{t_s\}_{s \in \mathbb{N}} \subset [0,1]$ has $\inf_{s \in \mathbb{N}} t_s \in [0,1]$. Furthermore, $\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}(\cdot\,; p)$ is non-decreasing and continuous on $[0,1]$ by Lemma 9, Parts 2 and 3. Hence, f is non-decreasing and continuous on $[0,1]$ (being the composition of two non-decreasing and continuous functions), and so $f(\inf_{s \in \mathbb{N}} t_s) = \inf_{s \in \mathbb{N}} f(t_s)$. Also, $1 \geq f(1)$ by the range of f. Hence, applying Kleene's fixed point theorem [73] on $([0,1], \geq)$,

$$\lim_{s \to \infty} f^{(s)}(1) \overset{(a)}{=} \inf_{s \in \mathbb{N}} f^{(s)}(1) = \max\{t \in [0,1] : f(t) = t\}\,,$$

where $f^{(s)}$ denotes the s-fold composition of f, and (a) holds because $f^{(s)}(1)$ is monotonically non-increasing in s.

With this groundwork established, we commence our main argument. Fix $\epsilon > 0$, and fix n sufficiently large such that for all $s \geq \lceil \log_b(n) \rceil$,

$$f^{(s)}(1) \leq \max\{t \in [0,1] : f(t) = t\} + \epsilon\,. \tag{122}$$

Consider any n-input circuit of noisy gates where each gate has at most b inputs. Define the *height* of an input vertex X_i as the length of the shortest directed path from X_i to the output vertex Y_m, where the existence of such a path is guaranteed for

each X_i by the formal model defined in Section III-B. Since the in-degree of each gate vertex is at most b, there must exist at least one input vertex whose height is $\lceil \log_b(n) \rceil$ or greater.[15] Without loss of generality, let X_1 be such an input vertex.

Fix any values $x_2, \dots, x_n \in \mathcal{Q}$ for the remaining inputs. We define some notation used throughout the remainder of our proof. To refer to input and gate vertices in a unified manner, let $W_k \triangleq X_{-k}$ for all $k \in \{-1, \dots, -n\}$, let $W_k \triangleq Y_k$ for all $k \in [m]$, and let $\mathcal{O}_j \triangleq \{-i : i \in \mathcal{N}_j\} \cup \mathcal{M}_j$ for all $j \in [m]$. Given a subset of indices $\mathcal{S} \subseteq \{-1, \dots, -n\} \cup [m]$, let $w_{\mathcal{S}} \triangleq \{w_k\}_{k \in \mathcal{S}}$ refer to the corresponding collection of subscripted variables. For any W, let $P_W^{(\ell)}$ be the marginal distribution of W induced by the circuit when setting $X_1 = \ell$ and $X_i = x_i$ for all $i > 1$. Lastly, let $a \wedge b \triangleq \min\{a, b\}$ for any scalars $a, b \in \mathbb{R}$.

Recall from Section III-B that we index the gate vertices in topological order, such that $\mathcal{M}_j \subseteq [j-1]$ for each $j \in [m]$. We follow the strategy in [25, Section 5.3]. Consider the following algorithm for constructing a coupling of the marginal distributions at each vertex in the circuit:

1) For each $i \in [n]$, let $\pi_{W_{-i}}$ be the maximal coupling of $P_{X_i}^{(1)}, \dots, P_{X_i}^{(q)}$.
2) For each $j \in [m]$:
   a) Let π_{Z_j} be the pushforward measure of $\bigotimes_{k \in \mathcal{O}_j} \pi_{W_k}$ through the function $\Gamma_j^{\otimes q} : (\mathbb{R}^d)^{b_j \times q} \to \mathcal{Q}^q$ which accepts q copies of input variables and independently applies Γ_j on each copy to produce q outputs, i.e., $Z_j^{(1:q)} = \Gamma_j^{\otimes q}\big(W_{\mathcal{O}_j}^{(1:q)}\big)$, where for each $\ell \in [q]$,
   $$Z_j^{(\ell)} = \Gamma_j\Big(W_{\mathcal{O}_j}^{(\ell)}\Big). \tag{123}$$
   Namely, π_{Z_j} is the coupling of $P_{Z_j}^{(1)}, \dots, P_{Z_j}^{(q)}$ induced by independently sampling $W_k^{(1:q)} \sim \pi_{W_k}$ for each $k \in \mathcal{O}_j$ and then computing (123) for each $\ell \in [q]$.
   b) Construct a coupling π_{W_j} of $P_{Y_j}^{(1)}, \dots, P_{Y_j}^{(q)}$ by applying Lemma 3 with input space $\mathcal{U} \triangleq \mathcal{Q}$, output space $\mathcal{V} \triangleq \mathbb{R}^d$, input coupling $\pi_U \triangleq \pi_{Z_j}$, output coupling $\pi_V \triangleq \pi_{W_j}$, and Φ being the circuit noise mechanism.

We analyze the coupling π_{W_m} constructed by the algorithm above to upper-bound $\rho\big([P_{Y_m}^{(1)}, \dots, P_{Y_m}^{(q)}]\big)$ using the maximal coupling characterization of Doeblin coefficients (Proposition 1). For notational convenience, for each $k \in \{-1, \dots, -n\} \cup [m]$, let

$$t_{W_k} \triangleq \mathop{\mathbb{P}}_{W_k^{(1:q)} \sim \pi_{W_k}} \Big(\neg\Big(W_k^{(1)} = \cdots = W_k^{(q)}\Big)\Big). \tag{124}$$

The value of $t_{W_{-i}}$ for each input $i \in [n]$ is

$$t_{W_{-1}} = 1\,, \tag{125}$$
$$\forall i > 1,\ \ t_{W_{-i}} = 0\,, \tag{126}$$

where (125) holds because $\pi_{W_{-1}}$ is a coupling of distinct Dirac measures $P_{X_1}^{(\ell)} = \delta_{\xi_\ell}$ for $\ell \in [q]$, and (126) holds because $\pi_{W_{-i}}$ for $i > 1$ is a coupling of identical Dirac measures $P_{X_i}^{(1)} = \cdots = P_{X_i}^{(q)} = \delta_{x_i}$. For each gate $j \in [m]$, t_{W_j} satisfies the recurrence (cf. [25, Eq. 149])

$$\begin{aligned}
t_{W_j} &\overset{(a)}{\le} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Bigg(\mathop{\mathbb{P}}_{Z_j^{(1:q)} \sim \pi_{Z_j}} \Big(\neg\Big(Z_j^{(1)} = \cdots = Z_j^{(q)}\Big)\Big); p\Bigg) \overset{(b)}{\le} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Bigg(\mathop{\mathbb{P}}_{W_{\mathcal{O}_j}^{(1:q)} \sim \bigotimes_{k \in \mathcal{O}_j} \pi_{W_k}} \Big(\neg\Big(W_{\mathcal{O}_j}^{(1)} = \cdots = W_{\mathcal{O}_j}^{(q)}\Big)\Big); p\Bigg) \\
&\overset{(c)}{\le} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Bigg(1 \wedge \sum_{k \in \mathcal{O}_j} \mathop{\mathbb{P}}_{W_k^{(1:q)} \sim \pi_{W_k}} \Big(\neg\Big(W_k^{(1)} = \cdots = W_k^{(q)}\Big)\Big); p\Bigg) \\
&\overset{(d)}{=} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Bigg(1 \wedge \sum_{k \in \mathcal{O}_j} t_{W_k}; p\Bigg) \overset{(e)}{\le} \hat{\mathrm{F}}_\Phi^{\mathsf{UA}}\Big(1 \wedge b \max_{k \in \mathcal{O}_j} t_{W_k}; p\Big) \overset{(f)}{=} f\Big(\max_{k \in \mathcal{O}_j} t_{W_k}\Big),
\end{aligned} \tag{127}$$

where (a) holds by (21); (b) holds by (123) because the gate operation is a deterministic function of its inputs, and because $\hat{\mathrm{F}}_\Phi^{\mathsf{UA}}$ is non-decreasing by Lemma 9, Part 2; (c) holds by the union bound; (d) holds by (124); (e) holds because $|\mathcal{O}_j| \le b$ since gates have at most b inputs; and (f) holds by definition of f.

For convenience, let $\mathcal{D} \subseteq [m]$ denote the set of gates which are descendants of X_1, i.e., there exists a directed path from X_1 to any gate $j \in \mathcal{D}$. Next, we will construct a directed path from X_1 to Y_m by searching backwards starting at Y_m, following the algorithm below (cf. [25, Section 5.3]):

[15] To see this, consider all circuits where the in-degree of each gate is at most b and the height of each input vertex is at most $h_{\max} \triangleq \lceil \log_b(n) - 1 \rceil$, and let C be such a circuit with the greatest number of input vertices. We may assume that no vertex in C has multiple outgoing edges, since we may equivalently consider the tree generated by running breadth-first traversal starting from Y_m and following edges in reverse direction, which preserves the heights of all input vertices. Furthermore, each gate vertex in C must have exactly b incoming edges; otherwise, additional input vertices may be added to the gate and hence to C. Lastly, each vertex in C with height less than $h_{\max}$ must be a gate, since if such a vertex was an input, replacing it with a gate taking b new input vertices would add $b-1$ input vertices to C overall. Hence, C is a perfect b-ary anti-arborescence where all input vertices have height $h_{\max}$, and so C has $b^{h_{\max}} < b^{\log_b(n)} = n$ inputs.

1) Let $j \in [m]$ denote the current gate vertex. Start at $j \triangleq m$.
2) If X_1 is an incoming neighbor of gate j (i.e., $1 \in \mathcal{N}_j$), move to X_1 and terminate.
3) Otherwise, find and move to any incoming gate neighbor $j' \in \mathcal{M}_j$ such that

$$j' \in \left(\arg\max_{j' \in \mathcal{M}_j} t_{W_{j'}}\right) \cap \mathcal{D}\,. \tag{128}$$

Set $j \triangleq j'$ and return to Step 2.

This algorithm is guaranteed to terminate at X_1, because the circuit has no directed cycles and the invariant $j \in \mathcal{D}$ holds throughout execution. (We start at the output gate $m \in \mathcal{D}$ in Step 1, and move to some gate $j' \in \mathcal{D}$ on each iteration of Step 3.) To establish well-definedness, it remains to show that the set in (128) is always non-empty if the algorithm did not terminate in Step 2 (i.e., $1 \notin \mathcal{N}_j$). Since $j \in \mathcal{D}$ and $1 \notin \mathcal{N}_j$, we must have $\mathcal{M}_j \neq \emptyset$. By (126) and (127) and the fact that $f(0) = \hat{\mathrm{F}}_{\Phi}^{\mathsf{UA}}(0;p) = \mathrm{F}_{\Phi}^{\mathsf{UA}}(0;p) = 0$ (where the second equality holds by Lemma 9, Part 1), it follows that any gate $j' \notin \mathcal{D}$ has $t_{W_{j'}} = 0$. Hence, $(\arg\max_{j' \in \mathcal{M}_j} t_{W_{j'}}) \not\subseteq \mathcal{D}^{\mathrm{c}}$, and so the set in (128) is non-empty, as desired.

Let $j_1 < \cdots < j_s = m$ be the sequence of gates from X_1 to Y_m constructed above, where s is the path length. We have $\max_{k \in \mathcal{O}_{j_1}} t_{W_k} = 1$, because $1 \in \mathcal{N}_{j_1}$ and $t_{W_{-1}} = 1$ by (125). For any $r \in \{2, \ldots, s\}$, we have

$$\max_{k \in \mathcal{O}_{j_r}} t_{W_k} \overset{(a)}{=} \max_{j' \in \mathcal{M}_{j_r}} t_{W_{j'}} \overset{(b)}{=} t_{W_{j_{r-1}}}\,,$$

where (a) holds because $1 \notin \mathcal{N}_{j_r}$ and so $t_{W_{-i}} = 0$ for all $i \in \mathcal{N}_{j_r}$ by (126), and (b) holds by Step 3 of the algorithm (128). Therefore,

$$\rho\big([P_{Y_m}^{(1)}, \ldots, P_{Y_m}^{(q)}]\big) \overset{(a)}{=} 1 - \sup_{\mathbb{P}: Y_m^{(\ell)} \sim P_{Y_m}^{(\ell)}} \mathbb{P}\Big(Y_m^{(1)} = \cdots = Y_m^{(q)}\Big) \overset{(b)}{\leq} 1 - \mathop{\mathbb{P}}_{Y_m^{(1:q)} \sim \pi_{W_m}}\Big(Y_m^{(1)} = \cdots = Y_m^{(q)}\Big) \overset{(c)}{=} t_{W_m} \overset{(d)}{\leq} f\big(t_{W_{j_{s-1}}}\big) \leq \cdots \overset{(e)}{\leq} f^{(s)}(1)\,,$$

where (a) holds by Proposition 1, where the supremum is taken over all couplings of $P_{Y_m}^{(1)}, \ldots, P_{Y_m}^{(q)}$; (b) holds by lower-bounding the supremum with a specific coupling π_{W_m}; (c) holds by definition of t_{W_k} (124); and (d) and (e) hold by repeatedly applying the recurrence (127) along with the above characterizations of $\max_{k \in \mathcal{O}_{j_r}}$. Since the height of X_1 is $\lceil \log_b(n) \rceil$ or greater, $s \geq \lceil \log_b(n) \rceil$. By (122), we have $\rho\big([P_{Y_m}^{(1)}, \ldots, P_{Y_m}^{(q)}]\big) \leq \max\{t \in [0,1] : f(t) = t\} + \epsilon$ as desired. □

We remark on an alternative way to guarantee that starting from Y_m and repeatedly moving to the gate maximizing $t_{W_{j'}}$ leads to a path terminating at X_1. For each $j \notin \mathcal{D}$, let N_j represent the randomness in the noise mechanism which produces Y_j from Z_j [74, Lemma 4.22]. Then,

$$\rho\big([P_{Y_m}^{(1)}, \ldots, P_{Y_m}^{(q)}]\big) \overset{(a)}{=} \rho\Bigg(\mathop{\mathbb{E}}_{\varphi_{\mathcal{D}^{\mathrm{c}}} \overset{\text{iid}}{\sim} \Phi}\Big[\big[P_{Y_m|N_{\mathcal{D}^{\mathrm{c}}}=\varphi_{\mathcal{D}^{\mathrm{c}}}}^{(1)}, \ldots, P_{Y_m|N_{\mathcal{D}^{\mathrm{c}}}=\varphi_{\mathcal{D}^{\mathrm{c}}}}^{(q)}\big]\Big]\Bigg) \overset{(b)}{\leq} \mathop{\mathbb{E}}_{\varphi_{\mathcal{D}^{\mathrm{c}}} \overset{\text{iid}}{\sim} \Phi}\Big[\rho\Big(\big[P_{Y_m|N_{\mathcal{D}^{\mathrm{c}}}=\varphi_{\mathcal{D}^{\mathrm{c}}}}^{(1)}, \ldots, P_{Y_m|N_{\mathcal{D}^{\mathrm{c}}}=\varphi_{\mathcal{D}^{\mathrm{c}}}}^{(q)}\big]\Big)\Big]$$

where (a) holds by the law of total probability, and (b) holds by convexity of complementary Doeblin coefficients [19, Theorem 1, Part 3] and Jensen's inequality. The marginal distributions in (b) correspond to analyzing the circuit where all gate vertices which are not descendants of X_1 are deterministic (after conditioning on $N_{\mathcal{D}^{\mathrm{c}}}$), and so may be removed from the circuit before proceeding with the coupling constructions. Since all directed paths in the circuit now start with X_1, the search algorithm is guaranteed to build a path back to X_1.

C. Proofs for Differential Privacy

In this section, we prove Theorems 5 and 6.

Proof of Theorem 5. Fix an absolutely continuous Markov kernel $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ with common dominating measure $\mu : \mathcal{F}_{\mathcal{Y}} \to \mathbb{R}_+$. Fix $n \geq 2$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}_+^n$ with $\epsilon_1 = 0$ and $\epsilon_i \leq \epsilon_j$ for all $i < j$.

Part 1: Fix a Markov kernel $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ from a Polish space $(\mathcal{U}, \mathcal{F}_{\mathcal{U}})$. We have

$$\rho_{\boldsymbol{\epsilon}}(WK, n) \overset{(a)}{=} 1 - \inf_{u_1,\ldots,u_n \in \mathcal{U}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} e^{\epsilon_i} WK(A_i \mid u_i) \overset{(b)}{=} 1 - \inf_{u_1,\ldots,u_n \in \mathcal{U}} \underbrace{\int_{\mathcal{Y}} \left(\min_{i \in [n]} e^{\epsilon_i} \frac{dWK}{d\mu}(\cdot \mid u_i)\right) d\mu}_{①}\,, \tag{129}$$

where (a) holds by definition of $\rho_{\boldsymbol{\epsilon}}$ (26) and (b) holds by Proposition 2.

Next, we lower-bound ①. Let $\nu : \mathcal{F}_{\mathcal{X}} \to \mathbb{R}_+$ given by $\nu(\cdot) \triangleq \bigwedge_{i \in [n]} e^{\epsilon_i} W(\cdot \mid u_i)$ be the weighted greatest common component of W restricted to $\{u_1, \ldots, u_n\}$. Let $\tau_0 \triangleq \nu(\mathcal{X})$ (for convenience, we elide the dependence of ν and τ_0 on $u_1, \ldots, u_n$ when denoting them). Notice that

$$\tau_0 \overset{(a)}{\leq} \min_{i \in [n]} e^{\epsilon_i} W(\mathcal{X} \mid u_i) \overset{(b)}{=} \min_{i \in [n]} e^{\epsilon_i} \overset{(c)}{=} 1\,, \tag{130}$$

where (a) holds by definition of greatest common component (3), (b) holds because W is a Markov kernel, and (c) holds because $\epsilon_1 = 0$ and $\epsilon_i \geq \epsilon_1$ for all $i > 1$. We consider two cases based on the value of τ_0.

Case 1: $\tau_0 = 1$. Proceeding from (129), we have

$$① \overset{(a)}{=} \int_{\mathcal{Y}} \left(\frac{d}{d\mu} \bigwedge_{i\in[n]} e^{\epsilon_i} WK(\cdot \mid u_i) \right) d\mu \overset{(b)}{=} \bigwedge_{i\in[n]} e^{\epsilon_i} WK(\mathcal{Y} \mid u_i)\,, \tag{131}$$

where (a) holds by Lemma 4 and because the lattice infimum of a finite family is the minimum, and (b) holds by the Radon-Nikodym theorem. For notational convenience, let $w_i : \mathcal{F}_{\mathcal{X}} \to [0,1]$ be the probability measure $w_i(\cdot) \triangleq W(\cdot \mid u_i)$. Observe that for any $i \in [n]$,

$$e^{\epsilon_i} WK(\cdot \mid u_i) = e^{\epsilon_i} w_i K(\cdot) = \left(e^{\epsilon_i} w_i - \nu\right) K(\cdot) + \nu K(\cdot)\,,$$

and so by Lemma 8,

$$\bigwedge_{i\in[n]} e^{\epsilon_i} WK(\cdot \mid u_i) = \bigwedge_{i\in[n]} \left(e^{\epsilon_i} w_i - \nu\right) K(\cdot) + \nu K(\cdot)\,. \tag{132}$$

Furthermore, it holds that $\nu = w_1$. To see this, first observe that $\nu \leq w_1$, because

$$\forall A \in \mathcal{F}_{\mathcal{X}},\ \ \nu(A) \overset{(a)}{\leq} e^{\epsilon_1} w_1(A) \overset{(b)}{=} w_1(A)\,,$$

where (a) holds by definition of greatest common component and (b) holds because $\epsilon_1 = 0$. Next, suppose for the sake of contradiction that $\nu(A) \neq w_1(A)$ for some $A \in \mathcal{F}_{\mathcal{X}}$. Then,

$$1 \overset{(a)}{=} \nu(\mathcal{X}) \overset{(b)}{=} \nu(A) + \nu(A^{\mathrm{c}}) \overset{(c)}{<} w_1(A) + \nu(A^{\mathrm{c}}) \overset{(d)}{\leq} w_1(A) + w_1(A^{\mathrm{c}}) \overset{(e)}{=} w_1(\mathcal{X}) \overset{(f)}{=} 1$$

and we obtain the contradiction $1 < 1$, where (a) holds by the assumption that $\tau_0 = 1$, (b) holds by additivity of measures, (c) holds by combining the supposition $\nu(A) \neq w_1(A)$ with the fact that $\nu \leq w_1$, (d) holds because $\nu \leq w_1$, (e) holds by additivity of measures, and (f) holds because w_1 is a probability measure. Finally, combining (132) with the fact that $\nu = w_1$, we have

$$\bigwedge_{i\in[n]} e^{\epsilon_i} WK(\cdot \mid u_i) = \nu K(\cdot)\,, \tag{133}$$

and combining (131) and (133), we have

$$① = \nu K(\mathcal{Y}) \overset{(a)}{=} (1 - \tau_0)\, \tau_{\boldsymbol{\epsilon}}(K, n) + \nu K(\mathcal{Y})\,, \tag{134}$$

where (a) holds by the assumption that $\tau_0 = 1$.

Case 2: $\tau_0 < 1$. Define a Markov kernel $V : \{u_1, \ldots, u_n\} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ given by

$$\forall i \in [n],\ \ V(\cdot \mid u_i) \triangleq \frac{e^{\epsilon_i} W(\cdot \mid u_i) - \nu(\cdot)}{e^{\epsilon_i} - \tau_0}\,. \tag{135}$$

This is a valid Markov kernel because for any $i \in [n]$, $V(\mathcal{X} \mid u_i) = \frac{e^{\epsilon_i} W(\mathcal{X}|u_i) - \nu(\mathcal{X})}{e^{\epsilon_i} - \tau_0} = \frac{e^{\epsilon_i} - \tau_0}{e^{\epsilon_i} - \tau_0} = 1$. By rearranging (135), W may be written in terms of V and ν as $e^{\epsilon_i} W(\cdot \mid u_i) = (e^{\epsilon_i} - \tau_0)\, V(\cdot \mid u_i) + \nu(\cdot)$, and therefore the composition WK may be written as

$$e^{\epsilon_i} WK(\cdot \mid u_i) = (e^{\epsilon_i} - \tau_0)\, VK(\cdot \mid u_i) + \nu K(\cdot)\,. \tag{136}$$

Proceeding from (129), we have

$$① \overset{(a)}{=} \int_{\mathcal{Y}} \min_{i\in[n]} \left\{ (e^{\epsilon_i} - \tau_0) \frac{dVK}{d\mu}(\cdot \mid u_i) + \frac{d\nu K}{d\mu} \right\} d\mu \overset{(b)}{=} \underbrace{\int_{\mathcal{Y}} \left(\min_{i\in[n]} (e^{\epsilon_i} - \tau_0) \frac{dVK}{d\mu}(\cdot \mid u_i) \right) d\mu}_{②} + \int_{\mathcal{Y}} \frac{d\nu K}{d\mu}\, d\mu\,, \tag{137}$$

where (a) holds by differentiating both sides of (136) and (b) holds by linearity of integration. Next, we lower-bound ②. We have

$$\begin{aligned}
② &\overset{(a)}{=} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} (e^{\epsilon_i} - \tau_0)\, VK(A_i \mid u_i) \overset{(b)}{=} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} (e^{\epsilon_i} - \tau_0) \int_{\mathcal{X}} K(A_i \mid x)\, V(dx \mid u_i) \\
&\overset{(c)}{\geq} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} (e^{\epsilon_i} - \tau_0) \inf_{x_i \in \mathcal{X}} K(A_i \mid x_i) \overset{(d)}{=} \inf_{x_1,\ldots,x_n \in \mathcal{X}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} (e^{\epsilon_i} - \tau_0)\, K(A_i \mid x_i) \\
&\overset{(e)}{=} (1 - \tau_0) \inf_{x_1,\ldots,x_n \in \mathcal{X}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} \frac{e^{\epsilon_i} - \tau_0}{1 - \tau_0} K(A_i \mid x_i) \overset{(f)}{\geq} (1 - \tau_0) \inf_{x_1,\ldots,x_n \in \mathcal{X}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} e^{\epsilon_i} K(A_i \mid x_i)
\end{aligned}$$

$$\overset{(g)}{=} (1-\tau_0)\,\tau_{\boldsymbol{\epsilon}}(K,n)\,, \tag{138}$$

where (a) holds by Proposition 2, (b) holds by definition of kernel composition (8), (c) holds by replacing a weighted average with an infimum, (d) holds by interchanging the order of infimums, (e) holds because $1-\tau_0>0$, (f) holds because $\epsilon_i \geq 0$ and so $\frac{e^{\epsilon_i}-\tau_0}{1-\tau_0} \geq \frac{e^{\epsilon_i}-e^{\epsilon_i}\tau_0}{1-\tau_0} = e^{\epsilon_i}$ for each $i \in [n]$, and (g) holds by definition of $\tau_{\boldsymbol{\epsilon}}$ (25). Combining (137) and (138), we have

$$① \geq (1-\tau_0)\,\tau_{\boldsymbol{\epsilon}}(K,n) + \int_{\mathcal{Y}} \frac{d\nu K}{d\mu}\,d\mu \overset{(a)}{=} (1-\tau_0)\,\tau_{\boldsymbol{\epsilon}}(K,n) + \nu K(\mathcal{Y})\,, \tag{139}$$

where (a) holds by the Radon-Nikodym theorem.

Proceeding from both cases: Now, by (134) and (139), we have the same lower bound for ① in both cases. For notational convenience, let $\hat{\nu}: \mathcal{F}_{\mathcal{X}} \to \mathbb{R}_+$ be the measure given by

$$\hat{\nu}(\cdot) \triangleq \begin{cases} \frac{\nu(\cdot)}{\tau_0}\,, & \text{if } \tau_0 > 0\,, \\ 0\,, & \text{if } \tau_0 = 0\,. \end{cases}$$

Proceeding onwards,

$$① \geq (1-\tau_0)\,\tau_{\boldsymbol{\epsilon}}(K,n) + \tau_0\,\hat{\nu}K(\mathcal{Y}) \overset{(a)}{=} (1-\tau_0)\,\tau_{\boldsymbol{\epsilon}}(K,n) + \tau_0\,, \tag{140}$$

where (a) holds in the case $\tau_0 > 0$ because $\hat{\nu}$ is a probability measure on $\mathcal{F}_{\mathcal{X}}$ in this case and so $\hat{\nu}K$ is a probability measure on $\mathcal{F}_{\mathcal{Y}}$. Combining (129) and (140), we have

$$\rho_{\boldsymbol{\epsilon}}(WK,n) \leq 1 - \inf_{u_1,\ldots,u_n \in \mathcal{U}} \{(1-\tau_0)\,\tau_{\boldsymbol{\epsilon}}(K,n) + \tau_0\} = \Bigg(1 - \underbrace{\inf_{u_1,\ldots,u_n \in \mathcal{U}} \tau_0}_{③}\Bigg)\,(1-\tau_{\boldsymbol{\epsilon}}(K,n))\,. \tag{141}$$

Next, we evaluate ③. Given any $u_1,\ldots,u_n \in \mathcal{U}$, let $\xi: \mathcal{F}_{\mathcal{X}} \to \mathbb{R}_+$ be the measure $\xi(\cdot) \triangleq \sum_{i=1}^n W(\cdot \mid u_i)$, where we elide the dependence on $u_1,\ldots,u_n$ in the notation ξ for convenience. We have

$$\begin{aligned} ③ &\overset{(a)}{=} \inf_{u_1,\ldots,u_n \in \mathcal{U}} \bigwedge_{i\in[n]} e^{\epsilon_i} W(\mathcal{X} \mid u_i) \overset{(b)}{=} \inf_{u_1,\ldots,u_n \in \mathcal{U}} \int_{\mathcal{X}} \left(\frac{d}{d\xi} \bigwedge_{i\in[n]} e^{\epsilon_i} W(\cdot \mid u_i)\right) d\xi \\ &\overset{(c)}{=} \inf_{u_1,\ldots,u_n \in \mathcal{U}} \int_{\mathcal{X}} \left(\min_{i\in[n]} e^{\epsilon_i} \frac{dW}{d\xi}(\cdot \mid u_i)\right) d\xi \overset{(d)}{=} \inf_{u_1,\ldots,u_n \in \mathcal{U}} \inf_{\substack{n\text{-partition of } \mathcal{X} \\ A_1,\ldots,A_n}} \sum_{i=1}^n e^{\epsilon_i} W(A_i \mid u_i) \overset{(e)}{=} \tau_{\boldsymbol{\epsilon}}(W,n)\,, \end{aligned} \tag{142}$$

where (a) holds by definition of τ_0, (b) holds by the Radon-Nikodym theorem because $\bigwedge_{i\in[n]} e^{\epsilon_i} W(\cdot \mid u_i) \ll \xi$, (c) holds by Lemma 4 and because the lattice infimum of a finite family is the minimum, (d) holds by Proposition 2, and (e) holds by definition of $\tau_{\boldsymbol{\epsilon}}$ (25). Combining (141) and (142), we have

$$\rho_{\boldsymbol{\epsilon}}(WK,n) \leq (1-\tau_{\boldsymbol{\epsilon}}(W,n))\,(1-\tau_{\boldsymbol{\epsilon}}(K,n)) \overset{(a)}{=} \rho_{\boldsymbol{\epsilon}}(W,n)\,\rho_{\boldsymbol{\epsilon}}(K,n)$$

as desired, where (a) holds by definition of $\rho_{\boldsymbol{\epsilon}}$ (26).

Part 2: First, observe that

$$K \text{ is } (\boldsymbol{\epsilon},\delta,n)\text{-LDP} \overset{(a)}{\iff} \inf_{x_1,\ldots,x_n \in \mathcal{X}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^n e^{\epsilon_i} K(A_i \mid x_i) \geq 1-\delta \overset{(b)}{\iff} \tau_{\boldsymbol{\epsilon}}(K,n) \geq 1-\delta \overset{(c)}{\iff} \rho_{\boldsymbol{\epsilon}}(K,n) \leq \delta\,, \tag{143}$$

where (a) holds by definition of $(\boldsymbol{\epsilon},\delta,n)$-LDP (24), (b) holds by definition of $\tau_{\boldsymbol{\epsilon}}$ (25), and (c) holds by subtracting both sides of the inequality from 1. We note that (143) generalizes [42, Theorem 1].

With (143) in mind, observe also that

$$\rho_{\boldsymbol{\epsilon}}(K,n) \leq \delta \implies \forall W,\ \rho_{\boldsymbol{\epsilon}}(WK,n) \leq \delta \rho_{\boldsymbol{\epsilon}}(W,n)\,, \tag{144}$$

because for any Markov kernel $W: \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ from any Polish space $(\mathcal{U},\mathcal{F}_{\mathcal{U}})$, we have

$$\rho_{\boldsymbol{\epsilon}}(WK,n) \overset{(a)}{\leq} \rho_{\boldsymbol{\epsilon}}(W,n)\,\rho_{\boldsymbol{\epsilon}}(K,n) \overset{(b)}{\leq} \delta \rho_{\boldsymbol{\epsilon}}(W,n)\,,$$

where (a) follows by Part 1 and (b) follows by the antecedent in (144). Next, we will show that the converse

$$\Big(\forall W,\ \rho_{\boldsymbol{\epsilon}}(WK,n) \leq \delta \rho_{\boldsymbol{\epsilon}}(W,n)\Big) \implies \rho_{\boldsymbol{\epsilon}}(K,n) \leq \delta \tag{145}$$

also holds. Note that if $\tau_{\boldsymbol{\epsilon}}(K,n)=1$, the consequent in (145) is trivially satisfied, because $\rho_{\boldsymbol{\epsilon}}(K,n)=1-\tau_{\boldsymbol{\epsilon}}(K,n)=0\le\delta$. Hence, we assume $\tau_{\boldsymbol{\epsilon}}(K,n)<1$ for the remainder of this argument. By definition of $\tau_{\boldsymbol{\epsilon}}$ as an infimum in (25), for any arbitrary $0<\gamma<1-\tau_{\boldsymbol{\epsilon}}(K,n)$, there exists $x_1^*,\dots,x_n^*\in\mathcal{X}$ (possibly depending on γ) such that

$$\inf_{\substack{n\text{-partition of }\mathcal{Y}\\ A_1,\dots,A_n}}\sum_{i=1}^n e^{\epsilon_i}K(A_i\mid x_i^*)\le\tau_{\boldsymbol{\epsilon}}(K,n)+\gamma\,. \tag{146}$$

Furthermore, there must be at least two distinct values within $(x_1^*,\dots,x_n^*)$, because if we suppose the contrary that $x_1^*=\cdots=x_n^*=x^*$ for some $x^*\in\mathcal{X}$, we obtain the contradiction

$$\inf_{\substack{n\text{-partition of }\mathcal{Y}\\ A_1,\dots,A_n}}\sum_{i=1}^n e^{\epsilon_i}K(A_i\mid x_i^*)=\inf_{\substack{n\text{-partition of }\mathcal{Y}\\ A_1,\dots,A_n}}\sum_{i=1}^n e^{\epsilon_i}K(A_i\mid x^*)\overset{(a)}{=}e^{\epsilon_1}K(\mathcal{Y}\mid x^*)\overset{(b)}{=}1>\tau_{\boldsymbol{\epsilon}}(K,n)+\gamma\,,$$

where (a) holds because $\epsilon_i\ge\epsilon_1$ for all $i>1$, and (b) holds because $\epsilon_1=0$ and K is a Markov kernel. Let $W^*:[n]\times\mathcal{F}_{\mathcal{X}}\to[0,1]$ be a Markov kernel such that $W^*(\cdot\mid i)\triangleq\delta_{x_i^*}(\cdot)$ for all $i\in[n]$. Then,

$$\begin{aligned}\rho_{\boldsymbol{\epsilon}}(K,n)&\overset{(a)}{=}1-\tau_{\boldsymbol{\epsilon}}(K,n)\overset{(b)}{\le}1-\inf_{\substack{n\text{-partition of }\mathcal{Y}\\ A_1,\dots,A_n}}\sum_{i=1}^n e^{\epsilon_i}K(A_i\mid x_i^*)+\gamma\overset{(c)}{=}1-\inf_{\substack{n\text{-partition of }\mathcal{Y}\\ A_1,\dots,A_n}}\sum_{i=1}^n e^{\epsilon_i}W^*K(A_i\mid i)+\gamma\\&\overset{(d)}{\le}1-\inf_{u_1,\dots,u_n\in[n]}\inf_{\substack{n\text{-partition of }\mathcal{Y}\\ A_1,\dots,A_n}}\sum_{i=1}^n e^{\epsilon_i}W^*K(A_i\mid u_i)+\gamma\overset{(e)}{=}\rho_{\boldsymbol{\epsilon}}(W^*K,n)+\gamma\overset{(f)}{\le}\delta\underbrace{\rho_{\boldsymbol{\epsilon}}(W^*,n)}_{①}+\gamma\,,\end{aligned} \tag{147}$$

where (a) holds by definition of $\rho_{\boldsymbol{\epsilon}}$ (26), (b) holds by (146), (c) holds by definition of W^* and because each $\{x_i^*\}$ is measurable (since $\mathcal{X}$ is Polish), (d) holds by lower-bounding the value of a particular instance with an infimum, (e) holds by definition of $\rho_{\boldsymbol{\epsilon}}$ (26), and (f) holds by the antecedent in (145). Next, we evaluate ①. Let $s,t\in[n]$ be indices such that $x_s^*\neq x_t^*$. We have

$$\begin{aligned}①&\overset{(a)}{=}1-\inf_{u_1,\dots,u_n\in\mathcal{U}}\inf_{\substack{n\text{-partition of }\mathcal{X}\\ A_1,\dots,A_n}}\sum_{i=1}^n e^{\epsilon_i}W^*(A_i\mid u_i)\overset{(b)}{=}1-\Big(e^{\epsilon_1}W^*(\mathcal{X}-\{x_s^*\}\mid s)+e^{\epsilon_2}W^*(\{x_s^*\}\mid t)\Big)\\&\overset{(c)}{=}1-\Big(e^{\epsilon_1}\delta_{x_s^*}(\mathcal{X}-\{x_s^*\})+e^{\epsilon_2}\delta_{x_t^*}(\{x_s^*\})\Big)\overset{(d)}{=}1\,,\end{aligned} \tag{148}$$

where (a) holds by definition of $\rho_{\boldsymbol{\epsilon}}$ (26); the events $\mathcal{X}-\{x_s^*\}$ and $\{x_s^*\}$ in (b) are measurable because $\mathcal{X}$ is Polish, and so $\{x\}$ is measurable for any $x\in\mathcal{X}$; (b) follows by replacing the infima with a specific instance, and equality holds because $\sum_{i=1}^n e^{\epsilon_i}W^*(A_i\mid u_i)\ge 0$ for all $u_1,\dots,u_n$ and $A_1,\dots,A_n$; (c) holds by definition of W^*; and (d) holds by definition of Dirac measure. Since $\gamma>0$ was arbitrary, combining (147) and (148) shows that $\rho_{\boldsymbol{\epsilon}}(K,n)\le\delta$, thus proving (145).

Finally, combining Equations (143) to (145), we have

$$K\text{ is }(\boldsymbol{\epsilon},\delta,n)\text{-LDP}\iff\forall W,\ \rho_{\boldsymbol{\epsilon}}(WK,n)\le\delta\rho_{\boldsymbol{\epsilon}}(W,n)$$

as desired.

Part 3: Fix a Markov kernel $W:\mathcal{U}\times\mathcal{F}_{\mathcal{X}}\to[0,1]$. Following the argument from (129), we have

$$\rho_{\boldsymbol{\epsilon}}(WK,n)\overset{(a)}{=}1-\inf_{u_1,\dots,u_n\in\mathcal{U}}\underbrace{\int_{\mathcal{Y}}\left(\min_{i\in[n]}e^{\epsilon_i}\frac{dWK}{d\mu}(\cdot\mid u_i)\right)d\mu}_{①}\,. \tag{149}$$

Note that if $\inf_{u_1,\dots,u_n\in\mathcal{U}}①=1$, the result (28) trivially follows because

$$\rho_{\boldsymbol{\epsilon}}(WK,n)=0\overset{(a)}{\le}\mathrm{F}_K^{\mathsf{UA}}(\rho_{\boldsymbol{\epsilon}}(W,n);p)\,,$$

where (a) holds by the range of $\mathrm{F}_K^{\mathsf{UA}}$. Hence, assume $\inf_{u_1,\dots,u_n\in\mathcal{U}}①<1$ for the remainder of this argument. Next, we lower-bound ①. Let $\nu:\mathcal{F}_{\mathcal{X}}\to\mathbb{R}_+$ given by $\nu(\cdot)\triangleq\bigwedge_{i\in[n]}e^{\epsilon_i}W(\cdot\mid u_i)$ be the weighted greatest common component of W restricted to $\{u_1,\dots,u_n\}$. Let $\tau_0\triangleq\nu(\mathcal{X})$ (as before, we elide the dependence of ν and τ_0 on $u_1,\dots,u_n$ when denoting them). Following the argument from (130), we have $\tau_0\le 1$. We consider two cases based on the value of τ_0.

Case 1: $\tau_0=1$. We have

$$①\overset{(a)}{=}\nu K(\mathcal{Y})\overset{(b)}{=}1\,, \tag{150}$$

where (a) holds by following the arguments from (131) and (133), and (b) holds because ν is a probability measure on $\mathcal{F}_{\mathcal{X}}$ (since $\nu(\mathcal{X})=\tau_0=1$) and so νK is a probability measure on $\mathcal{F}_{\mathcal{Y}}$.

Case 2: $\tau_0<1$. Define a Markov kernel $V:\{u_1,\dots,u_n\}\times\mathcal{F}_{\mathcal{X}}\to[0,1]$ given by

$$\forall i\in[n],\quad V(\cdot\mid u_i)\triangleq\frac{e^{\epsilon_i}W(\cdot\mid u_i)-\nu(\cdot)}{e^{\epsilon_i}-\tau_0}\,. \tag{151}$$

Following the argument from (137), we have

$$① = \underbrace{\int_{\mathcal{Y}} \left(\min_{i\in[n]} \left(e^{\epsilon_i} - \tau_0\right) \frac{dVK}{d\mu}(\cdot \mid u_i) \right) d\mu}_{②} + \nu K(\mathcal{Y})\,. \tag{152}$$

Next, we lower-bound ②. We have

$$② \overset{(a)}{\geq} (1-\tau_0) \int_{\mathcal{Y}} \left(\min_{i\in[n]} \frac{dVK}{d\mu}(\cdot \mid u_i) \right) d\mu \overset{(b)}{=} (1-\tau_0) \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} VK(A_i \mid u_i) \overset{(c)}{\geq} (1-\tau_0)\,\tau(VK)\,, \tag{153}$$

where (a) holds because $\epsilon_i \geq 0$ for all $i \in [n]$, (b) holds by Proposition 2, and (c) holds by Theorem 1. For notational convenience, let $\hat{\nu} : \mathcal{F}_{\mathcal{X}} \to \mathbb{R}_+$ be the measure given by

$$\hat{\nu}(\cdot) \triangleq \begin{cases} \frac{\nu(\cdot)}{\tau_0}\,, & \text{if } \tau_0 > 0\,, \\ 0\,, & \text{if } \tau_0 = 0\,. \end{cases}$$

Combining (152) and (153),

$$① \geq (1-\tau_0)\,\tau(VK) + \tau_0\,\hat{\nu}K(\mathcal{Y}) \overset{(a)}{=} (1-\tau_0)\,\tau(VK) + \tau_0\,, \tag{154}$$

where (a) holds in the case $\tau_0 > 0$ because $\hat{\nu}$ is a probability measure on $\mathcal{F}_{\mathcal{X}}$ in this case and so $\hat{\nu}K$ is a probability measure on $\mathcal{F}_{\mathcal{Y}}$.

Proceeding from both cases: Now, by (150) and (154), we have lower bounds on ① for any value of τ_0. Proceeding from (149), we have

$$\begin{aligned} \rho_{\boldsymbol{\epsilon}}(WK,n) &\overset{(a)}{=} 1 - \inf_{\substack{u_1,\ldots,u_n\in\mathcal{U}: \\ \tau_0<1}} ① \overset{(b)}{\leq} 1 - \inf_{\substack{u_1,\ldots,u_n\in\mathcal{U}: \\ \tau_0<1}} \left\{(1-\tau_0)\,\tau(VK) + \tau_0\right\} \overset{(c)}{\leq} 1 - \inf_{u_1,\ldots,u_n\in\mathcal{U}} \left\{(1-\tau_0)\,\tau(VK) + \tau_0\right\} \\ &= \sup_{u_1,\ldots,u_n\in\mathcal{U}} \left\{(1-\tau_0)\,\rho(VK)\right\} \overset{(d)}{\leq} \sup_{u_1,\ldots,u_n\in\mathcal{U}} \left\{(1-\tau_0)\,\mathrm{F}_K^{\mathsf{UA}}(\rho(V); |||V|||_{\mathsf{UA}})\right\} \overset{(e)}{\leq} \sup_{u_1,\ldots,u_n\in\mathcal{U}} \left\{(1-\tau_0)\,\mathrm{F}_K^{\mathsf{UA}}(1; |||V|||_{\mathsf{UA}})\right\}, \end{aligned} \tag{155}$$

where (a) holds because the infimum $\inf_{u_1,\ldots,u_n\in\mathcal{U}} ① < 1$ is not achieved by any $(u_1,\ldots,u_n)$ for which $\tau_0 = 1$, by (150); (b) holds by (154); (c) holds because an infimum does not increase when taken over a larger feasible set; (d) holds by definition of Doeblin curve (9); and (e) holds because Doeblin curves are non-decreasing in their first argument. The uniform average power of V may be bounded as

$$\begin{aligned} |||V|||_{\mathsf{UA}} &\overset{(a)}{=} \max_{i\in[n]} \int_{\mathcal{X}} M(\|x\|)\,V(dx \mid u_i) \overset{(b)}{=} \max_{i\in[n]} \int_{\mathcal{X}} M(\|x\|)\,\frac{e^{\epsilon_i} W(dx \mid u_i) - \nu(dx)}{e^{\epsilon_i} - \tau_0} \\ &\leq \max_{i\in[n]} \int_{\mathcal{X}} M(\|x\|)\,\frac{e^{\epsilon_i}}{e^{\epsilon_i} - \tau_0}\,W(dx \mid u_i) \overset{(c)}{\leq} p \max_{i\in[n]} \frac{e^{\epsilon_i}}{e^{\epsilon_i} - \tau_0} \overset{(d)}{=} \frac{p}{1-\tau_0}\,, \end{aligned} \tag{156}$$

where (a) holds by definition of uniform average power (12); (b) holds by definition of V (151); (c) holds because W satisfies $|||W|||_{\mathsf{UA}} \leq p$; and (d) holds because $s \mapsto s/(s-\tau_0)$ is decreasing for $s > \tau_0$, and $e^{\epsilon_i} \geq 1 > \tau_0$ for each $i \in [n]$. Combining (155) and (156),

$$\begin{aligned} \rho_{\boldsymbol{\epsilon}}(WK,n) &\leq \sup_{u_1,\ldots,u_n\in\mathcal{U}} \left\{(1-\tau_0)\,\mathrm{F}_K^{\mathsf{UA}}\left(1; \frac{p}{1-\tau_0}\right)\right\} \overset{(a)}{=} \sup_{u_1,\ldots,u_n\in\mathcal{U}} \mathrm{F}_K^{\mathsf{UA}}(1-\tau_0; p) \\ &\overset{(b)}{=} \mathrm{F}_K^{\mathsf{UA}}\left(\sup_{u_1,\ldots,u_n\in\mathcal{U}} \{1-\tau_0\}; p\right) \overset{(c)}{=} \mathrm{F}_K^{\mathsf{UA}}(1-\tau_{\boldsymbol{\epsilon}}(W,n); p) \overset{(d)}{=} \mathrm{F}_K^{\mathsf{UA}}(\rho_{\boldsymbol{\epsilon}}(W,n); p) \end{aligned}$$

as desired, where (a) holds by Proposition 5, (b) holds because Doeblin curves are non-decreasing in their first argument, (c) holds by following the argument in (142), and (d) holds by definition of $\rho_{\boldsymbol{\epsilon}}$ (26). □

Next, we prove Theorem 6.

Proof of Theorem 6. For each $t \in [T]$ and each initialization $i \in [n]$, define the following random variables representing intermediate computations within the update rule:

$$U_t^{(i)} \triangleq W_{t-1}^{(i)} - \eta_t \nabla g_t\big(W_{t-1}^{(i)}\big)\,,\quad V_t^{(i)} \triangleq U_t^{(i)} + m_t N_t\,,\quad W_t^{(i)} \triangleq \operatorname*{proj}_{\mathcal{W}}\big(V_t^{(i)}\big)\,.$$

Clearly, the random variables form the Markov chain (similar to the proof of Lemma 2 and [41], [67])

$$W_0^{(i)} \to \underline{U_1^{(i)} \to V_1^{(i)} \to W_1^{(i)}} \to \cdots \to \underline{U_T^{(i)} \to V_T^{(i)} \to W_T^{(i)}}$$

for each $i \in [n]$. We have

$$\rho_{\boldsymbol{\epsilon}}\big([P_{W_T}^{(1)}, \ldots, P_{W_T}^{(n)}], n\big) \overset{(a)}{\leq} \rho_{\boldsymbol{\epsilon}}\big([P_{V_T}^{(1)}, \ldots, P_{V_T}^{(n)}], n\big) \overset{(b)}{=} \rho_{\boldsymbol{\epsilon}}\big([P_{U_T}^{(1)} * \mathsf{Normal}(0, m_T^2 I), \ldots, P_{U_T}^{(n)} * \mathsf{Normal}(0, m_T^2 I)], n\big)\,, \quad (157)$$

where (a) holds by the data processing inequality for $\rho_{\boldsymbol{\epsilon}}$ (29), and (b) holds by definition of V_T. The power of the input distributions is bounded as

$$\begin{aligned}\Big|\Big|\Big|[P_{U_T}^{(1)}, \ldots, P_{U_T}^{(n)}]\Big|\Big|\Big|_{\mathsf{UA}} &\overset{(a)}{=} \max_{i\in[n]} \mathbb{E}\left[\|U_T^{(i)}\|_2^2\right] \overset{(b)}{=} \max_{i\in[n]} \mathbb{E}\left[\left\|W_{T-1}^{(i)} - \eta_T \nabla g_T\big(W_{T-1}^{(i)}\big)\right\|_2^2\right] \\ &\overset{(c)}{\leq} \max_{i\in[n]} \left\{2\mathbb{E}\left[\|W_{T-1}^{(i)}\|_2^2\right] + 2\mathbb{E}\left[\left\|\eta_T \nabla g_T\big(W_{T-1}^{(i)}\big)\right\|_2^2\right]\right\} \overset{(d)}{\leq} 2\max_{i\in[n]} \mathbb{E}\left[\|W_{T-1}^{(i)}\|_2^2\right] + 2\eta_T^2 L^2 \overset{(e)}{=} m_T^2 p_T\,,\end{aligned}$$

where (a) holds by definition of uniform average power (12), (b) holds by definition of U_T, (c) holds by the identity $\|x+y\|_2^2 \leq \|x+y\|_2^2 + \|x-y\|_2^2 = 2\|x\|_2^2 + 2\|y\|_2^2$ and linearity of expectation, (d) holds by the bound on the objective gradients (31), and (e) holds by definition of p_T (32). Continuing from (157), we have

$$\rho_{\boldsymbol{\epsilon}}\big([P_{W_T}^{(1)}, \ldots, P_{W_T}^{(n)}], n\big) \overset{(a)}{\leq} \mathrm{F}_\Phi^{\mathsf{UA}}\big(\rho_{\boldsymbol{\epsilon}}\big([P_{U_T}^{(1)}, \ldots, P_{U_T}^{(n)}], n\big); p_T\big) \overset{(b)}{\leq} \mathrm{F}_\Phi^{\mathsf{UA}}\big(\rho_{\boldsymbol{\epsilon}}\big([P_{W_{T-1}}^{(1)}, \ldots, P_{W_{T-1}}^{(n)}], n\big); p_T\big)\,, \quad (158)$$

where (a) holds by Theorem 5 and Proposition 6, and (b) holds by the data processing inequality for $\rho_{\boldsymbol{\epsilon}}$ (29). Finally, by recursively applying the arguments in (157) and (158) above, we obtain

$$\rho_{\boldsymbol{\epsilon}}\big([P_{W_T}^{(1)}, \ldots, P_{W_T}^{(n)}], n\big) \leq \mathrm{F}_\Phi^{\mathsf{UA}}\big(\cdots \mathrm{F}_\Phi^{\mathsf{UA}}\big(\rho_{\boldsymbol{\epsilon}}\big([P_{W_0}^{(1)}, \ldots, P_{W_0}^{(n)}], n\big); p_1\big) \cdots ; p_T\big)$$

as desired. □

VI. Conclusion

In closing, we review our main contributions and propose some directions for follow-up work. In this paper, we formulated the notion of a Doeblin curve to quantify information contraction of Markov kernels on collections of arbitrarily many input distributions with specific divergence and power levels, building upon existing literature on Doeblin coefficients and nonlinear information contraction. After introducing a new variational characterization of Doeblin coefficients, we established several properties of Doeblin curves and derived bounds on Doeblin curves under canonical power constraints and regularity conditions. With this more nuanced measure of information contraction in place, we presented three theoretical applications in noisy iterative optimization, reliable computation, and differential privacy, leveraging Doeblin curves to generalize results in these areas to multi-way or unbounded settings where Doeblin coefficients fail to capture information contraction.

We suggest two potential extensions of our present work. Firstly, our current understanding of Doeblin curves may be further enriched by establishing precise conditions for concavity and by studying examples of closed-form uniform average Doeblin curves. Secondly, future research may explore and broaden our proposed definition of group differential privacy (24), since the standard definition of differential privacy is sometimes considered very stringent for many applications [75], [76].

Appendix A
Technical Lemmata

In this appendix, we state and prove two miscellaneous results used throughout our paper.

Lemma 8 (Greatest Common Component Under Affine Transformation)**.** *For any kernel $W : \mathcal{U} \times \mathcal{F}_\mathcal{X} \to \mathbb{R}_+$ and signed measure $\pi : \mathcal{F}_\mathcal{X} \to \mathbb{R}$, the greatest common component of the kernel $\hat{W} : \mathcal{U} \times \mathcal{F}_\mathcal{X} \to \mathbb{R}_+$ given by*

$$\forall u \in \mathcal{U},\ \ \hat{W}(\cdot \mid u) \triangleq \alpha W(\cdot \mid u) + \pi(\cdot)\,, \quad (159)$$

where $\alpha \geq 0$ and π are such that $\hat{W} \geq 0$, is

$$\bigwedge_{u\in\mathcal{U}} \hat{W}(\cdot \mid u) = \alpha \bigwedge_{u\in\mathcal{U}} W(\cdot \mid u) + \pi(\cdot)\,. \quad (160)$$

Proof. The lemma trivially holds for $\alpha = 0$, so assume $\alpha > 0$ henceforth. For notational convenience, let μ^* denote the measure given by the right-hand side of (160). By definition of greatest common component as a supremum in (3), we seek to verify two properties of μ^*.

Firstly, μ^* satisfies the condition of the supremum $\hat{W} \geq \mu^*$, because for any $u \in \mathcal{U}$,

$$\hat{W}(\cdot \mid u) \overset{(a)}{=} \alpha W(\cdot \mid u) + \pi(\cdot) \overset{(b)}{\geq} \alpha \bigwedge_{u\in\mathcal{U}} W(\cdot \mid u) + \pi(\cdot) \overset{(c)}{=} \mu^*(\cdot)\,,$$

where (a) holds by definition of $\hat{W}$ (159), (b) holds because the greatest common component of a kernel is a lower bound on the kernel, and (c) holds by definition of μ^* as the right-hand side of (160).

Secondly, we will prove that for any measure μ satisfying $\hat{W} \geq \mu$, we have $\mu^* \geq \mu$. Fix a measure μ such that $\hat{W} \geq \mu$. Observe that

$$\forall u \in \mathcal{U},\ \ W(\cdot \mid u) \overset{(a)}{=} \frac{\hat{W}(\cdot \mid u) - \pi(\cdot)}{\alpha} \geq \frac{\mu(\cdot) - \pi(\cdot)}{\alpha}\,, \tag{161}$$

where (a) holds by rearranging (159). Consider a Hahn-Jordan decomposition of $\mu - \pi$, i.e.,

$$\forall A \in \mathcal{F}_{\mathcal{X}},\ \ (\mu-\pi)^+(A) \triangleq (\mu-\pi)(A \cap P)\,,\ \ (\mu-\pi)^-(A) \triangleq -(\mu-\pi)(A \cap N)\,, \tag{162}$$

where $P \subseteq \mathcal{X}$ and $N = P^{\mathrm{c}}$ are the supports (modulo null sets) of the positive and negative parts of $\mu - \pi$, respectively. Then,

$$W(A \mid u) \overset{(a)}{=} W(A \cap P \mid u) + W(A \cap N \mid u) \overset{(b)}{\geq} \frac{(\mu-\pi)(A \cap P)}{\alpha} + 0 \overset{(c)}{=} \frac{(\mu-\pi)^+(A)}{\alpha} \tag{163}$$

for all $u \in \mathcal{U}$ and $A \in \mathcal{F}_{\mathcal{X}}$, where (a) holds by additivity of measures because $N = P^{\mathrm{c}}$, (b) holds because $W \geq (\mu - \pi)/\alpha$ (by (161)) and $W \geq 0$, and (c) holds by (162). Now, suppose for the sake of contradiction that there exists $A \in \mathcal{F}_{\mathcal{X}}$ such that $\mu^*(A) < \mu(A)$. Then,

$$\bigwedge_{u \in \mathcal{U}} W(A \mid u) \overset{(a)}{=} \sup\{\nu(A) : W \geq \nu\} \overset{(b)}{\geq} \frac{(\mu-\pi)^+(A)}{\alpha} \geq \frac{\mu(A) - \pi(A)}{\alpha} > \frac{\mu^*(A) - \pi(A)}{\alpha} \overset{(c)}{=} \bigwedge_{u \in \mathcal{U}} W(A \mid u)$$

and we obtain the contradiction $\bigwedge_{u \in \mathcal{U}} W(A \mid u) > \bigwedge_{u \in \mathcal{U}} W(A \mid u)$, where (a) holds by definition of greatest common component (3), (b) holds because $W \geq (\mu - \pi)^+/\alpha$ by (163) and so we may lower-bound the supremum by this specific instance, and (c) holds because μ^* is the right-hand side of (160). □

Lemma 9 (Properties of Upper Concave Envelope)**.** *Let $f : \mathcal{I} \to [0,1]$ be a function defined on a (possibly infinite) interval $\mathcal{I} \subseteq \mathbb{R}_+$. The upper concave envelope $\hat{f} : \mathcal{I} \to [0,1]$ of f satisfies the following properties:*

1) *For all boundary points t of $\mathcal{I}$, we have $\hat{f}(t) = f(t)$.*
2) *If f is non-decreasing, then $\hat{f}$ is non-decreasing.*
3) *If f is non-decreasing, $\mathcal{I} = [0,1]$, and $f(t) \leq t$ for all $t \in \mathcal{I}$, then $\hat{f}$ is continuous on $\mathcal{I}$.*

Proof.

Part 1: Consider a boundary point $a \in \mathcal{I}$. Suppose for the sake of contradiction that $\hat{f}(a) \neq f(a)$. Since $\hat{f} \geq f$ everywhere on $\mathcal{I}$ by definition of upper concave envelope, we must have $\hat{f}(a) > f(a)$. Consider the function $g : \mathcal{I} \to [0,1]$ given by

$$\forall t \in \mathcal{I},\ \ g(t) \triangleq \begin{cases} f(a)\,, & \text{if } t = a\,, \\ \hat{f}(t)\,, & \text{if } t \neq a\,. \end{cases} \tag{164}$$

Clearly, $f \leq g \leq \hat{f}$ everywhere on $\mathcal{I}$. Moreover, g is concave, because for all distinct $s, t \in \mathcal{I}$ and all $\theta \in (0,1)$,

$$g(\theta s + (1-\theta)\,t) \overset{(a)}{=} \hat{f}(\theta s + (1-\theta)\,t) \overset{(b)}{\geq} \theta\,\hat{f}(s) + (1-\theta)\,\hat{f}(t) \overset{(c)}{\geq} \theta\,g(s) + (1-\theta)\,g(t)\,,$$

where (a) holds by the second case in (164) because $\theta s + (1-\theta)t$ is an interior point of $\mathcal{I}$, (b) holds because $\hat{f}$ is concave, and (c) holds because $g \leq \hat{f}$ everywhere, as mentioned above. This contradicts the fact that $\hat{f}$ is the pointwise infimum of all concave upper bounds of f.

Part 2: Let $\bar{f} : \mathcal{I} \to \mathbb{R}$ be the closed upper concave envelope (or "closed convex hull") of f [44, Chapter B, Proposition 2.5.2, Definition 2.5.3], i.e.,

$$\forall t \in \mathcal{I},\ \ \bar{f}(t) \triangleq \inf\{\alpha t + \beta : \alpha \geq 0,\ \beta \in \mathbb{R} \text{ such that } \forall s \in \mathcal{I},\ \alpha s + \beta \geq f(s)\}\,, \tag{165}$$

where it suffices to consider only $\alpha \geq 0$ because f is non-decreasing. Fix $s, t \in \mathcal{I}$ such that $s < t$. By definition of $\bar{f}$ as an infimum (165), there exist sequences of coefficients $\{\alpha_n\}_{n \in \mathbb{N}} \subset \mathbb{R}_+$ and $\{\beta_n\}_{n \in \mathbb{N}} \subset \mathbb{R}$, satisfying $\alpha_n s + \beta_n \geq f(s)$ for all $s \in \mathcal{I}$ and $n \in \mathbb{N}$, such that $\bar{f}(t) = \lim_{n \to \infty}\{\alpha_n t + \beta_n\}$. It follows that

$$\bar{f}(s) \overset{(a)}{\leq} \inf_{n \in \mathbb{N}}\{\alpha_n s + \beta_n\} \leq \lim_{n \to \infty}\{\alpha_n s + \beta_n\} \overset{(b)}{\leq} \lim_{n \to \infty}\{\alpha_n t + \beta_n\} = \bar{f}(t)\,,$$

where (a) holds by taking the infimum in (165) over a smaller set, and (b) holds because $s < t$ and $\alpha_n \geq 0$ for all $n \in \mathbb{N}$. This establishes that $\bar{f}$ is non-decreasing on $\mathcal{I}$. Since $\bar{f}$ is the closure of $\hat{f}$ [44, Chapter B, Proposition 2.5.2], $\bar{f}$ and $\hat{f}$ agree on the interior of $\mathcal{I}$ [44, Chapter B, Proposition 1.2.6] and $\hat{f} \leq \bar{f}$ on any boundary points of $\mathcal{I}$. Hence, $\hat{f}$ is non-decreasing on $\mathcal{I} - \{\sup \mathcal{I}\}$. If $\mathcal{I}$ is finite and includes its right boundary point $b \in \mathcal{I}$, then $\hat{f}(b) \leq \bar{f}(b) \leq f(b)$, where the second inequality holds by upper-bounding the infimum in (165) with the specific majorant $\alpha = 0$ and $\beta = f(b)$. By Part 1, $\hat{f}(b) = f(b)$, and so $\hat{f}(b) = \bar{f}(b)$. Thus, $\hat{f}$ is non-decreasing on all of $\mathcal{I}$, as desired.

Part 3: By concavity, $\hat{f}$ is continuous on the interior of $\mathcal{I}$. To establish the continuity of $\hat{f}$ at 0, observe that

$$\forall t \in \mathcal{I},\ \ \hat{f}(t) \overset{(a)}{\geq} \hat{f}(0) \overset{(b)}{=} f(0) \overset{(c)}{=} 0\,,$$

where (a) holds by Part 2, (b) holds by Part 1, and (c) holds by the assumption $f(t) \le t$ and the range of f. Furthermore,

$$\forall t \in \mathcal{I},\ \ \hat{f}(t) \le \bar{f}(t) \overset{(a)}{\le} t\,,$$

where (a) holds by upper-bounding the infimum in (165) with the specific majorant $\alpha = 1$ and $\beta = 0$ due to the assumption $f(t) \le t$. Hence, we have $\lim_{t\to 0^+} \hat{f}(t) = 0 = \hat{f}(0)$ as desired. To establish the continuity of $\hat{f}$ at 1, observe that $\lim_{t\to 1^-} \hat{f}(t) \le \hat{f}(1)$ because $\hat{f}$ is non-decreasing by Part 2, and $\lim_{t\to 1^-} \hat{f}(t) \ge \lim_{t\to 1^-} \{t\,\hat{f}(1) + (1-t)\,\hat{f}(0)\} = \hat{f}(1)$ because $\hat{f}$ is concave. Hence, $\lim_{t\to 1^-} \hat{f}(t) = \hat{f}(1)$ as desired. □

Appendix B
Variational Characterization Under Equicontinuity

In this appendix, we provide an alternative statement and proof of Theorem 1 without the use of lattice infima, under the assumption of equicontinuity of the kernel K.

Theorem 7 (Variational Characterization of Doeblin Coefficient)**.** *Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ be Polish spaces, where $\mathcal{X}$ is endowed with the metric $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Let $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ be an absolutely continuous Markov kernel with respect to the σ-finite measure $\mu : \mathcal{F}_{\mathcal{Y}} \to \mathbb{R}_+$. Assume the family of functions $\{x \mapsto \frac{dK}{d\mu}(y \mid x) : y \in \mathcal{Y}\}$ is equicontinuous [77, Definition 7.22], i.e.,*

$$\forall \epsilon > 0,\ \exists \delta > 0,\ \forall x, x' \in \mathcal{X},\ \forall y \in \mathcal{Y},\ d_{\mathcal{X}}(x, x') < \delta \implies \left| \frac{dK}{d\mu}(y \mid x) - \frac{dK}{d\mu}(y \mid x') \right| < \epsilon\,. \tag{166}$$

Then, the Doeblin coefficient of K admits the characterization

$$\tau(K) = \inf_{n\in\mathbb{N}} \inf_{x_1,\ldots,x_n \in \mathcal{X}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} K(A_i \mid x_i)\,. \tag{167}$$

Proof. Fix a Markov kernel $K : \mathcal{X} \times \mathcal{F}_{\mathcal{Y}} \to [0,1]$ satisfying the assumptions in Theorem 7. For notational convenience, denote the right-hand side of (167) as

$$① \triangleq \inf_{n\in\mathbb{N}} \inf_{x_1,\ldots,x_n \in \mathcal{X}} \inf_{\substack{n\text{-partition of } \mathcal{Y} \\ A_1,\ldots,A_n}} \sum_{i=1}^{n} K(A_i \mid x_i)\,.$$

First, we will show that $\tau(K) \le ①$. By definition of Doeblin coefficient as a supremum (4) which is achieved as per the discussion following (4), there exists some probability measure $\pi^* : \mathcal{F}_{\mathcal{Y}} \to [0,1]$ such that $K \ge \tau(K)\,\pi^*$. For any $n \in \mathbb{N}$, any $x_1, \ldots, x_n \in \mathcal{X}$, and any partition $A_1, \ldots, A_n$ of $\mathcal{Y}$, we have

$$\tau(K) \overset{(a)}{=} \tau(K)\,\pi^*(\mathcal{Y}) \overset{(b)}{=} \tau(K) \sum_{i=1}^{n} \pi^*(A_i) \overset{(c)}{\le} \sum_{i=1}^{n} K(A_i \mid x_i)\,,$$

where (a) holds because π^* is a probability measure, (b) holds because $A_1, \ldots, A_n$ is a partition of $\mathcal{Y}$, and (c) holds because $K \ge \tau(K)\,\pi^*$. Since n, $x_1, \ldots, x_n$, and $A_1, \ldots, A_n$ were arbitrary, it follows that $\tau(K) \le ①$ as desired.

Next, we will show that $① \le \tau(K)$. Since $\mathcal{X}$ is Polish and thus separable, there exists a countable dense net $\{x_1^*, x_2^*, \ldots\} \subseteq \mathcal{X}$, i.e.,

$$\forall x \in \mathcal{X},\ \forall \delta > 0,\ \exists i \in \mathbb{N},\ d_{\mathcal{X}}(x, x_i^*) \le \delta\,. \tag{168}$$

For notational convenience, define a sequence of functions $g_n : \mathcal{Y} \to \mathbb{R}_+$ as

$$\forall n \in \mathbb{N},\ \ g_n(y) \triangleq \min_{i\in[n]} \frac{dK}{d\mu}(y \mid x_i^*)\,.$$

We have

$$① \overset{(a)}{=} \inf_{n\in\mathbb{N}} \inf_{x_1,\ldots,x_n \in \mathcal{X}} \int_{\mathcal{Y}} \left(\min_{i\in[n]} \frac{dK}{d\mu}(\cdot \mid x_i) \right) d\mu \overset{(b)}{\le} \inf_{n\in\mathbb{N}} \int_{\mathcal{Y}} g_n\,d\mu \overset{(c)}{=} \lim_{n\to\infty} \int_{\mathcal{Y}} g_n\,d\mu\,, \tag{169}$$

where (a) holds by Proposition 2, (b) holds by upper-bounding the inner infimum with the specific instance $\{x_1^*, \ldots, x_n^*\}$, and (c) holds because $\{g_n\}_{n\in\mathbb{N}}$ is a non-increasing sequence, since each successive g_n is the minimum over a larger set of i. Next, define a function $g : \mathcal{Y} \to \mathbb{R}_+$ as

$$g(y) \triangleq \lim_{n\to\infty} g_n(y) = \inf_{i\in\mathbb{N}} \frac{dK}{d\mu}(y \mid x_i^*)\,. \tag{170}$$

Observe that g is measurable, because it is the countable infimum of measurable functions. Furthermore, observe that $g_n \le g_1$ for all $n \in \mathbb{N}$, $g \le g_1$, and $\int_{\mathcal{Y}} g_1\,d\mu = K(\mathcal{Y} \mid x_1^*) = 1 < \infty$. Hence, proceeding from (169) and applying the dominated convergence theorem, we have

$$① \le \int_{\mathcal{Y}} \left(\lim_{n\to\infty} g_n \right) d\mu = \underbrace{\int_{\mathcal{Y}} g\,d\mu}_{=\alpha^*}\,, \tag{171}$$

where we define $\alpha^* \in [0,1]$ above for convenience hereafter. If $\alpha^* = 0$, we trivially have $① = 0 \le \tau(K)$ as desired. Hence, assume $\alpha^* > 0$ for the remainder of this argument. Define a probability measure $\pi^* : \mathcal{F}_{\mathcal{Y}} \to [0,1]$ as $\pi^*(A) \triangleq \frac{1}{\alpha^*} \int_A g \, d\mu$ for all $A \in \mathcal{F}_{\mathcal{Y}}$. Fix an arbitrary $\epsilon > 0$. Consider $\delta > 0$ (possibly depending on ϵ) such that (166) holds. Observe that for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have

$$g(y) \overset{(a)}{\le} \frac{dK}{d\mu}(y \mid x_i^*) \overset{(b)}{\le} \frac{dK}{d\mu}(y \mid x) + \epsilon \,, \tag{172}$$

where (a) holds for *all* $i \in \mathbb{N}$ by definition of g (170), and (b) holds for *some* $i \in \mathbb{N}$ (possibly depending on x) such that $d_{\mathcal{X}}(x, x_i^*) < \delta$, by equicontinuity (166); we remark that the existence of such an i is guaranteed by (168). Hence, for any $x \in \mathcal{X}$ and $A \in \mathcal{F}_{\mathcal{Y}}$,

$$\alpha^* \pi^*(A) = \int_A g \, d\mu \overset{(a)}{\le} \int_A \left(\frac{dK}{d\mu}(\cdot \mid x) \right) d\mu = K(A \mid x) \,, \tag{173}$$

where (a) holds by (172) because $\epsilon > 0$ was arbitrary. Proceeding from (171), we have

$$① \le \alpha^* \overset{(a)}{\le} \sup\{\alpha \in \mathbb{R} : \exists \pi \in \mathscr{P}, \, K \ge \alpha\pi\} \overset{(b)}{=} \tau(K)$$

as desired, where (a) holds because $K \ge \alpha^* \pi^*$ by (173), and (b) holds by definition of Doeblin coefficient (4). □

We remark that the equicontinuity assumption on K allows us to circumvent measurability issues in the proof by defining g as the pointwise infimum over a countable dense subset of $\mathcal{X}$, which is guaranteed to be measurable while approximating the infimum over all of general uncountable $\mathcal{X}$ to any $\epsilon > 0$ accuracy. Otherwise, g would have to be defined as the lattice or essential infimum, as was done in the proof of Theorem 1 in Section IV-A.

Appendix C
Markov Kernel Examples

In this appendix, we provide examples of non-trivial Markov kernels satisfying the preconditions for various results in our paper. Throughout this appendix, let $\Phi : \mathbb{R} \to (0,1)$ denote the standard Gaussian CDF $\Phi(x) \triangleq \int_{-\infty}^{x} (1/\sqrt{2\pi}) \exp(-t^2/2) \, dt$.

First, we present a Markov kernel $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ with a countably infinite source space $\mathcal{U}$, unbounded support on the target space $\mathcal{X}$, and finite average extremal power.

Proposition 10 (Average Extremal Power Example)**.** *Let $\mathcal{U}$ be countable (without loss of generality, $\mathcal{U} \triangleq \mathbb{N}$) and $\mathcal{X} \triangleq \mathbb{R}$. Let $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ be a Markov kernel such that for each $i \in \mathcal{U}$, $X_i \sim W(\cdot \mid i)$ is sub-Gaussian with mean 0 and variance factor $\sigma_i^2 \le c / \log_e(i+1)$, i.e.,*

$$\forall \lambda \in \mathbb{R}, \;\; \mathbb{E}\Big[e^{\lambda(X_i - \mathbb{E}[X_i])}\Big] \le \exp\left(\frac{\sigma_i^2 \lambda^2}{2}\right),$$

where $c > 0$ is a fixed constant. Then, under the norm $\|x\| \triangleq |x|$ and power function $M(z) \triangleq z^2$, the average extremal power of W satisfies $|||W|||_{\mathsf{AE}} \le 4c(1 + \pi^2/6)$.

Proof. Let $\mathbb{P}$ be any coupling of random variables $\{X_i\}_{i \in \mathbb{N}}$ with $X_i \sim W(\cdot \mid i)$ for each $i \in \mathbb{N}$. In the following, all expectations are taken with respect to $\{X_i\}_{i \in \mathbb{N}} \sim \mathbb{P}$. We have

$$|||W|||_{\mathsf{AE}} \overset{(a)}{=} \mathbb{E}\left[\sup_{i \in \mathbb{N}} M(\|X_i\|)\right] \overset{(b)}{=} \mathbb{E}\left[\sup_{i \in \mathbb{N}} |X_i|^2\right] = \mathbb{E}\left[\lim_{n \to \infty} \max_{i \in [n]} |X_i|^2\right] \overset{(c)}{=} \lim_{n \to \infty} \underbrace{\mathbb{E}\left[\max_{i \in [n]} |X_i|^2\right]}_{①} , \tag{174}$$

where (a) holds by definition of average extremal power (13), (b) holds by definition of $\|\cdot\|$ and M, and (c) holds by the monotone convergence theorem because the sequence of functions $\{g_n\}_{n \in \mathbb{N}}$ given by $g_n(\{x_i\}_{i \in \mathbb{N}}) \triangleq \max_{i \in [n]} |x_i|^2$ is monotonically non-decreasing in n. Next, we upper-bound ① by adapting the result in [78, Lemma 2.3]. For any $t_0 \ge 0$, we have

$$\begin{aligned}
① &\overset{(a)}{\le} \int_0^\infty \mathbb{P}\left(\max_{i \in [n]} |X_i|^2 \ge t\right) dt \overset{(b)}{=} \int_0^{t_0} \mathbb{P}\left(\max_{i \in [n]} |X_i|^2 \ge t\right) dt + \int_{t_0}^\infty \mathbb{P}\left(\max_{i \in [n]} |X_i|^2 \ge t\right) dt \\
&\overset{(c)}{\le} t_0 + \int_{t_0}^\infty \mathbb{P}\left(\max_{i \in [n]} |X_i|^2 \ge t\right) dt \overset{(d)}{\le} t_0 + \int_{t_0}^\infty \sum_{i=1}^n \mathbb{P}\Big(|X_i|^2 \ge t\Big) dt = t_0 + \sum_{i=1}^n \int_{t_0}^\infty \mathbb{P}\Big(|X_i| \ge t^{\frac{1}{2}}\Big) dt \\
&\overset{(e)}{\le} t_0 + \sum_{i=1}^n \int_{t_0}^\infty 2 \exp\left(-\frac{t}{2\sigma_i^2}\right) dt \overset{(f)}{=} t_0 + 4 \sum_{i=1}^n \sigma_i^2 \exp\left(-\frac{t_0}{2\sigma_i^2}\right),
\end{aligned} \tag{175}$$

where (a) holds by the layer cake representation [79, Lemma 1.2.1], (b) holds for any $t_0 \in [0, \infty]$ by splitting the region of integration into two intervals, (c) holds because probability values are bounded by 1, (d) holds by the union bound, (e) holds

by [72, Eq. 2.9] because X_i is sub-Gaussian with mean 0 and variance factor σ_i^2, and (f) holds by evaluating the integral. Choosing $t_0 \triangleq 4\max_{i\in[n]}\left\{\sigma_i^2 \log_e(i+1)\right\}$ and continuing from (175), we have

$$\begin{aligned}
① &\overset{(a)}{\leq} 4\max_{i\in[n]}\left\{\sigma_i^2\log_e(i+1)\right\} + 4\sum_{i=1}^{n}\sigma_i^2\exp\left(-\frac{4\sigma_i^2\log_e(i+1)}{2\sigma_i^2}\right) \\
&= 4\max_{i\in[n]}\left\{\sigma_i^2\log_e(i+1)\right\} + 4\sum_{i=1}^{n}\frac{\sigma_i^2}{(i+1)^2} \overset{(b)}{\leq} 4c + 4\sum_{i=1}^{n}\frac{c}{(i+1)^2\log_e(i+1)}\,,
\end{aligned} \tag{176}$$

where (a) holds by the choice of t_0 and (b) holds because $\sigma_i^2 \leq c/\log_e(i+1)$ for each $i \in \mathbb{N}$. Combining (174) and (176), we have

$$|||W|||_{\mathsf{AE}} \leq 4c\left(1 + \sum_{i=1}^{\infty}\frac{1}{(i+1)^2\log_e(i+1)}\right) \leq 4c\left(1+\sum_{i=1}^{\infty}\frac{1}{i^2}\right) = 4c\left(1+\frac{\pi^2}{6}\right)$$

as desired. □

Next, we present a Markov kernel $W : \mathcal{U}\times\mathcal{F}_{\mathcal{X}} \to [0,1]$ with an uncountable source space $\mathcal{U}$, unbounded support on the target space $\mathcal{X}$, and finite average extremal power.

Proposition 11 (Average Extremal Power Example)**.** *Let $\mathcal{U} \triangleq (0,\infty)$ and $\mathcal{X} \triangleq \mathbb{R}$. Let $W : \mathcal{U}\times\mathcal{F}_{\mathcal{X}} \to [0,1]$ be the Markov kernel such that for each $u \in \mathcal{U}$, $X_u \sim W(\cdot \mid u)$ is given by*

$$X_u \sim \begin{cases} \mathsf{Normal}(0,1)\,, & \text{if } J_u = 1\,, \\ 2\,\mathsf{Beta}(u,u) - 1\,, & \text{if } J_u = 0\,, \end{cases}$$

where $J_u \sim \mathsf{Bernoulli}(1/2)$ is independent of the Beta and Gaussian random variables. Namely, the cumulative distribution function $f_{X_u} : \mathbb{R} \to (0,1)$ is

$$f_{X_u}(x) = \frac{1}{2}\,\Phi(x) + \begin{cases} 0\,, & \text{if } x < -1\,, \\ \frac{1}{4\mathrm{B}(u,u)}\int_{-1}^{x}\left(\frac{t+1}{2}\right)^{u-1}\left(1-\frac{t+1}{2}\right)^{u-1}dt\,, & \text{if } -1 \leq x \leq 1\,, \\ \frac{1}{2}\,, & \text{if } x > 1\,, \end{cases}$$

where $\mathrm{B}(\alpha,\beta)$ denotes the Beta function $\mathrm{B}(\alpha,\beta) \triangleq \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$, and the probability density function p_{X_u} is

$$p_{X_u}(x) = \begin{cases} \frac{1}{4\mathrm{B}(u,u)}\left(\frac{x+1}{2}\right)^{u-1}\left(1-\frac{x+1}{2}\right)^{u-1} + \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right), & \text{if } |x| < 1\,, \\ \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right), & \text{if } |x| > 1\,. \end{cases} \tag{177}$$

Then, under the norm $\|x\| \triangleq |x|$ and power function $M(z) \triangleq z^2$, the average extremal power of W satisfies $|||W|||_{\mathsf{AE}} \leq 1$.

Proof. First, we compute the pointwise infimum of the probability densities $\{p_{X_u}\}_{u\in\mathcal{U}}$. We have

$$\inf_{u\in\mathcal{U}} p_{X_u}(x) \overset{(a)}{=} \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) + \mathbb{1}\{|x|<1\}\underbrace{\inf_{u\in\mathcal{U}}\frac{1}{4\mathrm{B}(u,u)}\left(\frac{x+1}{2}\right)^{u-1}\left(\frac{1-x}{2}\right)^{u-1}}_{①}, \tag{178}$$

where (a) holds by (177). For any $x \in (-1,1)$,

$$\begin{aligned}
① &\overset{(a)}{\leq} \lim_{u\to 0}\frac{1}{4\mathrm{B}(u,u)}\left(\frac{x+1}{2}\right)^{u-1}\left(\frac{1-x}{2}\right)^{u-1} = \frac{1}{4}\lim_{u\to 0}\left\{\frac{1}{\mathrm{B}(u,u)}\right\}\lim_{u\to 0}\left\{\left(\frac{x+1}{2}\right)^{u-1}\left(\frac{1-x}{2}\right)^{u-1}\right\} \\
&\overset{(b)}{=} \frac{1}{4}\left(\frac{4}{1-x^2}\right)\lim_{u\to 0}\frac{1}{\mathrm{B}(u,u)} \overset{(c)}{=} \frac{1}{1-x^2}\lim_{u\to 0}\frac{\Gamma(2u)}{\Gamma(u)^2} \overset{(d)}{=} \frac{1}{1-x^2}\lim_{u\to 0}\frac{2^{2u-1}\Gamma\left(u+\frac{1}{2}\right)}{\sqrt{\pi}\,\Gamma(u)} = \frac{1}{2\,(1-x^2)}\lim_{u\to 0}\frac{1}{\Gamma(u)} = 0\,,
\end{aligned} \tag{179}$$

where (a) holds because $u = 0$ is a limit point of $\mathcal{U}$, (b) holds because $x \in (-1,1)$ and so $((x+1)/2)^{-1}$ and $((1-x)/2)^{-1}$ are well-defined, (c) holds by the identity $\mathrm{B}(\alpha,\beta) = \Gamma(\alpha)\,\Gamma(\beta)/\Gamma(\alpha+\beta)$, where Γ denotes the Gamma function $\Gamma(\alpha) \triangleq \int_0^\infty t^{\alpha-1}e^{-t}dt$, and (d) holds by the *Legendre duplication formula* $\Gamma(2\alpha) = 2^{2\alpha-1}\Gamma(\alpha)\,\Gamma\left(\alpha+\frac{1}{2}\right)/\sqrt{\pi}$. Also, clearly $① \geq 0$ by inspection. Hence, combining (178) and (179) yields

$$\inf_{u\in\mathcal{U}} p_{X_u}(x) = \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right).$$

Following from the above, the greatest common component of W is

$$\bigwedge_{u\in\mathcal{U}} W(A \mid u) \overset{(a)}{=} \int_A \left(\inf_{u\in\mathcal{U}} p_{X_u}(x)\right)dx = \int_A \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx\,,$$

where (a) holds by Lemma 4 and because the lattice infimum reduces to the pointwise infimum when the latter is measurable. Hence, the maximal coupling of random variables $\{X_u\}_{u\in\mathcal{U}}$ with $X_u \sim W(\cdot \mid u)$ for each $u \in \mathcal{U}$, defined in (34), is

$$\forall u \in \mathcal{U},\ \ X_u \triangleq \begin{cases} X^*\,, & \text{if } I = 1\,, \\ \tilde{X}_u\,, & \text{if } I = 0\,, \end{cases} \tag{180}$$

where the random variables I, X^*, and $\{\tilde{X}_u\}_{u\in\mathcal{U}}$ are sampled independently from the distributions

$$I \sim \mathsf{Bernoulli}\left(\frac{1}{2}\right),$$

$$X^* \sim \mathsf{Normal}(0,1)\,, \tag{181}$$

$$\forall u \in \mathcal{U},\ \ \tilde{X}_u \sim 2\,\mathsf{Beta}(u,u) - 1\,. \tag{182}$$

Therefore, the average extremal power of W is

$$|||W|||_{\mathsf{AE}} \overset{(a)}{=} \mathbb{P}(I=1)\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, I=1\right] + \mathbb{P}(I=0)\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} M(\|X_u\|) \,\middle|\, I=0\right] \overset{(b)}{=} \underbrace{\frac{1}{2}\,\mathbb{E}\left[(X^*)^2\right]}_{=\frac{1}{2}\ (181)} + \underbrace{\frac{1}{2}\,\mathbb{E}\left[\sup_{u\in\mathcal{U}} \tilde{X}_u^2\right]}_{\leq\frac{1}{2}\ (182)} \leq 1$$

as desired, where all expectations are taken with respect to the maximal coupling (180), (a) holds by definition of average extremal power (13) and the law of total expectation, and (b) holds by definitions of $\|\cdot\|$, M, and X_u. □

Next, we present a Markov kernel $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ with unbounded support on the target space $\mathcal{X}$ which satisfies the preconditions for Proposition 7.

Proposition 12 (Upper Bound Example)**.** *Let $n \in \mathbb{N}$ and $\varsigma > 0$ be fixed constants. Let $\mathcal{U} \triangleq [n]$ and let $\mathcal{X} \triangleq \mathbb{R}_+$ be equipped with the norm $\|\cdot\| \triangleq |\cdot|$. Let $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ be the Markov kernel such that for each $i \in \mathcal{U}$, $X_i \sim W(\cdot \mid i)$ is given by*

$$X_i \triangleq \begin{cases} X_{i|1} \sim \mu_i + \mathsf{HalfNormal}(\varsigma)\,, & \text{if } J_i = 1\,, \\ X_{i|0} \sim \mathsf{Uniform}(i, \mu_i)\,, & \text{if } J_i = 0\,, \end{cases}$$

where $J_i \sim \mathsf{Bernoulli}(1/2)$, $X_{i|1}$, and $X_{i|0}$ are independent, and $\mu_i \triangleq i + 2\varsigma\sqrt{2/\pi}$. Namely, the cumulative distribution function $f_{X_i} : \mathbb{R} \to [0,1)$ is

$$f_{X_i}(x) = \begin{cases} 0\,, & \text{if } x < i\,, \\ \frac{x-i}{4\varsigma}\sqrt{\frac{\pi}{2}}\,, & \text{if } i \leq x < \mu_i\,, \\ \Phi\left(\frac{x-\mu_i}{\varsigma}\right), & \text{if } x \geq \mu_i\,, \end{cases}$$

and the probability density function p_{X_i} is

$$p_{X_i}(x) = \begin{cases} 0\,, & \text{if } x < i\,, \\ \frac{1}{4\varsigma}\sqrt{\frac{\pi}{2}}\,, & \text{if } i \leq x < \mu_i\,, \\ \frac{1}{\varsigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_i)^2}{2\varsigma^2}\right), & \text{if } x > \mu_i\,. \end{cases}$$

Then, under the definition of $\mathcal{G}$ in Proposition 7 with power function $M(z) \triangleq z$, $p \triangleq \mu_n$, and $\sigma \triangleq \sqrt{2}\,(4\varsigma/\pi)$, we have $W \in \mathcal{G}$.

Proof. Throughout this proof, for notational convenience, let $Z_i \triangleq M(\|X_i\|) \overset{(a)}{=} X_i$ for all $i \in \mathcal{U}$, where (a) holds because $X_i \geq i > 0$.

Uniform average power constraint: Observe that for each $i \in \mathcal{U}$, we have

$$\mathbb{E}[Z_i] = \mathbb{E}[X_i] \overset{(a)}{=} \mathbb{P}(J_i=0)\mathbb{E}\big[X_{i|0}\big] + \mathbb{P}(J_i=1)\mathbb{E}\big[X_{i|1}\big] \overset{(b)}{=} \frac{1}{2}\mathbb{E}\big[X_{i|0}\big] + \frac{1}{2}\mathbb{E}\big[X_{i|1}\big] \overset{(c)}{=} \frac{1}{2}\left(\frac{i+\mu_i}{2}\right) + \frac{1}{2}\left(\mu_i + \varsigma\sqrt{2/\pi}\right) = \mu_i\,, \tag{183}$$

where (a) holds by the law of total expectation, (b) holds because $J_i \sim \mathsf{Bernoulli}(1/2)$, and (c) holds because $X_{i|0} \sim \mathsf{Uniform}(i,\mu_i)$ and $X_{i|1} \sim \mu_i + \mathsf{HalfNormal}(\varsigma)$. Therefore,

$$|||W|||_{\mathsf{UA}} \overset{(a)}{=} \sup_{i\in\mathcal{U}} \mathbb{E}[Z_i] \overset{(b)}{=} \mu_n = p$$

as desired, where (a) holds by definition of uniform average power (12) and (b) holds from (183) because $\mathcal{U} = [n]$.

Sub-Gaussianity: Fix $i \in \mathcal{U}$. For any $t \geq 0$, we have

$$\mathbb{P}(|Z_i - \mathbb{E}[Z_i]| \geq t) = \mathbb{P}(|X_i - \mathbb{E}[X_i]| \geq t) \overset{(a)}{=} \mathbb{P}(J_i=0)\,\mathbb{P}\big(\big|X_{i|0} - \mathbb{E}[X_i]\big| \geq t\big) + \mathbb{P}(J_i=1)\,\mathbb{P}\big(\big|X_{i|1} - \mathbb{E}[X_i]\big| \geq t\big)$$

$$\overset{(b)}{=} \frac{1}{2}\,\mathbb{P}\big(\big|X_{i|0} - \mathbb{E}[X_i]\big| \geq t\big) + \frac{1}{2}\,\mathbb{P}\big(\big|X_{i|1} - \mathbb{E}[X_i]\big| \geq t\big) \overset{(c)}{=} \frac{1}{2}\,\mathbb{P}\big(\big|X_{i|0} - \mu_i\big| \geq t\big) + \frac{1}{2}\,\mathbb{P}\big(\big|X_{i|1} - \mu_i\big| \geq t\big)$$
$$\overset{(d)}{=} \underbrace{\frac{1}{2}\,\mathbb{P}\big(X_{i|0} \leq \mu_i - t\big)}_{①} + \underbrace{\frac{1}{2}\,\mathbb{P}\big(X_{i|1} \geq \mu_i + t\big)}_{②}, \tag{184}$$

where (a) holds by the law of total probability, (b) holds because $J_i \sim \mathsf{Bernoulli}(1/2)$, (c) holds by (183), and (d) holds because $X_{i|0} \leq \mu_i$ and $X_{i|1} \geq \mu_i$. Next, we upper-bound ① and ②. Let $N_i \sim \mathsf{Normal}(0, (4\varsigma/\pi)^2)$ be an independent Gaussian random variable. Then,

$$① \overset{(a)}{=} \frac{1}{2}\max\left\{0, 1 - \int_{-t}^{0} \frac{1}{2\varsigma\sqrt{2/\pi}}\,dx\right\} = \max\left\{0, \frac{1}{2} - \int_{-t}^{0} \frac{1}{4\varsigma\sqrt{2/\pi}}\,dx\right\}$$
$$\overset{(b)}{\leq} \max\left\{0, \frac{1}{2} - \int_{-t}^{0} \frac{1}{(4\varsigma/\pi)\sqrt{2\pi}} \exp\left(-\frac{x^2}{2(4\varsigma/\pi)^2}\right) dx\right\} \overset{(c)}{=} \mathbb{P}(N_i \leq -t)\,, \tag{185}$$

where (a) holds by the PDF of a uniform random variable supported on an interval of length $\mu_i - i = 2\varsigma\sqrt{2/\pi}$, (b) holds by lower-bounding the integrand pointwise, and (c) holds by the PDF of $N_i \sim \mathsf{Normal}(0, (4\varsigma/\pi)^2)$. Similarly, we also have

$$② \overset{(a)}{=} \frac{1}{2}\int_{t}^{\infty} \frac{\sqrt{2}}{\varsigma\sqrt{\pi}} \exp\left(-\frac{x^2}{2\varsigma^2}\right) dx = 1 - \Phi\left(\frac{t}{\varsigma}\right) \overset{(b)}{\leq} 1 - \Phi\left(\frac{t}{4\varsigma/\pi}\right) = \mathbb{P}(N_i \geq t)\,, \tag{186}$$

where (a) holds by the PDF of a $\mathsf{HalfNormal}(\varsigma)$ random variable and (b) holds because CDFs are non-decreasing. Combining Equations (184) to (186), we thus have $\mathbb{P}(|Z_i - \mathbb{E}[Z_i]| \geq t) \leq \mathbb{P}(|N_i| \geq t)$, and so by [72, Theorem 2.6] we have sub-Gaussianity with variance factor $\sigma = \sqrt{2}\,(4\varsigma/\pi)$ as desired, i.e., $\mathbb{E}\big[e^{\lambda(Z_i - \mathbb{E}[Z_i])}\big] \leq \exp\left(\frac{16\varsigma^2\lambda^2}{\pi^2}\right)$ for all $\lambda \in \mathbb{R}$. □

Finally, we present a Markov kernel $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ with an uncountable source space $\mathcal{U}$ and unbounded support on the target space $\mathcal{X}$ which satisfies the preconditions for Proposition 8.

Proposition 13 (Upper Bound Example)**.** *Let $a > 0$ and $\varsigma > 0$ be fixed constants. Let $\mathcal{U} \triangleq [-a, a]$ be equipped with the metric $d_{\mathcal{U}}(u, v) \triangleq |u - v|$. Let $\mathcal{X} \triangleq \mathbb{R}$ be equipped with the norm $\|\cdot\| \triangleq |\cdot|$. Let $W : \mathcal{U} \times \mathcal{F}_{\mathcal{X}} \to [0,1]$ be the Markov kernel such that for each $u \in \mathcal{U}$, $X_u \sim W(\cdot \mid u)$ is given by*

$$X_u \triangleq \begin{cases} X_{u|1} \sim \mathsf{Normal}\big(0, \varsigma^2\big)\,, & \text{if } J_u = 1\,, \\ X_{u|0} \sim \delta_u\,, & \text{if } J_u = 0\,, \end{cases}$$

where $J_u \sim \mathsf{Bernoulli}(1/2)$, $X_{u|1}$, and $X_{u|0}$ are independent. Namely, the probability measure $W(\cdot \mid u) : \mathcal{F}_{\mathcal{X}} \to [0,1]$ is

$$\forall A \in \mathcal{F}_{\mathcal{X}},\ \ W(A \mid u) = \int_A \frac{1}{2\varsigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\varsigma^2}\right) dx + \frac{1}{2}\,\delta_u(A)\,,$$

and the cumulative distribution function $f_{X_u} : \mathbb{R} \to (0,1)$ is

$$f_{X_u}(x) = \begin{cases} \frac{1}{2}\Phi\big(\frac{x}{\varsigma}\big)\,, & \text{if } x < u\,, \\ \frac{1}{2}\Phi\big(\frac{x}{\varsigma}\big) + \frac{1}{2}\,, & \text{if } x \geq u\,. \end{cases}$$

Then, under the definition of $\mathcal{G}$ in Proposition 8 with power function $M(z) \triangleq z^2$, $p \triangleq (a^2 + \varsigma^2)/2$, and $\sigma \triangleq a$, we have $W \in \mathcal{G}$.

This example resembles an erasure channel wherein erasures are replaced by independent Gaussian variables, akin to the construction of symmetric channels using erasure channels on finite alphabets (cf. [80], [81]).

Proof. Throughout this proof, for notational convenience, let $Z_u \triangleq M(\|X_u\|) = X_u^2$ for all $u \in \mathcal{U}$.

Uniform average power constraint: Observe that for each $u \in \mathcal{U}$, we have

$$\mathbb{E}[Z_u] = \mathbb{E}\big[X_u^2\big] \overset{(a)}{=} \mathbb{P}(J_u = 0)\,\mathbb{E}\Big[X_{u|0}^2\Big] + \mathbb{P}(J_u = 1)\,\mathbb{E}\Big[X_{u|1}^2\Big] \overset{(b)}{=} \frac{1}{2}\,\mathbb{E}\Big[X_{u|0}^2\Big] + \frac{1}{2}\,\mathbb{E}\Big[X_{u|1}^2\Big] \overset{(c)}{=} \frac{u^2 + \varsigma^2}{2}\,, \tag{187}$$

where (a) holds by the law of total expectation, (b) holds because $J_u \sim \mathsf{Bernoulli}(1/2)$, and (c) holds because $X_{u|0} \sim \delta_u$ and $X_{u|1} \sim \mathsf{Normal}(0, \varsigma^2)$. Therefore, $|||W|||_{\mathsf{UA}} \overset{(a)}{=} \sup_{u \in \mathcal{U}} \mathbb{E}[Z_u] \overset{(b)}{=} (a^2 + \varsigma^2)/2 = p$ as desired, where (a) holds by definition of uniform average power (12) and (b) holds from (187) because $\mathcal{U} = [-a, a]$.

Sub-Gaussian increments: Let $\{X_u\}_{u \in \mathcal{U}}$ be distributed according to the maximal coupling defined in (34). Fix $u, v \in \mathcal{U}$. Since

$$\mathbb{E}\big[e^{\lambda((Z_u - Z_v) - \mathbb{E}[Z_u - Z_v])}\big] = \underbrace{\mathbb{E}\big[e^{\lambda(Z_u - Z_v)}\big]}_{①}\ \underbrace{e^{\lambda(\mathbb{E}[Z_v] - \mathbb{E}[Z_u])}}_{②}, \tag{188}$$

we separately consider ① and ②. To evaluate ①, notice that the greatest common component of W is

$$\bigwedge_{u\in\mathcal{U}} W(A \mid u) \overset{(a)}{=} \int_A \frac{1}{2\varsigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\varsigma^2}\right) dx + \frac{1}{2} \bigwedge_{u\in\mathcal{U}} \delta_u(A) = \int_A \frac{1}{2\varsigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\varsigma^2}\right) dx\,,$$

where (a) holds by Lemma 8, and so $\bigwedge_{u\in\mathcal{U}} W(\mathcal{X} \mid u) = \int_{-\infty}^{\infty} \frac{1}{2\varsigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\varsigma^2}\right) dx = \frac{1}{2}$. Thus, the maximal coupling $\{X_u\}_{u\in\mathcal{U}}$ such that $X_u \sim W(\cdot \mid u)$ for all $u \in \mathcal{U}$ is

$$X_u \triangleq \begin{cases} X^*\,, & \text{if } I = 1\,, \\ \tilde{X}_u\,, & \text{if } I = 0\,, \end{cases}$$

where the random variables I, X^*, and $\{\tilde{X}_u\}_{u\in\mathcal{U}}$ are sampled independently from the probability measures

$$I \sim \mathsf{Bernoulli}\left(\frac{1}{2}\right), \quad X^* \sim \mathsf{Normal}\left(0, \sigma^2\right), \quad \tilde{X}_u \sim \delta_u\,.$$

Hence,

$$\begin{aligned} ① &= \mathbb{E}\left[e^{\lambda\left(X_u^2 - X_v^2\right)}\right] \overset{(a)}{=} \mathbb{P}(I=1)\,\mathbb{E}\left[e^{\lambda\left(X^{*2} - X^{*2}\right)}\right] + \mathbb{P}(I=0)\,\mathbb{E}\left[e^{\lambda\left(\tilde{X}_u^2 - \tilde{X}_v^2\right)}\right] \\ &\overset{(b)}{=} \frac{1}{2} + \frac{1}{2}\,\mathbb{E}\left[e^{\lambda\left(\tilde{X}_u^2 - \tilde{X}_v^2\right)}\right] \overset{(c)}{=} \frac{1}{2} + \frac{1}{2}\,\mathbb{E}\left[e^{\lambda\tilde{X}_u^2}\right]\mathbb{E}\left[e^{-\lambda\tilde{X}_v^2}\right] \overset{(d)}{=} \frac{1}{2} + \frac{1}{2}\,e^{\lambda u^2} e^{-\lambda v^2} = \frac{1 + e^{-\left(v^2-u^2\right)\lambda}}{2}\,, \end{aligned} \tag{189}$$

where (a) holds by the law of total expectation, (b) holds because $I \sim \mathsf{Bernoulli}(1/2)$, (c) holds because $\tilde{X}_u$ and $\tilde{X}_v$ are independent, and (d) holds because $\tilde{X}_u \sim \delta_u$ and $\tilde{X}_v \sim \delta_v$. Next, we evaluate ② and obtain

$$② \overset{(a)}{=} \exp\left(\lambda\left(\frac{v^2+\varsigma^2}{2} - \frac{u^2+\varsigma^2}{2}\right)\right) = \frac{1}{\exp\left(-\frac{(v^2-u^2)\lambda}{2}\right)}\,, \tag{190}$$

where (a) holds by (187). Combining Equations (188) to (190), we have

$$\begin{aligned} \mathbb{E}\left[e^{\lambda((Z_u - Z_v) - \mathbb{E}[Z_u - Z_v])}\right] &\overset{(a)}{=} \cosh\left(\frac{\left(v^2-u^2\right)\lambda}{2}\right) \overset{(b)}{\leq} \exp\left(\frac{\left(u^2-v^2\right)^2\lambda^2}{8}\right) = \exp\left(\frac{(u+v)^2(u-v)^2\lambda^2}{8}\right) \\ &\overset{(c)}{\leq} \exp\left(\frac{a^2(u-v)^2\lambda^2}{2}\right) \overset{(d)}{=} \exp\left(\frac{\sigma^2 d_{\mathcal{U}}(u,v)^2\lambda^2}{2}\right) \end{aligned}$$

as desired, where (a) holds by the identity $\cosh(x) = (1+e^{-2x})/(2e^{-x})$, (b) holds by the inequality $\cosh(x) \leq \exp(x^2/2)$, (c) holds because $\mathcal{U} = [-a, a]$, and (d) holds by definition of $d_{\mathcal{U}}$ and σ.

Measurability: Observe that if $I = 1$, we have

$$\sup_{u\in\mathcal{U}} \{Z_u - \mathbb{E}[Z_u]\} = \sup_{u\in[-a,a]} \left\{X^{*2} - \frac{u^2+\varsigma^2}{2}\right\} = X^{*2} - \frac{\varsigma^2}{2}\,,$$

and if $I = 0$, we have

$$\sup_{u\in\mathcal{U}} \{Z_u - \mathbb{E}[Z_u]\} = \sup_{u\in[-a,a]} \left\{\tilde{X}_u^2 - \frac{u^2+\varsigma^2}{2}\right\} = \sup_{u\in[-a,a]} \left\{u^2 - \frac{u^2+\varsigma^2}{2}\right\} = \frac{a^2+\varsigma^2}{2}\,.$$

Hence, the pre-image of any measurable set $A \subseteq \mathbb{R}$ under $\sup_{u\in\mathcal{U}}\{Z_u - \mathbb{E}[Z_u]\}$ is the measurable event $(\{I=1\} \cap \{X^* \in \mathcal{S}_A\}) \cup \{I=0\}$ if $(a^2+\varsigma^2)/2 \in A$, and the measurable event $\{I=1\} \cap \{X^* \in \mathcal{S}_A\}$ otherwise, where $\mathcal{S}_A \subseteq \mathbb{R}$ is the set $\mathcal{S}_A = \{-\sqrt{z+\varsigma^2/2}, \sqrt{z+\varsigma^2/2} : z \in A,\, z \geq -\varsigma^2/2\}$. □

Appendix D
Doeblin Curves and Markov Chains

Lastly, we conclude with a discussion relating Doeblin curves back to the classic setting of Markov chain ergodicity. As a minor departure from our previous exclusive focus on Doeblin curves, we will also discuss contraction properties of Doeblin coefficients here. Throughout this appendix, we focus on the special case of *discrete-time finite-state time-homogeneous* Markov chains, which we represent as a *row* stochastic matrix $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$ whose entries $[\mathbf{K}]_{i,j}$ denote the probability of transitioning from state i to state j (which, by homogeneity, is the same at any time). We assume that norms operate on row vectors analogously to their usual behavior on column vectors.

Below, we present an overview of pertinent results on Markov chain convergence from the literature and add new results discussing how Doeblin curves relate to these existing characterizations.

Proposition 14 (Markov Chain Properties)**.** *Given a discrete-time finite-state time-homogeneous Markov chain represented as a row stochastic matrix $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$, the following statements are equivalent:*

a) (Positive spectral gap [82]) $|\lambda_2(\mathbf{K})| < 1$.
b) (Aperiodic unichain [82]) *The Markov chain represented by* $\mathbf{K}$ *has exactly one recurrent class, and its recurrent class is aperiodic.*
c) (Strong ergodicity [82], [83]) *The Markov chain represented by* $\mathbf{K}$ *converges to a fixed steady-state distribution* $\boldsymbol{\pi}^* \in \mathscr{P}_{d-1}$ *regardless of the initial distribution, i.e.,* $\lim_{n\to\infty} \boldsymbol{\pi}_0 \mathbf{K}^n = \boldsymbol{\pi}^*$ *for all* $\boldsymbol{\pi}_0 \in \mathscr{P}_{d-1}$. *Furthermore, the convergence is exponentially fast, i.e., there exist constants* $C > 0$ *and* $0 < \alpha < 1$ *independent of* $\boldsymbol{\pi}_0$ *such that* $\|\boldsymbol{\pi}_0 \mathbf{K}^n - \boldsymbol{\pi}^*\|_1 \leq C\alpha^n$ *for all* $n \in \mathbb{N}$.
d) (Weak ergodicity [83], [84]) *The rows of* $\mathbf{K}^n$ *equalize as time* $n \to \infty$, *i.e.,* $\lim_{n\to\infty} \|[\mathbf{K}^n]_{\langle i \rangle} - [\mathbf{K}^n]_{\langle j \rangle}\|_\infty = 0$ *for all* $i, j \in [d]$.
e) (Doeblin characterization of weak ergodicity) *For all* $n \geq d^2$, $\tau(\mathbf{K}^n) > 0$.
f) (Doeblin curves) $\mathrm{F}_{\mathbf{K}^n}\left(t; \mathbb{R}_{\mathsf{sto}}^{d\times d}\right) < t$ *for all* $n \geq d^2$ *and* $t \in (0, 1]$.

Proof. The equivalences a) through f) hold because:

- b) $\implies$ a) by [82, Theorem 4.4.2]
- $\neg$ b) $\implies$ $\neg$ a) by [82, Section 4.4.2, p. 178]
- b) $\implies$ c) by [82, Theorem 4.3.7]
- $\neg$ b) $\implies$ $\neg$ c) by [82, Section 4.3.5, p. 175]
- c) $\iff$ d) by [83, p. 867]
- e) $\implies$ d) by [28, Eq. 6] or [84, Theorem 4.8]

Proof of d) $\implies$ e): By [84, Theorem 4.8], d) implies that there exists some $N \in \mathbb{N}$ such that $\tau(\mathbf{K}^N) > 0$ (i.e., a weaker version of e) that lacks an explicit threshold), but the threshold d^2 is not usually explicitly derived in the literature. To remedy this, we supply a proof of the threshold below.

Lemma 10 (Threshold For Positivity of $\tau(\mathbf{K}^n)$)**.** *If there exists some* $N \in \mathbb{N}$ *such that* $\tau(\mathbf{K}^N) > 0$, $\tau(\mathbf{K}^n) > 0$ *for all* $n \geq d^2$.

Proof. If $\tau(\mathbf{K}^N) > 0$, there must exist some state i in the Markov chain represented by $\mathbf{K}$ that can be reached in exactly N steps from any state including itself. Thus, i must be a part of some cycle of length $p \leq d$ in the graph representation of $\mathbf{K}$.

We argue that unless $p = 1$, there must exist another cycle that includes state i with length $q \leq d$ such that $p \perp q$ (i.e., p and q are coprime). If this is not the case, every cycle that includes state i must have a length that is a multiple of some integer $k \geq 2$. Let j be the predecessor state of i in one or more of these cycles. Then, all paths from j to i must have length $\equiv 1 \mod k$. However, all paths from i to i must have length $\equiv 0 \mod k$, thus no path from j to i can have the same length as any path from i to i, which is a contradiction.

Thus, state i is included in two (possibly equal) cycles of coprime lengths p and q. From the solution of the Frobenius coin problem with two coins [85], it is known that any integer $\geq (p-1)(q-1)$ can be represented as $ap + bq$ where a, b are nonnegative integers. Given any state l, let $d(l) < d$ be the distance from l to i (i must be reachable from l by assumption). Then, for any integer $m \geq d(l) + (p-1)(q-1)$, there must exist a trajectory of exactly length m from l to i. Since $d(l) + (p-1)(q-1) \leq (d-1) + (d-1)(d-2) = d^2 - 2d + 1 \leq d^2$ since $p, q \leq d$ and $p \perp q$, $\tau(\mathbf{K}^n) > 0$ for all $n \geq d^2$. □

Note that this result bears a strong resemblance to [86, Corollary 8.5.8] which states a similar threshold for primitive matrices. Furthermore, kernels represented by primitive matrices are ergodic with no transient states (i.e., every state is in the recurrent class), thus the class of primitive matrices is a subset of the class of ergodic Markov chains.

Although c) $\iff$ d) is known in the literature [83, p. 867], this equivalence is seldom directly argued for homogeneous Markov chains, since the distinction between strong and weak forms of ergodicity is traditionally only made in the context of inhomogeneous chains. So, for completeness, we provide a proof of c) $\iff$ d) from first principles below.

Proof of c) $\iff$ d): Fix a time-homogeneous Markov chain $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$. First, we will prove the forward direction; the argument is straightforward and standard. Assume $\mathbf{K}$ is strongly ergodic, with $\boldsymbol{\pi}^*$ denoting its steady-state distribution. Then, for any $i, j \in [d]$,

$$\lim_{n\to\infty} \left\| [\mathbf{K}^n]_{\langle i \rangle} - [\mathbf{K}^n]_{\langle j \rangle} \right\|_\infty = \lim_{n\to\infty} \|\mathbf{e}_i \mathbf{K}^n - \mathbf{e}_j \mathbf{K}^n\|_\infty \overset{(a)}{\leq} \lim_{n\to\infty} \|\mathbf{e}_i \mathbf{K}^n - \boldsymbol{\pi}^*\|_\infty + \lim_{n\to\infty} \|\mathbf{e}_j \mathbf{K}^n - \boldsymbol{\pi}^*\|_\infty \overset{(b)}{=} 0\,,$$

where (a) holds by the triangle inequality and additivity of limits, and (b) holds by strong ergodicity. Hence, $\mathbf{K}$ is weakly ergodic, as desired.

Next, we will prove the reverse direction. Assume $\mathbf{K}$ is weakly ergodic. Consider the sequence of powers $\{\mathbf{K}^n\}_{n=1}^\infty$. Since the set of all row stochastic matrices $\mathbb{R}_{\mathsf{sto}}^{d\times d}$ is compact (e.g., under the Frobenius norm), there exists a subsequence $\{\mathbf{K}^{n_\ell}\}_{\ell=1}^\infty$ which converges to a limit $\mathbf{K}^* \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$, i.e.,

$$\lim_{\ell\to\infty} \|\mathbf{K}^{n_\ell} - \mathbf{K}^*\|_{\mathrm{F}} = 0\,. \tag{191}$$

Furthermore, for any $i, j \in [d]$, we have

$$\begin{aligned}\left\|[\mathbf{K}^*]_{\langle i\rangle} - [\mathbf{K}^*]_{\langle j\rangle}\right\|_\infty &\overset{(a)}{\leq} \lim_{\ell\to\infty}\left\|[\mathbf{K}^{n_\ell}]_{\langle i\rangle} - [\mathbf{K}^*]_{\langle i\rangle}\right\|_\infty + \lim_{\ell\to\infty}\left\|[\mathbf{K}^{n_\ell}]_{\langle j\rangle} - [\mathbf{K}^*]_{\langle j\rangle}\right\|_\infty + \lim_{\ell\to\infty}\left\|[\mathbf{K}^{n_\ell}]_{\langle i\rangle} - [\mathbf{K}^{n_\ell}]_{\langle j\rangle}\right\|_\infty \\ &\overset{(b)}{=} \lim_{\ell\to\infty}\left\|[\mathbf{K}^{n_\ell}]_{\langle i\rangle} - [\mathbf{K}^{n_\ell}]_{\langle j\rangle}\right\|_\infty \overset{(c)}{=} \lim_{n\to\infty}\left\|[\mathbf{K}^{n}]_{\langle i\rangle} - [\mathbf{K}^{n}]_{\langle j\rangle}\right\|_\infty \overset{(d)}{=} 0\,,\end{aligned}$$

where (a) holds by the triangle inequality and additivity of limits, (b) holds by (191), (c) holds because the limit of any subsequence of a convergent sequence is equal to the limit of the entire sequence, and (d) holds by weak ergodicity. Hence, $\mathbf{K}^*$ has all identical rows, and so $\operatorname{rank}(\mathbf{K}^*) = 1$. Therefore, $|\lambda_1(\mathbf{K}^*)| > 0$ and $\lambda_2(\mathbf{K}^*) = \cdots = \lambda_d(\mathbf{K}^*) = 0$.

Moreover, by the Perron-Frobenius theorem [84, Chapter 1], we have $\lambda_1(\mathbf{K}^*) = 1$ and $\lambda_1(\mathbf{K}^n) = 1$ for every $n \in \mathbb{N}$. This, along with the continuity of eigenvalues with respect to the entries of a matrix [87, Chapter IV, Theorem 1.1], yields $\lim_{\ell\to\infty}\lambda_i(\mathbf{K}^{n_\ell}) = \lambda_i(\mathbf{K}^*) = 0$ for all $i \in \{2,\dots,d\}$. Since $\lambda_i(\mathbf{K}^{n_\ell}) = \lambda_i(\mathbf{K})^{n_\ell}$ for all $i \in [d]$, it follows that $|\lambda_i(\mathbf{K})| < 1$ for all $i \in \{2,\dots,d\}$. Hence, consider the Jordan canonical form $\mathbf{K} = \mathbf{X}^{-1}\mathbf{J}\mathbf{X}$ [86, Chapter 3], where the Jordan matrix $\mathbf{J}$ is lower triangular with $[\mathbf{J}]_{1,1} = 1$ and all other diagonal entries strictly less than 1 in magnitude, and the rows of $\mathbf{X}$ contain the generalized eigenvectors of $\mathbf{K}$, scaled such that $\|[\mathbf{X}]_{\langle 1\rangle}\|_1 = 1$. For any initial distribution $\boldsymbol{\pi}_0 = \mathbf{c}\mathbf{X} \in \mathscr{P}_{d-1}$, we have

$$\lim_{n\to\infty}\boldsymbol{\pi}_0\mathbf{K}^n = \lim_{n\to\infty}\left\{(\mathbf{c}\mathbf{X})\left(\mathbf{X}^{-1}\mathbf{J}^n\mathbf{X}\right)\right\} = \mathbf{c}\left(\lim_{n\to\infty}\mathbf{J}^n\right)\mathbf{X} \overset{(a)}{=} \mathbf{c}\mathbf{e}_1^{\mathrm{T}}\mathbf{e}_1\mathbf{X} = [\mathbf{c}]_1\,[\mathbf{X}]_{\langle 1\rangle}\,,$$

where (a) holds because the powers of Jordan blocks corresponding to eigenvalues with magnitude less than 1 vanish entry-wise in the limit. Finally, since $\|\boldsymbol{\pi}_0\mathbf{K}^n\|_1 = 1$ for each $n \in \mathbb{N}$, we have $[\mathbf{c}]_1 = 1$ for any choice of $\boldsymbol{\pi}_0 \in \mathscr{P}_{d-1}$,[16] and thus, $\lim_{n\to\infty}\boldsymbol{\pi}_0\mathbf{K}^n = [\mathbf{X}]_{\langle 1\rangle}$. Hence, $\mathbf{K}$ is strongly ergodic, as desired.

Proof of e) $\iff$ f): The set of all $d\times d$ row stochastic matrices $\mathbb{R}_{\mathsf{sto}}^{d\times d}$ is convex, contains the identity kernel (i.e., the $d\times d$ identity matrix), and contains a constant kernel (e.g., the $d\times d$ matrix where each row is $\mathbf{e}_1$). Thus, by Proposition 3, Part 2,

$$\mathrm{F}_{\mathbf{K}^n}\left(t;\mathbb{R}_{\mathsf{sto}}^{d\times d}\right) = \rho(\mathbf{K}^n)\,t\,. \tag{192}$$

Hence, $\mathrm{F}_{\mathbf{K}^n}\left(t;\mathbb{R}_{\mathsf{sto}}^{d\times d}\right) < t$ holds iff $\rho(\mathbf{K}^n) < 1$, or equivalently $\tau(\mathbf{K}^n) > 0$, as desired. □

Proposition 14 states that any Markov chain $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$ has positive spectral gap (i.e., $|\lambda_2(\mathbf{K})| < 1$) iff $\mathrm{F}_{\mathbf{K}^n}\left(t;\mathbb{R}_{\mathsf{sto}}^{d\times d}\right) < t$ holds for sufficiently large n. We note that taking into account whether the inequality holds for *sufficiently high power* $\mathbf{K}^n$, not necessarily $\mathbf{K}$ itself, is crucial to this equivalence, as $|\lambda_2(\mathbf{K})| < 1$ does not imply in general that the inequality holds for $n = 1$. As a counterexample, consider a directed cycle on $d \geq 4$ vertices with self-loops at each vertex:

$$\forall i,j \in [d],\ \ [\mathbf{K}]_{i,j} = \begin{cases}\frac{1}{2}\,, & \text{if } j = i\,,\\ \frac{1}{2}\,, & \text{if } j = i+1 \text{ or } (i,j) = (d,1)\,,\\ 0\,, & \text{otherwise}\,.\end{cases}$$

Clearly, $\mathbf{K}$ has positive spectral gap because it is a unichain (by virtue of its cycle structure) and aperiodic (by virtue of its self-loops). Now, let $\mathbf{w}$ and $\mathbf{v}$ be point mass distributions at vertices at least distance 2 apart on the cycle, e.g., $\mathbf{w} = \mathbf{e}_1$ and $\mathbf{v} = \mathbf{e}_3$. Then, by matrix multiplication, the distributions after one step of $\mathbf{K}$ are $\mathbf{w}\mathbf{K} = (1/2)\mathbf{e}_1 + (1/2)\mathbf{e}_2$ and $\mathbf{v}\mathbf{K} = (1/2)\mathbf{e}_3 + (1/2)\mathbf{e}_4$. It follows that

$$\rho(\mathbf{K}) \overset{(a)}{=} \rho\big([\mathbf{w},\mathbf{v}]\big)\,\rho(\mathbf{K}) \overset{(b)}{\geq} \rho\big([\mathbf{w}\mathbf{K},\mathbf{v}\mathbf{K}]\big) \overset{(c)}{=} 1\,,$$

where (a) holds because $\mathbf{w} = \mathbf{e}_1$ and $\mathbf{v} = \mathbf{e}_3$ have disjoint support and thus $\rho([\mathbf{w},\mathbf{v}]) = 1$, (b) holds by submultiplicativity of complementary Doeblin coefficients, and (c) holds because $\mathbf{w}\mathbf{K}$ and $\mathbf{v}\mathbf{K}$ have disjoint support. By (192), it follows that $\mathrm{F}_{\mathbf{K}}(t;\mathbb{R}_{\mathsf{sto}}^{d\times d}) = t$, even though $\mathbf{K}$ has positive spectral gap.

Next, we present the following proposition which exactly characterizes when contraction occurs between two input distributions after one step of the Markov chain $\mathbf{K}$.[17]

Proposition 15 (Strict Contraction of Doeblin Coefficient)**.** *Let $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$ be a Markov chain. For any subset of states $\mathcal{S} \subseteq [d]$, let $\mathcal{N}_{\mathbf{K}}(\mathcal{S})$ denote the set of states reachable from $\mathcal{S}$ in one time step, i.e.,*

$$\mathcal{N}_{\mathbf{K}}(\mathcal{S}) = \Big\{j \in [d] : \exists i \in [d],\ [\mathbf{K}]_{i,j} > 0\Big\}\,. \tag{193}$$

Then, for any pair of distributions $\mathbf{w},\mathbf{v} \in \mathscr{P}_{d-1}$, we have $\rho\big([\mathbf{w},\mathbf{v}]\,\mathbf{K}\big) < \rho\big([\mathbf{w},\mathbf{v}]\big)$ iff $\mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}>\mathbf{v}}) \cap \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}<\mathbf{v}})$ is non-empty, where we define

$$\mathcal{S}_{\mathbf{w}>\mathbf{v}} = \{i \in [d] : [\mathbf{w}]_i > [\mathbf{v}]_i\}\,,\ \ \mathcal{S}_{\mathbf{w}<\mathbf{v}} = \{i \in [d] : [\mathbf{w}]_i < [\mathbf{v}]_i\}\,. \tag{194}$$

[16] This occurs because $[\mathbf{X}]_{\langle 1\rangle}$ is a point on $\mathscr{P}_{d-1}$, and the remaining rows of $\mathbf{X}$ span the direction space associated with $\mathscr{P}_{d-1}$.

[17] Since we consider two input distributions, the Doeblin coefficient ρ reduces to TV distance.

Proof. Fix a Markov chain $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$ and distributions $\mathbf{w}, \mathbf{v} \in \mathscr{P}_{d-1}$. We have

$$\begin{aligned}\rho\big([\mathbf{w},\mathbf{v}]\,\mathbf{K}\big) &\overset{(a)}{=} 1 - \sum_{j=1}^{d} \min\Big\{[\mathbf{w}\mathbf{K}]_j\,, [\mathbf{v}\mathbf{K}]_j\Big\} \overset{(b)}{=} 1 - \sum_{j=1}^{d} \min\left\{\sum_{i=1}^{d} [\mathbf{w}]_i\,[\mathbf{K}]_{i,j}\,, \sum_{i=1}^{d} [\mathbf{v}]_i\,[\mathbf{K}]_{i,j}\right\} \\ &\overset{(c)}{\le} 1 - \sum_{j=1}^{d}\sum_{i=1}^{d} \min\Big\{[\mathbf{w}]_i\,[\mathbf{K}]_{i,j}\,, [\mathbf{v}]_i\,[\mathbf{K}]_{i,j}\Big\} \overset{(d)}{=} 1 - \sum_{i=1}^{d} \min\{[\mathbf{w}]_i\,, [\mathbf{v}]_i\} \sum_{j=1}^{d} [\mathbf{K}]_{i,j} \\ &\overset{(e)}{=} 1 - \sum_{i=1}^{d} \min\{[\mathbf{w}]_i\,, [\mathbf{v}]_i\} \overset{(f)}{=} \rho\big([\mathbf{w},\mathbf{v}]\big)\,, \end{aligned} \tag{195}$$

where (a) holds by definition of Doeblin coefficient (6), (b) holds by matrix algebra, (c) holds by the identity

$$\min\left\{\sum_i x_i, \sum_i y_i\right\} \ge \sum_i \min\{x_i, y_i\}\,, \tag{196}$$

(d) holds by factoring $[\mathbf{K}]_{i,j}$ outside the minimum (because $[\mathbf{K}]_{i,j} \ge 0$) and interchanging the order of summation, (e) holds because $\mathbf{K}$ is row stochastic, and (f) holds by definition of Doeblin coefficient (6). The inequality in (196) is strict iff $x_i > y_i$ and $x_{i'} < y_{i'}$ for some i and i'. Hence, the inequality in step (c) of (195) is strict iff there exist $i, i', j \in [d]$ such that $[\mathbf{w}]_i > [\mathbf{v}]_i$ and $[\mathbf{K}]_{i,j} > 0$ (i.e., $j \in \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}>\mathbf{v}})$), and $[\mathbf{w}]_{i'} < [\mathbf{v}]_{i'}$ and $[\mathbf{K}]_{i',j} > 0$ (i.e., $j \in \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}<\mathbf{v}})$).[18] Clearly, this is equivalent to the condition that $\mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}>\mathbf{v}}) \cap \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}<\mathbf{v}}) \neq \emptyset$, as desired. □

Proposition 15 exactly characterizes the condition on $\mathbf{K}$, $\mathbf{w}$, $\mathbf{v}$ such that strict contraction occurs in one step. We remark that two natural corollaries of this result give intuitive conditions on $\mathbf{K}$ under which strict contraction occurs for entire classes of $\mathbf{w}$ and $\mathbf{v}$.

Corollary 3 (Strict Contraction and Gramian)**.** *Let $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$ be a Markov chain. The joint range of $\mathbf{K}$ satisfies $\mathfrak{F}\big(\mathbf{K}; \mathbb{R}_{\mathsf{sto}}^{2\times d}\big) \subseteq \{(t,y) \in [0,1]^2 : y < t\} \cup \{(0,0)\}$, or, equivalently, we have $\rho\big([\mathbf{w},\mathbf{v}]\,\mathbf{K}\big) < \rho\big([\mathbf{w},\mathbf{v}]\big)$ for all pairs of distinct distributions $\mathbf{w}, \mathbf{v} \in \mathscr{P}_{d-1}$, iff the Gramian matrix $\mathbf{K}\mathbf{K}^{\mathrm{T}}$ is entry-wise strictly positive.*

Proof. Fix $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$.

First, we will prove the forward direction. Assume $\mathbf{K}\mathbf{K}^{\mathrm{T}}$ is entry-wise strictly positive. Fix any distinct distributions $\mathbf{w}, \mathbf{v} \in \mathscr{P}_{d-1}$. Recall the definitions of $\mathcal{S}_{\mathbf{w}>\mathbf{v}}$ and $\mathcal{S}_{\mathbf{w}<\mathbf{v}}$ from (194). Since probability distributions sum to 1, both $\mathcal{S}_{\mathbf{w}>\mathbf{v}}$ and $\mathcal{S}_{\mathbf{w}<\mathbf{v}}$ are non-empty. Consider any $i \in \mathcal{S}_{\mathbf{w}>\mathbf{v}}$ and $i' \in \mathcal{S}_{\mathbf{v}<\mathbf{w}}$. By assumption, we have $[\mathbf{K}\mathbf{K}^{\mathrm{T}}]_{i,i'} = \sum_{j=1}^{d} [\mathbf{K}]_{i,j}[\mathbf{K}]_{i',j} > 0$, so there exists $j^* \in [d]$ such that $[\mathbf{K}]_{i,j^*} > 0$ and $[\mathbf{K}]_{i',j^*} > 0$. By (193), it holds that $j^* \in \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}>\mathbf{v}})$ and $j^* \in \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}<\mathbf{v}})$, and so $\mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}>\mathbf{v}}) \cap \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}<\mathbf{v}}) \neq \emptyset$. By Proposition 15, we have $\rho([\mathbf{w},\mathbf{v}]\,\mathbf{K}) < \rho([\mathbf{w},\mathbf{v}])$ as desired.

Next, we will prove the reverse direction. Assume that $\rho([\mathbf{w},\mathbf{v}]\,\mathbf{K}) < \rho([\mathbf{w},\mathbf{v}])$ for any distinct distributions $\mathbf{w}, \mathbf{v} \in \mathscr{P}_{d-1}$. Fix unspecified distinct $i, i' \in [d]$. Choosing $\mathbf{w} = \mathbf{e}_i$ and $\mathbf{v} = \mathbf{e}_{i'}$ and applying Proposition 15, we have $\mathcal{N}_{\mathbf{K}}(\{i\}) \cap \mathcal{N}_{\mathbf{K}}(\{i'\}) \neq \emptyset$. Consider any $j^* \in \mathcal{N}_{\mathbf{K}}(\{i\}) \cap \mathcal{N}_{\mathbf{K}}(\{i'\})$. We have

$$\big[\mathbf{K}\mathbf{K}^{\mathrm{T}}\big]_{i,i'} = \sum_{j=1}^{d} [\mathbf{K}]_{i,j}\,[\mathbf{K}]_{i',j} \ge [\mathbf{K}]_{i,j^*}\,[\mathbf{K}]_{i',j^*} \overset{(a)}{>} 0\,,$$

where (a) holds by definition of $\mathcal{N}_{\mathbf{K}}$ from (193), and so the off-diagonal entries of $\mathbf{K}\mathbf{K}^{\mathrm{T}}$ are positive. Furthermore, for any $i \in [d]$, we have $[\mathbf{K}\mathbf{K}^{\mathrm{T}}]_{i,i} = \sum_{j=1}^{d} [\mathbf{K}]_{i,j}^2 > 0$, where the final inequality holds because $\mathbf{K}$ is row stochastic, and so the diagonal entries of $\mathbf{K}\mathbf{K}^{\mathrm{T}}$ are positive. Hence, $\mathbf{K}\mathbf{K}^{\mathrm{T}}$ is entry-wise strictly positive, as desired. □

Corollary 4 (Strict Contraction and Laziness)**.** *Let $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$ be a lazy Markov chain (i.e., $[\mathbf{K}]_{i,i} > 0$ for all $i \in [d]$) with positive spectral gap (i.e., $|\lambda_2(\mathbf{K})| < 1$). Then, for all pairs of distributions $\mathbf{w}, \mathbf{v} \in \mathscr{P}_{d-1}$ which differ at each entry (i.e., $[\mathbf{w}]_i \neq [\mathbf{v}]_i$ for all $i \in [d]$), we have $\rho\big([\mathbf{w},\mathbf{v}]\,\mathbf{K}\big) < \rho\big([\mathbf{w},\mathbf{v}]\big)$.*

Proof. Fix a Markov chain $\mathbf{K} \in \mathbb{R}_{\mathsf{sto}}^{d\times d}$ and distributions $\mathbf{w}, \mathbf{v} \in \mathscr{P}_{d-1}$ satisfying the conditions in Corollary 4. Since $\mathbf{w}$ and $\mathbf{v}$ differ at each entry, we have $\mathcal{S}^{\mathrm{c}}_{\mathbf{w}>\mathbf{v}} = \mathcal{S}_{\mathbf{w}<\mathbf{v}}$. Since $\mathbf{K}$ has positive spectral gap, $\mathbf{K}$ has exactly one recurrent class (by Proposition 14, Part (b)), and so the directed graph representation of $\mathbf{K}$ is connected (in an undirected sense). Hence, there exists an edge between $\mathcal{S}_{\mathbf{w}>\mathbf{v}}$ and $\mathcal{S}^{\mathrm{c}}_{\mathbf{w}>\mathbf{v}}$, i.e., for some $i \in \mathcal{S}_{\mathbf{w}>\mathbf{v}}$ and $i' \in \mathcal{S}_{\mathbf{w}<\mathbf{v}}$, we have $[\mathbf{K}]_{i,i'} > 0$ or $[\mathbf{K}]_{i',i} > 0$. Without loss of generality, assume $[\mathbf{K}]_{i,i'} > 0$. Since $\mathbf{K}$ is lazy, we have $[\mathbf{K}]_{i',i'} > 0$. By (193), it holds that $i' \in \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}>\mathbf{v}})$ and $i' \in \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}<\mathbf{v}})$, and so $\mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}>\mathbf{v}}) \cap \mathcal{N}_{\mathbf{K}}(\mathcal{S}_{\mathbf{w}<\mathbf{v}}) \neq \emptyset$. By Proposition 15, we have $\rho([\mathbf{w},\mathbf{v}]\,\mathbf{K}) < \rho([\mathbf{w},\mathbf{v}])$ as desired. □

[18] This condition for equality can also be verified by using the ℓ^1-norm characterization of TV distance in (a), applying the triangle inequality to deduce (e), and then using known conditions for equality in the triangle inequality.

REFERENCES

[1] W. Lu, A. Makur, and J. Singh, "On Doeblin curves and their properties," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Athens, Greece, July 7–12 2024, pp. 2544–2549.

[2] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.

[3] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 2, no. 1–4, pp. 191–213, March 1972.

[4] H. O. Hirschfeld, "A connection between correlation and contingency," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 31, no. 4, pp. 520—-524, October 1935.

[5] R. L. Dobrushin, "Central limit theorem for nonstationary Markov chains. i," *Theory of Probability & Its Applications*, vol. 1, no. 1, pp. 65–80, 1956.

[6] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3–4, pp. 441–451, September 1959.

[7] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *The Annals of Probability*, vol. 4, no. 6, pp. 925–939, December 1976.

[8] J. E. Cohen, J. H. B. Kemperman, and G. Zbăganu, *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population Sciences*. Boston, MA, USA: Birkhäuser, 1998.

[9] W. S. Evans and L. J. Schulman, "Signal propagation and noisy circuits," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2367–2373, November 1999.

[10] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and a data processing inequality," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Honolulu, HI, USA, June 29–July 4 2014, pp. 3022–3026.

[11] A. Makur and L. Zheng, "Bounds between contraction coefficients," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 29–October 2 2015, pp. 1422–1429.

[12] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, June 2016.

[13] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and Bayesian networks," in *Convexity and Concentration*, 2017, pp. 211–249.

[14] F. du Pin Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities for input constrained additive noise channels," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1879–1892, March 2018.

[15] A. Makur, "Information contraction and decomposition," Sc.D. Thesis in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, May 2019.

[16] A. Makur and L. Zheng, "Comparison of contraction coefficients for f-divergences," *Problems of Information Transmission*, vol. 56, no. 2, pp. 103–156, April 2020.

[17] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "Universal features for high-dimensional learning and inference," *Foundations and Trends in Communications and Information Theory*, vol. 21, no. 1–2, pp. 1–299, February 2024.

[18] D. Lee and A. Makur, "Contraction coefficients of product symmetric channels," in *Proceedings of the 60th Annual Allerton Conference on Communication, Control, and Computing*, Urbana, IL, USA, September 24–27 2024, pp. 1–8.

[19] A. Makur and J. Singh, "Doeblin coefficients and related measures," *IEEE Transactions on Information Theory*, vol. 70, no. 7, pp. 4667–4692, July 2024.

[20] W. Doeblin, "Sur les proprietes asymptotiques de mouvement régis par certains types de chaînes simples," *Bulletin Mathématique de la Société Roumaine des Sciences*, vol. 39, no. 1, pp. 57–115, 1937, in French.

[21] W. Doeblin, "Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états," *Revue Mathématique de l'Union Interbalkanique*, vol. 2, pp. 77–105, 1938, in French.

[22] M. E. Lladser and S. R. Chestnut, "Approximation of sojourn-times via maximal couplings: motif frequency distributions," *Journal of Mathematical Biology*, vol. 69, no. 1, pp. 147–182, July 2014.

[23] G. O. Roberts and J. S. Rosenthal, "General state space Markov chains and MCMC algorithms," *Probability Surveys*, vol. 1, pp. 20–71, 2004.

[24] 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.

[25] 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, January 2016.

[26] S. R. Kimbleton, "A stable limit theorem for Markov chains," *The Annals of Mathematical Statistics*, vol. 40, no. 4, pp. 1467–1473, August 1969.

[27] R. Bhattacharya and E. C. Waymire, "Iterated random maps and some classes of Markov processes," in *Stochastic Processes: Theory and Methods*, ser. Handbook of Statistics, D. N. Shanbhag and C. R. Rao, Eds. Amsterdam, Netherlands: Elsevier, 2001, vol. 19, pp. 145–170.

[28] S. Chestnut and M. E. Lladser, "Occupancy distributions in Markov chains via Doeblin's ergodicity coefficient," in *Proceedings of the 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms*, Vienna, Austria, June 28–July 2 2010, pp. 79–92.

[29] A. Gohari, O. Günlü, and G. Kramer, "Coding for positive rate in the source model key agreement problem," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6303–6323, October 2020.

[30] A. Makur, "Coding theorems for noisy permutation channels," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 6723–6748, November 2020.

[31] H. Chen, J. Tang, and A. Gupta, "Change detection of Markov kernels with unknown pre and post change kernel," in *Proceedings of the 61st IEEE Conference on Decision and Control (CDC)*, Cancún, Mexico, December 6–9 2022, pp. 4814–4820.

[32] V. Moulos, "Finite-time analysis of round-robin Kullback-Leibler upper confidence bounds for optimal adaptive allocation with multiple plays and Markovian rewards," in *Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS)*, December 6–12 2020, pp. 7863–7874.

[33] J. S. Rosenthal, "Minorization conditions and convergence rates for Markov chain Monte Carlo," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 558–566, June 1995.

[34] W. Doeblin, "Éléments d'une théorie générale des chaînes simples constantes de Markoff," *Annales scientifiques de l'École Normale Supérieure, Serie 3*, vol. 57, pp. 61–111, 1940.

[35] I. Kontoyiannis and Y. M. Suhov, "Prefixes and the entropy rate for long-range sources," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Trondheim, Norway, June 27–July 1 1994, p. 194.

[36] A. Makur and Y. Polyanskiy, "Comparison of channels: Criteria for domination by a symmetric channel," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5704–5725, August 2018.

[37] A. A. Gushchin, "On an extension of the notion of f-divergence," *Theory of Probability & Its Applications*, vol. 52, no. 3, pp. 439–455, 2008.

[38] R. C. Williamson and Z. Cranko, "Information processing equalities and the information–risk bridge," *Journal of Machine Learning Research*, vol. 25, no. 103, 2024.

[39] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1625–1657, March 2020.

[40] A. Makur and J. Singh, "Bounds on maximal leakage over Bayesian networks," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Ann Arbor, MI, USA, June 22–27 2025, pp. 1–6.

[41] H. Wang, R. Gao, and F. P. Calmon, "Generalization bounds for noisy iterative algorithms using properties of additive noise channels," *Journal of Machine Learning Research*, vol. 24, no. 1, pp. 1–43, January 2023.

[42] S. Asoodeh, M. Aliakbarpour, and F. P. Calmon, "Local differential privacy is equivalent to contraction of an f-divergence," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Melbourne, Australia, July 12–20 2021, pp. 545–550.

[43] M. Diaz, L. Paull, and P. S. Castro, "LOCO: Adaptive exploration in reinforcement learning via local estimation of contraction coefficients," in *Self-Supervision for Reinforcement Learning Workshop - ICLR*, May 7 2021, pp. 1–18.

[44] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, ser. Grundlehren Text Editions. Berlin, Heidelberg, Germany: Springer, 2001.

[45] E. Çinlar, *Probability and Stochastics*. New York, NY, USA: Springer, 2011.

[46] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge, UK: Cambridge University Press, 2025.

[47] H. Thorisson, *Coupling, Stationarity, and Regeneration*. New York, NY, USA: Springer New York, 2000.

[48] A. Makur and J. Singh, "On properties of Doeblin coefficients," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, June 25–30 2023, pp. 72–77.

[49] O. Kallenberg, *Foundations of Modern Probability*. Cham, Switzerland: Springer, 1997.

[50] I. M. Gelfand, A. N. Kolmogorov, and A. M. Yaglom, "Towards the general definition of the amount of information," *Doklady Akademii Nauk SSSR*, vol. 111, pp. 745–748, 1956, in Russian.

[51] G. L. Gilardoni, "On a Gel'fand-Yaglom-Peres theorem for f-divergences," November 2009, arXiv:0911.1934 [cs.IT]. [Online]. Available: https://arxiv.org/abs/0911.1934

[52] H. Jung, "Ueber die kleinste kugel, die eine räumliche figur einschliesst." *Journal für die reine und angewandte Mathematik*, vol. 123, pp. 241–257, 1901.

[53] Y. Q. Yan, "On the exact value of Jung constants of Orlicz sequence spaces," *Collectanea Mathematica*, vol. 55, no. 2, pp. 163–170, 2004.

[54] F. Bohnenblust, "Convex regions and projections in Minkowski spaces," *Annals of Mathematics*, vol. 39, no. 2, pp. 301–308, April 1938.

[55] S. Umarov, C. Tsallis, and S. Steinberg, "On a q-central limit theorem consistent with nonextensive statistical mechanics," *Milan Journal of Mathematics*, vol. 76, pp. 307–328, 2008.

[56] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis," in *Proceedings of the 2017 Conference on Learning Theory*, Amsterdam, Netherlands, July 7–10 2017, pp. 1674–1703.

[57] J. von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components," in *Automata Studies*, ser. Annals of Mathematics Studies, C. E. Shannon and J. McCarthy, Eds. Princeton, NJ, USA: Princeton University Press, 1956, vol. 34, pp. 43–98.

[58] B. Hajek and T. Weller, "On the maximum tolerable noise for reliable computation by formulas," *IEEE Transactions on Infomation Theory*, vol. 37, no. 2, pp. 388–391, March 1991.

[59] W. S. Evans and L. J. Schulman, "On the maximum tolerable noise of k-input gates for reliable computation by formulas," *IEEE Transactions on Information Theory*, vol. 49, no. 11, pp. 3094–3098, November 2003.

[60] A. K. Tan, M. H. Ho, and I. L. Chuang, "Reliable computation by large-alphabet formulas in the presence of noise," *IEEE Transactions on Information Theory*, vol. 70, no. 12, December 2024.

[61] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, 1st ed. Providence, RI, USA: American Mathematical Society, 2009.

[62] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[63] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, San Diego, CA, USA, June 9–11 2003, pp. 211–222.

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

[65] 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, March 2021.

[66] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France, July 6–11 2015, pp. 1376–1385.

[67] S. Asoodeh, M. Diaz, and F. P. Calmon, "Privacy amplification of iterative algorithms via contraction coefficients," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Los Angeles, CA, USA, June 21–26 2020, pp. 896–901.

[68] M. H. DeGroot, "Uncertainty, information, and sequential experiments," *The Annals of Mathematical Statistics*, vol. 33, no. 2, June 1962.

[69] P. Hajłasz and J. Malý, "Approximation in Sobolev spaces of nonlinear expressions involving the gradient," *Arkiv för Matematik*, vol. 40, no. 2, pp. 245–274, October 2002.

[70] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2009.

[71] Y. Chu and M. Raginsky, "Talagrand meets Talagrand: Upper and lower bounds on expected soft maxima of Gaussian processes with finite index sets," February 2025, arXiv:2502.06709 [math.PR]. [Online]. Available: https://arxiv.org/abs/2502.06709

[72] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, UK: Cambridge University Press, 2019.

[73] A. Baranga, "The contraction principle as a particular case of Kleene's fixed point theorem," *Discrete Mathematics*, vol. 98, no. 1, pp. 75–79, December 1991.

[74] O. Kallenberg, *Foundations of Modern Probability*, 3rd ed. Springer, 2021.

[75] V. Feldman and T. Zrnic, "Individual privacy accounting via a Rényi filter," in *Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS)*, December 6–14 2021, pp. 28 080–28 091.

[76] D. Levy, Z. Sun, K. Amin, S. Kale, A. Kulesza, M. Mohri, and A. T. Suresh, "Learning with user-level privacy," in *Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS)*, December 6–14 2021, pp. 12 466–12 479.

[77] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York, NY, USA: McGraw-Hill, 1976.

[78] R. Van Handel, "On the spectral norm of Gaussian random matrices," *Transactions of the American Mathematical Society*, vol. 369, no. 11, pp. 8161–8178, November 2017.

[79] R. Vershynin, *High-Dimensional Probability*, 1st ed. Cambridge, UK: Cambridge University Press, 2018.

[80] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman, "Broadcasting on trees and the Ising model," *The Annals of Applied Probability*, vol. 10, no. 2, pp. 410–433, May 2000.

[81] A. Makur, E. Mossel, and Y. Polyanskiy, "Broadcasting on two-dimensional regular grids," *IEEE Transactions on Information Theory*, vol. 68, no. 10, pp. 6297–6334, October 2022.

[82] R. G. Gallager, *Stochastic Processes: Theory for Applications*. New York, NY, USA: Cambridge University Press, 2013.

[83] S. Anily and A. Federgruen, "Ergodicity in parametric nonstationary Markov chains: An application to simulated annealing methods," *Operations Research*, vol. 35, no. 6, pp. 867–874, December 1987.

[84] E. Seneta, *Non-Negative Matrices and Markov Chains*, 2nd ed. New York, NY, USA: Springer, 1981.

[85] J. L. R. Alfonsín, *The Diophantine Frobenius Problem*. Oxford, UK: Oxford University Press, 2005.

[86] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge University Press, 2013.

[87] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*. Cambridge, MA, USA: Academic Press, 1990.