On Properties of Doeblin Coefficients

Anuran Makur* and Japneet Singh†
*Department of Computer Science and *,†School of Electrical and Computer Engineering,
Purdue University, West Lafayette, IN 47907
Email: {amakur, sing1041}@purdue.edu

***Abstract*—Doeblin coefficients are a classical tool to study the ergodicity of Markov chains. Propelled by recent works on contraction coefficients of strong data processing inequalities, we investigate whether Doeblin coefficients also exhibit some of the notable properties of canonical contraction coefficients. Specifically, we present various new structural and geometric properties of Doeblin coefficients. Then, by establishing an extremal coupling characterization, we show that Doeblin coefficients generalize the well-known total variation (TV) distance to a multi-way divergence, enabling us to measure the distance between multiple distributions rather than just two. We also demonstrate that Doeblin coefficients exhibit contraction properties over Bayesian networks similar to other canonical contraction coefficients. Finally, we discuss how Doeblin coefficients can be used to define a new rule for fusion of probability mass functions.**

I. Introduction

Recently, there has been a flurry of research activity on *strong data processing inequalities* (SDPIs), which provide quantitative bounds on the contraction of information in channels or Markov kernels as measured by various f-divergences. Specifically, such contraction of information is mathematically captured by so-called *contraction coefficients* and has been studied extensively, cf. [1]–[5]. Furthermore, the resulting ideas have been used in distributed estimation [6], differential privacy [7], [8], and secret key generation [9], among other applications. In [2], it was shown that contraction coefficients for Kullback-Leibler (KL) divergence of a given channel are defined by the extremal erasure probabilities so that an erasure channel dominates the given channel in the *less noisy* preorder sense [10]. This observation has been generalized for a broader class of f-divergences in [4], and to broader classes of dominating channels such as symmetric channels in [3]. In this paper, we study the extremal erasure probabilities so that an erasure channel dominates a given channel in the (output) *degradation* sense, cf. [3], [11], [12]. These extremal erasure probabilities are known as *Doeblin coefficients* [13], [14]. In particular, our objective is to illustrate that Doeblin coefficients share most of the nice properties of canonical contraction coefficients, such as their behavior over Bayesian networks.

Formally, let $\mathbb{R}_{\mathsf{sto}}^{n\times m}$ be the set of all $n \times m$ row stochastic matrices, or equivalently, channels with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ such that $|\mathcal{X}| = n$ and $|\mathcal{Y}| = m$. Then, for any channel $W \in \mathbb{R}_{\mathsf{sto}}^{n\times m}$, its *Doeblin coefficient* of ergodicity $\tau(W)$ is defined as

$$\tau(W) \triangleq \sum_{j=1}^{m} \min_{i\in\{1,\ldots,n\}} W_{ij}. \tag{1}$$

The author ordering is alphabetical.

Classically, the Doeblin coefficient was used in the study of ergodicity of Markov chains. Indeed, for a given channel (or Markov kernel) W, its Doeblin coefficient is the largest constant $\alpha \in [0, 1]$ such that the *Doeblin minorization* condition holds [15], i.e., there exists a probability distribution μ over $\mathcal{Y}$ such that

$$W_{ij} \geq \alpha\mu_j \tag{2}$$

for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. A Markov chain with a large Doeblin coefficient is said to be "well-minorized," because it can be shown that the chain converges quickly to its invariant distribution.

A. Related Literature

As mentioned above, Doeblin minorization was originally developed to study the ergodicity of Markov chains. In [15], Doeblin characterized the conditions for weak ergodicity for possibly inhomogeneous Markov chains. In [16], Doeblin introduced coupling as a technique for proving uniform geometric rates of convergence of Markov chains to their stationary distributions in terms of *total variation (TV) distance* (see [17] for more details). The Doeblin minorization condition has also been studied in the context of information theory, where it was used to derive universal upper bounds on contraction coefficients for any f-divergence [1, Remark 3.2], [3, Section I-D]. In fact, Doeblin minorization is known to be equivalent to degradation by erasure channels (see [17, Theorem 3.1], [14, Lemma 5] and [18, Section IV-D] for more details). In the theory of Harris chains, this idea can be viewed as a specialization of the *regeneration* or Nummelin splitting technique [19], [20]. Moreover, this is why Doeblin coefficients can be characterized using degradation by erasure channels [13], [14]. As noted earlier, in this sense, Doeblin coefficients can be construed as extensions of the broader theory of contraction coefficients [21], [22, Definition 5.1], [5, Section 3.2], which are characterized using domination by erasure channels under the less noisy preorder. Several properties of Doeblin coefficients have been explored in [13]. Finally, in a different vein, we will show that Doeblin coefficients have the flavor of *multi-way divergences*, which are measures of discrimination between two or more probability distributions. We refer readers to [23] and the references therein for a detailed overview of multi-way divergences and the references therein.

B. Main Contributions

In this paper, we delve into various structural and geometric properties of Doeblin coefficients. Firstly, we collect several

known properties and prove various new properties of Doeblin coefficients in Section II for the readers' convenience. For example, we show that the Doeblin coefficient has desirable properties that make it suitable for measuring some notion of "distance" between multiple distributions. In particular, it generalizes TV distance to a multi-way metric, which is an interesting observation in its own right. Secondly, we study the information contraction properties of Doeblin coefficients over Bayesian networks (or directed graphical models) in Section IV. To do this, we develop several extremal-coupling-based characterizations of Doeblin coefficients in Section III that generalize known maximal coupling and simultaneous coupling results for TV distance. Finally, on a more applied front, we briefly illustrate how Doeblin coefficients can be used to design a new notion of fusion or aggregation of probability mass functions (PMFs) in Section V. Specifically, we show that Doeblin coefficients allow us to develop an optimization-based approach for the aggregation of PMFs.

II. Properties of Doeblin Coefficients

In this section, we introduce several new properties of the Doeblin coefficient, starting with a brief review of the notation used throughout this work.

A. Notation

We briefly collect some notation that is used throughout this work. For $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, 2, \ldots, n\}$. Denote $\mathbf{1}$ as a column vector with all entries equal to 1 of appropriate dimension. Let $\mathcal{P}_n$ denote the $(n-1)$-dimensional probability simplex of row vectors in $\mathbb{R}^n$, i.e., the set $\{P \in \mathbb{R}^n : P \geq 0 \text{ entry-wise and } P\mathbf{1} = 1\}$. For an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ with $|\mathcal{X}| = n$ and $|\mathcal{Y}| = m$, let X and Y be random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, and let $\mathbb{R}_{\mathsf{sto}}^{n\times m}$ denote the set of all $n \times m$ row stochastic matrices in $\mathbb{R}^{n\times m}$, which are conditional distributions of Y given X. For any matrix $W \in \mathbb{R}_{\mathsf{sto}}^{n\times m}$ such that $W = \begin{bmatrix} P_1^{\mathrm{T}} & P_2^{\mathrm{T}} & \cdots & P_n^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ is formed by stacking the row vectors $P_1, P_2, \ldots, P_n \in \mathcal{P}_m$, we will interchangeably use the notation $\tau(P_1, P_2, \ldots, P_n)$ to denote the Doeblin coefficient $\tau(W)$.

B. Properties

The following theorem summarizes various properties of Doeblin coefficients.

Theorem 1 (Properties of Doeblin Coefficients): Let $W = \begin{bmatrix} P_1^{\mathrm{T}} & \cdots & P_n^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}_{\mathsf{sto}}^{n\times m}$ be a channel formed by stacking the PMFs $P_1, \ldots, P_n \in \mathcal{P}_m$. Then, the Doeblin coefficient $\tau(W)$ of W satisfies the following properties:

1) *(Normalization)* We have $0 \leq \tau(W) \leq 1$, where $\tau(W) = 1$ if and only if $P_1 = \cdots = P_n$ (entry-wise), and $\tau(W) = 0$ if and only if at least one of $P_1(y), \ldots, P_n(y)$ is zero for all $y \in \mathcal{Y}$.
2) *(n-way Metric)* The complement of the Doeblin coefficient $\gamma(P_1, \ldots, P_n) \triangleq 1 - \tau(P_1, \ldots, P_n) \in [0, 1]$ is an n-way metric, i.e., it satisfies the following properties:
   a) *(Total Symmetry)* For any permutation π of $[n]$, we have
   $$\gamma(P_{\pi(1)}, \ldots, P_{\pi(n)}) = \gamma(P_1, \ldots, P_n).$$
   b) *(Positive Definiteness)* $\gamma(P_1, \ldots, P_n) = 0$ if and only if $P_1 = \cdots = P_n$.
   c) *(Polyhedron Inequality)* For any $P_{n+1} \in \mathcal{P}_m$, $\gamma(P_1, \ldots, P_n)$ satisfies the inequality
   $$(n-1)\gamma(P_1, \ldots, P_n) \leq \sum_{i=1}^{n} \gamma(P_1, \ldots, P_{i-1}, P_{i+1}, \ldots, P_{n+1}).$$
3) *(Concavity)* The map $W \mapsto \tau(W)$ is concave.
4) *(Sub-multiplicativity [13])* For channels $V \in \mathbb{R}_{\mathsf{sto}}^{k\times n}$ and $W \in \mathbb{R}_{\mathsf{sto}}^{n\times m}$, the complement of the Doeblin coefficient is sub-multiplicative:
   $$1 - \tau(VW) \leq (1 - \tau(V))(1 - \tau(W)).$$
5) *(Tensorization)* For channels $W \in \mathbb{R}_{\mathsf{sto}}^{n\times m}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{l\times k}$, we have
   $$\tau(W \otimes V) = \tau(W)\tau(V),$$
   where $\otimes$ denotes the Kronecker product of matrices.
6) *(Upper Bound [1])* $\tau(W)$ satisfies the following bound:
   $$\tau(W) \leq 1 - \eta_{\mathsf{TV}}(W),$$
   where $\eta_{\mathsf{TV}}(W) = \frac{1}{2} \max_{i,j\in[n]} \|P_i - P_j\|_1$ denotes the Dobrushin contraction coefficient for TV distance and $\|\cdot\|_1$ is the ℓ^1-norm.
7) *(Minimum Trace Characterization)* $\tau(W)$ is the solution to the following optimization problem:
   $$\tau(W) = \min_{P \in \mathbb{R}_{\mathsf{sto}}^{m\times n}} \mathsf{Tr}(PW),$$
   where the minimum is over all possible $m \times n$ row stochastic matrices P, and $\mathsf{Tr}(\cdot)$ denotes the trace of a matrix.
8) *(Optimal Estimator)* Consider a hidden random variable X that is uniformly distributed on $\mathcal{X}$, i.e., $X \sim \mathrm{unif}(\mathcal{X})$, and the fixed channel (or observation model) $W = P_{Y|X} \in \mathbb{R}_{\mathsf{sto}}^{n\times m}$. Let $\hat{X} \in \mathcal{X}$ denote any (possibly randomized) estimator of X based on the observation Y, which is defined by a kernel $P_{\hat{X}|Y} \in \mathbb{R}_{\mathsf{sto}}^{m\times n}$ such that $X \to Y \to \hat{X}$ forms a Markov chain. Then, we have
   $$\frac{\tau(P_{Y|X})}{n} = \min_{P_{\hat{X}|Y} \in \mathbb{R}_{\mathsf{sto}}^{m\times n}} \mathbb{P}(\hat{X} = X),$$
   where the minimum is computed over all estimators $\hat{X}$, or equivalently, over all kernels $P_{\hat{X}|Y} \in \mathbb{R}_{\mathsf{sto}}^{m\times n}$, and $\mathbb{P}(\cdot)$ denotes the probability law of $(X, Y, \hat{X})$.

The proof is provided in the extended version of this paper [24]. Properties 1 and 3 in Theorem 1 are straightforward to prove (and property 1 is well-known). Property 4 was already proved in [13], but we provide a new proof. The upper bound in property 6 is also well-known, cf. [1, Remark III.2]. The rest of the properties are new to our knowledge. It is worth reiterating that, interestingly, the complement of the Doeblin coefficient $\gamma(\cdot)$ acts as a multi-way metric, which allows us to measure "distance" between more than two distributions. In Theorem 2, we will establish that $\gamma(\cdot)$ is actually a multi-way generalization of TV distance.

III. Extremal Coupling Characterizations

In this section, we develop several extremal-coupling-based characterizations of Doeblin coefficients that generalize known maximal coupling and simultaneous coupling results for TV distance. In particular, as mentioned earlier, our maximal coupling result will show that the complement of the Doeblin coefficient $\gamma(P_1, \ldots, P_n)$ is in fact a multi-way generalization of TV distance between multiple probability distributions.

A. Maximal Coupling

Recall that for random variables X and Y taking values in some common alphabet $\mathcal{X}$ with probability distributions P and Q, respectively, a *coupling* of P and Q is a joint distribution $P_{X,Y}$ on the product space $\mathcal{X} \times \mathcal{X}$ such that the corresponding marginals distributions of X and Y are equal to P and Q, respectively. A *maximal coupling* of two distributions P and Q on $\mathcal{X}$ is a coupling that maximizes the probability that the two random variables are equal. It is well-known that the TV distance between P and Q can be characterized by Dobrushin's maximal coupling of P and Q [25, Prop. 4.7]

$$1 - \|P - Q\|_{\mathsf{TV}} = \max_{P_{X,Y}:\, P_X = P,\, P_Y = Q} \mathbb{P}(X = Y), \quad (3)$$

where $\|\cdot\|_{\mathsf{TV}}$ denotes the TV distance, the maximum is over all couplings of P and Q, and $\mathbb{P}(\cdot)$ is the probability law corresponding to the coupling $P_{X,Y}$. Moreover, under the maximal coupling, for any $x \in \mathcal{X}$,

$$\mathbb{P}(X = Y = x) = \min\{P(x), Q(x)\}. \quad (4)$$

We will now generalize the result for the case where we have n distributions $P_1, P_2, \ldots, P_n \in \mathcal{P}_m$. A coupling of random variables $X_1, \ldots, X_n$ distributed as $P_1, \ldots, P_n$ is defined as a joint distribution $P_{X_1,\ldots,X_n}$ that preserves the marginals. Specifically, under the maximal coupling of $P_1, P_2, \ldots, P_n$, we will show that the measure of the "diagonal set" of $\mathcal{X}^n$ equals $\tau(P_1, \ldots, P_n)$.

Theorem 2 (Maximal Coupling): For any random variables $X_1, \ldots, X_n \in \mathcal{Y}$ distributed according to the PMFs $P_1, \ldots, P_n \in \mathcal{P}_m$, respectively, let $W \in \mathbb{R}^{n\times m}_{\mathsf{sto}}$ be a channel defined by $W = \left[P_1^{\mathrm{T}}\, P_2^{\mathrm{T}}\, \cdots\, P_n^{\mathrm{T}}\right]^{\mathrm{T}}$. Then, we have

$$\tau(W) = 1 - \gamma(W) = \max_{\substack{P_{X_1,\ldots,X_n}:\\ P_{X_1} = P_1,\, \ldots,\, P_{X_n} = P_n}} \mathbb{P}(X_1 = \cdots = X_n),$$

where the maximum is over all couplings of $P_1, \ldots, P_n$, and $\mathbb{P}(\cdot)$ denotes the probability law corresponding to a coupling. Moreover, under the maximal coupling, for any $x \in \mathcal{Y}$

$$\mathbb{P}(X_1 = \cdots = X_n = x) = \min\{P_1(x), \ldots, P_n(x)\}.$$

The proof is provided in [24]. Theorem 2 illustrates that $\gamma(\cdot)$ is a generalization of TV distance and for the special case of $n = 2$, $\gamma(\cdot)$ reduces to the TV distance. Moreover, by property 2 in Theorem 1, $\gamma(\cdot)$ is a n-way metric just like TV distance is a metric between two distributions.

In [1], it was shown that the TV distance could be used to define *DeGroot distance* [26], which is a measure of Bayes statistical information. We will now generalize the DeGroot distance to the case when we have multiple probability distributions. We show that we get two different notions of statistical information: min-DeGroot and max-DeGroot distance.

1) Min-DeGroot Distance: Given any prior PMF $\lambda \in \mathcal{P}_n$, consider a hidden random variable $X \in \mathcal{X}$ and an observed random variable $Y \in \mathcal{Y}$ such that $|\mathcal{X}| = n$, $|\mathcal{Y}| = m$, and

$$X \sim \lambda \text{ and } P_{Y|X=x_i} = P_i \quad (5)$$

for $i \in [n]$, where $P_1, \ldots, P_n \in \mathcal{P}_m$ determine the channel (or observation model) $P_{Y|X}$, and we let $\mathcal{X} = \{x_1, \ldots, x_n\}$ without loss of generality. Let $\hat{X} \in \mathcal{X}$ denote any (possibly randomized) estimator of X based on Y, which is defined by a kernel $P_{\hat{X}|Y} \in \mathbb{R}^{m\times n}_{\mathsf{sto}}$ such that $X \to Y \to \hat{X}$ forms a Markov chain. Suppose the goal is to minimize the *Bayes risk* for the loss function $l(X, \hat{X}) = \mathbb{1}_{\hat{X}=X}$, where $\mathbb{1}$ denotes the indicator function. When we have no observations related to X, the Bayes optimal estimator is to choose the least likely $x \in \mathcal{X}$, i.e., $\hat{X}^* = x_{i^*}$ with $i^* = \arg\min_{i\in[n]} \lambda_i$, and the corresponding Bayes risk R^*_λ is

$$R^*_\lambda = \min\{\lambda_1, \ldots, \lambda_n\}. \quad (6)$$

On the other hand, when Y is observed, the Bayes optimal estimator $\hat{X}^*$ is the solution to the following problem:

$$\hat{X}^* = \arg\min_{P_{\hat{X}|Y} \in \mathbb{R}^{m\times n}_{\mathsf{sto}}} \mathbb{P}(\hat{X} = X), \quad (7)$$

where the minimum is over all possible estimators. In statistical decision theory, is it well-known that for an observation model $P_{Y|X} \in \mathbb{R}^{n\times m}_{\mathsf{sto}}$, prior PMF $P_X \in \mathcal{P}_n$, loss function $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and randomized estimator $P_{\hat{X}|Y} \in \mathbb{R}^{m\times n}_{\mathsf{sto}}$, the risk function is defined as

$$\begin{aligned} R_{P_X}(P_{Y|X}, l, P_{\hat{X}|Y}) &\triangleq \mathbb{E}[l(X, \hat{X})] \\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \sum_{\hat{x}\in\mathcal{X}} l(x, \hat{x}) P_X(x) P_{Y|X}(y|x) P_{\hat{X}|Y}(\hat{x}|y) \\ &= \mathsf{Tr}(L^{\mathrm{T}} \operatorname{diag}(P_X) P_{Y|X} P_{\hat{X}|Y}), \end{aligned} \quad (8)$$

where $L \in \mathbb{R}^{n\times n}$ is the matrix representation of the loss function $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e., the (i, j)th entry of L is $l(x_i, \hat{x}_j)$ for $i, j \in [n]$ and $x_i, \hat{x}_j \in \mathcal{X}$. By our earlier choice of loss function, we have $L = I$, where I is identity matrix. Moreover, $P_X = \lambda$ and let $\Lambda = \operatorname{diag}(\lambda)$ as the diagonal matrix with λ along its principal diagonal. This gives the Bayes risk

$$\begin{aligned} R^*_\lambda(P_{Y|X}) &= \min_{P_{\hat{X}|Y} \in \mathbb{R}^{m\times n}_{\mathsf{sto}}} \mathbb{P}(\hat{X} = X) \\ &= \min_{P_{\hat{X}|Y} \in \mathbb{R}^{m\times n}_{\mathsf{sto}}} \operatorname{tr}\Big(\Lambda P_{Y|X} P_{\hat{X}|Y}\Big) \\ &= \sum_{y\in\mathcal{Y}} \min\{\lambda_1 P_{Y|X}(y|x_1), \ldots, \lambda_n P_{Y|X}(y|x_n)\}. \end{aligned} \quad (9)$$

Now define the *min-DeGroot distance* as $\tilde{\tau}_{\min}(\lambda, P_{Y|X}) \triangleq R^*_\lambda - R^*_\lambda(P_{Y|X})$, which gives

$$\begin{aligned} \tilde{\tau}_{\min}(\lambda, P_{Y|X}) &= \min\{\lambda_1, \ldots, \lambda_n\} \\ &\quad - \sum_{y\in\mathcal{Y}} \min\{\lambda_1 P_{Y|X}(y|x_1), \ldots, \lambda_n P_{Y|X}(y|x_n)\}. \end{aligned} \quad (10)$$

Note that $\tilde{\tau}(P_{Y|X}) \geq 0$ is a measure of the statistical information obtained about X after observing Y.

2) Max-DeGroot Distance: Consider the same setting as Section III-A1, where we are given a pair of random variables (X, Y) distributed as in (5). This time, fix the loss function as $l(X, \hat{X}) = \mathbb{1}_{\hat{X} \neq X}$. Then, the Bayes risk R_λ^* in estimating X when we do not observe Y is

$$R_\lambda^* = 1 - \max\{\lambda_1, \ldots, \lambda_n\}. \tag{11}$$

Now when Y is observed, since $L = \mathbf{1}\mathbf{1}^{\mathrm{T}} - I$ and $\mathrm{diag}(P_X) = \Lambda$ in the context of (8), the Bayes risk $R_\lambda^*(P_{Y|X})$ is given by

$$\begin{aligned} R_\lambda^*(P_{Y|X}) &= \min_{P_{\hat{X}|Y} \in \mathbb{R}_{\mathsf{sto}}^{m \times n}} \mathsf{Tr}\Big((\mathbf{1}\mathbf{1}^T - I)\Lambda P_{Y|X} P_{\hat{X}|Y}\Big) \\ &= \mathsf{Tr}(\Lambda P_{Y|X}\mathbf{1}\mathbf{1}^T) - \max_{P_{\hat{X}|Y} \in \mathbb{R}_{\mathsf{sto}}^{m \times n}} \mathsf{Tr}(\Lambda P_{Y|X} P_{\hat{X}|Y}) \\ &= 1 - \sum_{y \in \mathcal{Y}} \max\{\lambda_1 P_{Y|X}(y|x_1), \ldots, \lambda_n P_{Y|X}(y|x_n)\}. \end{aligned} \tag{12}$$

Hence, we define the quantity $\tilde{\tau}_{\max}(\lambda, P_{Y|X}) \triangleq R_\lambda^* - R_\lambda^*(P_{Y|X})$ as the *max-DeGroot distance*:

$$\begin{aligned} \tilde{\tau}_{\max}(\lambda, P_{Y|X}) = &\sum_{y \in \mathcal{Y}} \max\{\lambda_1 P_{Y|X}(y|x_1), \ldots, \lambda_n P_{Y|X}(y|x_n)\} \\ &- \max\{\lambda_1, \ldots, \lambda_n\}. \end{aligned} \tag{13}$$

Observe that for the special case where $n = 2$, both min-DeGroot and max-DeGroot distances are equivalent and the same as classical DeGroot distance [1]. Indeed, we have

$$\begin{aligned} &\min\{\lambda, \bar{\lambda}\} - \sum_{y \in \mathcal{Y}} \min\{\lambda P_{Y|X}(y|x_1), \bar{\lambda} P_{Y|X}(y|x_2)\} \\ &= \sum_{y \in \mathcal{Y}} \max\{\lambda P_{Y|X}(y|x_1), \bar{\lambda} P_{Y|X}(y|x_2)\} - \max\{\lambda, \bar{\lambda}\} \\ &= \big\|\lambda P_{Y|X=x_1} - \bar{\lambda} P_{Y|X=x_2}\big\|_1 - \frac{1}{2}|1 - 2\lambda|, \end{aligned} \tag{14}$$

where $\bar{\lambda} = 1 - \lambda$ for $\lambda \in [0, 1]$.

B. Simultaneously Maximal Coupling

In this section, we will generalize Goldstein's simultaneous coupling result, cf. [2], [27], to the case where we have n distributions rather than just two. Given a finite collection of probability distributions $P_{X_1,Y_1}, P_{X_2,Y_2}, \ldots, P_{X_n,Y_n}$, the next theorem constructs a joint coupling which is simultaneously maximal with respect to the joint distributions of (X_i, Y_i) for $i \in [n]$ and the marginal distributions of $X_1, \ldots, X_n$.

Theorem 3 (Simultaneously Maximal Coupling): Given n probability distributions $P_{X_1,Y_1}, P_{X_2,Y_2}, \ldots, P_{X_n,Y_n}$, over finite alphabets with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$ for $i \in [n]$, there exists a coupling $P_{X_1,Y_1,X_2,Y_2,\ldots,X_n,Y_n}$ that is simultaneously maximal in the following sense

$$\begin{aligned} \tau(P_{X_1,Y_1}, \ldots, P_{X_n,Y_n}) &= \mathbb{P}(X_1 = \cdots = X_n, Y_1 = \cdots = Y_n), \\ \tau(P_{X_1}, \ldots, P_{X_n}) &= \mathbb{P}(X_1 = \cdots = X_n), \end{aligned}$$

where $\mathbb{P}(\cdot)$ denotes the probability law under the coupling.
The proof is provided in [24]. We will use Theorem 3 to establish the information contraction properties of Doeblin coefficients over Bayesian networks.

IV. Contraction over Bayesian Networks

In this section, building on the ideas in [2], we derive the contraction properties of Doeblin coefficients over Bayesian networks. Specifically, given Doeblin coefficients for each of the channels comprising a Bayesian network, we obtain a bound on the Doeblin coefficient of the composite channel from a single source to any sink of nodes. To this end, consider a Bayesian network represented by a finite directed acyclic graph with vertex set $\mathcal{V}$. Every vertex U in $\mathcal{V}$ is a random variable that takes values in a finite alphabet set. Let the source node be denoted by X, and suppose each vertex U (other than the source node) is associated with the conditional distribution $P_{U|\mathrm{pa}(U)}$, where $\mathrm{pa}(U)$ denotes the set of parents of U. These conditional distributions collectively define the Bayesian network [28]. We assume that the vertices are topologically sorted, and for any node U and *subset* of nodes $V \subseteq \mathcal{V}$, we write $U > V$ when there is no directed path from U to V. Moreover, in the spirit of [2], we define a *site percolation* process on $\mathcal{V}$ so that each vertex U in $\mathcal{V}$ is removed independently with probability τ_U, where $\tau_U = \tau(P_{U|\mathrm{pa}(U)})$ denotes the Doeblin coefficient at U. Finally, let $\mathrm{perc}(V)$ be the probability (with repsect to the percolation process) that there is an (open) path from X to a subset of nodes $V \subseteq \mathcal{V}$.

Theorem 4 (Doeblin Coefficients in Bayesian Networks): For any node U in $\mathcal{V}$ and any subset of nodes $V \subseteq \mathcal{V}$ such that $U > V$, we have

$$\tau\big(P_{V \cup \{U\}|X}\big) \geq \tau_U \tau\big(P_{V|X}\big) + (1 - \tau_U)\tau(P_{V \cup \mathrm{pa}(U)|X}).$$

Furthermore, under the aforementioned site percolation model, we have for every $V \subseteq \mathcal{V}$,

$$\gamma\big(P_{V|X}\big) = 1 - \tau\big(P_{V|X}\big) \leq \mathrm{perc}(V).$$

The proof is provided in [24]. Note that $\tau(\cdot)$ is tending towards one over the network, which is intuitive since a stochastic matrix typically transforms input probability distributions into "more uniform" distributions. Therefore, a Doeblin coefficient closer to one implies that the corresponding channel is *losing information* contained in source distributions. Moreover, the percolation bound in Theorem 4 is a strengthening of that in [2] as $1 - \tau(P_{V|X}) \geq \eta_{\mathsf{TV}}(P_{V|X})$.

A. Samorodnitsky's SDPI for Doeblin Coefficients

We next establish a generalization of Samorodnitsky's SDPI for Doeblin coefficients in analogy with the results in [2], [4]. To this end, consider discrete random variables $U, X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ on finite alphabets. Suppose $U \to X^n \to Y^n$ forms a Markov chain for a channel $P_{X^n|U}$ and a *memoryless* channel $P_{Y^n|X^n}$, which is a tensor product of the channels $P_{Y_j|X_j}$ for $j \in [n]$:

$$P_{Y^n|X^n} = \prod_{j=1}^n P_{Y_j|X_j}. \tag{15}$$

Define the Doeblin coefficient of each individual channel as $\tau_j = \tau(P_{Y_j|X_j})$ for $j \in [n]$. Using the tensorization and

sub-multiplicativity properties in Theorem 1, we obtain the following bound on the Doeblin coefficient of $P_{Y^n|U}$:

$$\tau(P_{Y^n|U}) \geq \tau(P_{Y^n|X^n}) + \tau(P_{X^n|U})(1 - \tau(P_{Y^n|X^n})) \quad (16)$$

$$\geq \prod_{j=1}^{n} \tau_j. \quad (17)$$

This bound parallels the bound on contraction coefficients for operator convex f-divergences in [4, Equation 61]. In [2, Section 6.2] and [4, Section 3.4], the authors argue that stronger bounds can be obtained using more refined knowledge of the distribution P_{U,X^n}. In this vein, the ensuing theorem presents tighter bounds on $\tau(P_{Y^n|U})$ in terms of the single-letter Doeblin coefficients $\{\tau_j : j \in [n]\}$.

Theorem 5 (Samorodnitsky's SDPI for Doeblin Coefficients): For the Markov chain $U \to X^n \to Y^n$ with a given memoryless channel $P_{Y^n|X^n} = \prod_{j=1}^{n} P_{Y_j|X_j}$, we have

$$\tau(P_{Y^n|U}) \geq \sum_{T \subset [n]} P(T)\tau(P_{X_T|U}),$$

where $P(T)$ is the probability of a subset T of $[n]$ generated by independently drawing each element $i \in [n]$ with probability $1 - \tau_i$, and $X_T = \{X_i : i \in T\}$ for any subset $T \subseteq [n]$. Specifically, when $\tau_j = \tau$ for all $j \in [n]$, the following bound holds:

$$\tau(P_{Y^n|U}) \geq \sum_{T \subseteq [n]} \tau^{n-|T|}(1-\tau)^{|T|}\tau(P_{X_T|U}).$$

The proof is provided in [24].

V. Application to PMF Fusion

In this section, we will apply the techniques developed above to the general problem of aggregation or *fusion* of PMFs [29]. Pooling or fusion of PMFs is often used in statistics and machine learning to combine multiple sources of information or data into a single, more accurate estimate of a probability distribution. This can be useful in many applications, such as in resource-constrained Bayesian inference, where multiple sources of data or prior beliefs need to be combined to produce a single posterior distribution [30]. We next propose a new PMF fusion framework using an optimization-based approach.

A. PMF Fusion: Optimization-based Approach

Suppose we have n agents, each with their own PMFs $P_1, \ldots, P_n \in \mathcal{P}_m$, representing their beliefs about some phenomenon of interest. The goal is to generate a combined belief by aggregating their individual PMFs. One approach to achieving this is to use an optimization method to find the aggregated PMF that maximizes or minimizes some measure of discrepancy between the individual PMFs. This approach has been studied using various measures, such as weighted KL divergence and weighted α-divergences (see [29] for details).

Inspired by Doeblin coefficients, we propose a new approach that focuses on maximizing the probability of agreement among the agents under all posssible coupling of their beliefs. More precisely, we find the joint distribution or coupling that maximizes the likelihood that the agents' beliefs, represented as random variables $X_1, \ldots, X_n$, agree with each other over all possible couplings of their beliefs, i.e.,

$$\max_{\substack{P_{X_1,\ldots,X_n}: \\ P_{X_1}=P_1,\ldots,P_{X_n}=P_n}} \mathbb{P}(X_1 = \cdots = X_n), \quad (18)$$

where $\mathbb{P}(\cdot)$ denotes the probability law corresponding to a coupling. Hence, we define the optimal fused PMF as the conditional PMF under the coupling conditioned on the agreement among the agents i.e., $P_* \in \mathcal{P}_m$ via $P_*(x) \propto \mathbb{P}(X_1 = \cdots = X_n = x)$ for the maximal coupling. Since Theorem 2 tells us that $\mathbb{P}(X_1 = \cdots = X_n = x) = \min\{P_1(x), \ldots, P_n(x)\}$ for the maximal coupling, we can write the fused PMF as:

$$\forall x \in \mathcal{Y}, \;\; P_*(x) = \frac{\min\{P_1(x), \ldots, P_n(x)\}}{\sum_{x \in \mathcal{X}} \min\{P_1(x), \ldots, P_n(x)\}}. \quad (19)$$

We refer to the above fusion method as the *min-rule*. Such a min-rule had been proposed in the literature, cf. [31], but lacked a formal probabilistic understanding. Indeed, the min-rule was used for distributed hypothesis testing in [31] based on the intuition that if there is a true state of the world $x^* \in \mathcal{Y}$, and if there is an agent that can distinguish x from x^* for every false state $x \in \mathcal{Y}\backslash\{x^*\}$, then the min-rule can drive the beliefs of distributed agents to zero on each of the false states. While [31] established that the min-rule enjoys faster asymptotic convergence than other belief averaging-based approaches, such as arithmetic pooling [32], [33] and geometric pooling [34]–[37], our development reveals the underlying probabilistic interpretation of the min-rule.

It is important to note that the min-rule in (19) is only suitable under certain settings, such as group decision problems [38], where different experts are given equal importance and each expert is trying to construct optimal estimates based on their knowledge. Clearly, if some of the agents are either adversarial or have imperfect knowledge about the environment, then the min-rule may not be appropriate.

VI. Conclusion

In this work, we have developed new insights into the structure and geometry of Doeblin coefficients. For instance, we have illustrated using extremal coupling characterizations how Doeblin coefficients generalize TV distance to measure "distances" between multiple probability distributions. We have also demonstrated the contraction properties of Doeblin coefficients over Bayesian networks, similar to other established contraction coefficients, such as those for KL divergence and TV distance. Furthermore, we have introduced a new approach for fusing PMFs based on Doeblin coefficients dubbed the *min-rule*, which serves as a means of aggregating the beliefs of agents in certain applications. Finally, some immediate directions for future research include identifying the axioms that uniquely characterize the min-rule and developing robust versions of it that take into account the reliability of agents.

REFERENCES

[1] 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.

[2] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and bayesian networks," in *Convexity and Concentration*. New York, NY: Springer New York, 2017, pp. 211–249.

[3] 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.

[4] 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.

[5] A. Makur, "Information contraction and decomposition," Sc.D. Thesis in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, May 2019.

[6] A. Xu and M. Raginsky, "Converses for distributed estimation via strong data processing inequalities," *IEEE International Symposium on Information Theory - Proceedings*, vol. 2015-June, pp. 2376–2380, 9 2015.

[7] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pp. 429–438, 2013.

[8] S. Asoodeh, M. Aliakbarpour, and F. P. Calmon, "Local differential privacy is equivalent to contraction of an f-divergence," *IEEE International Symposium on Information Theory - Proceedings*, vol. 2021-July, pp. 545–550, 7 2021.

[9] J. Liu, P. Cuff, and S. Verdu, "Secret key generation with one communicator and a one-shot converse via hypercontractivity," *IEEE International Symposium on Information Theory - Proceedings*, vol. 2015-June, pp. 710–714, 9 2015.

[10] J. Körner and K. Marton, "Comparison of two noisy channels," in *Topics in Information Theory*, I. Csiszr and P. Elias, Eds., Amsterdam: North-Holland, 1977, pp. 411–423.

[11] P. P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Transactions on Information Theory*, vol. 19, pp. 197–207, 1973.

[12] T. M. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. IT-18, no. 1, pp. 2–14, January 1972.

[13] S. Chestnut and M. E. Lladser, "Occupancy distributions in markov chains via doeblin's ergodicity coefficient," *Discrete Mathematics & Theoretical Computer Science*, pp. 79–92, 2010.

[14] 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.

[15] 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.

[16] Wolfgang 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.

[17] 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., vol. 19. Amsterdam, Netherlands: North-Holland, Elsevier, 2001, pp. 145–170.

[18] A. Makur, "Coding theorems for noisy permutation channels," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 6723–6748, November 2020.

[19] K. B. Athreya and P. Ney, "A new approach to the limit theory of recurrent Markov chains," *Transactions of the American Mathematical Society*, vol. 245, pp. 493–501, November 1978.

[20] E. Nummelin, "A splitting technique for Harris recurrent Markov chains," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 43, no. 4, pp. 309–318, December 1978.

[21] J. E. Cohen, J. H. B. Kemperman, and G. Zbăganu, *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population Sciences*. Ann Arbor, MI, USA: Birkhäuser, 1998.

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

[23] R. C. Williamson and Z. Cranko, "Information processing equalities and the information-risk bridge," 7 2022. [Online]. Available: https://arxiv.org/abs/2207.11987v1

[24] A. Makur and J. Singh, "Doeblin coefficients and related quantities," 2023, in preparation.

[25] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, 1st ed. Providence, RI, USA: American Mathematical Society, 2009.

[26] M. H. DeGroot, "Uncertainty, information, and sequential experiments," *https://doi.org/10.1214/aoms/1177704567*, vol. 33, pp. 404–419, 6 1962.

[27] S. Goldstein, "Maximal coupling," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 46, pp. 193–204, 1979.

[28] M. Kevin, "A brief introduction to graphical models and bayesian networks," *A Tutorial on Bayesian Networks*, 1998.

[29] G. Koliander, Y. El-Laham, P. M. Djuric, and F. Hlawatsch, "Fusion of probability density functions," *Proceedings of the IEEE*, vol. 110, pp. 404–453, 4 2022.

[30] C. Genest and J. V. Zidek, "Combining Probability Distributions: A Critique and an Annotated Bibliography," *Statistical Science*, vol. 1, no. 1, pp. 114 – 135, 1986. [Online]. Available: https://doi.org/10.1214/ss/1177013825

[31] A. Mitra, J. A. Richards, and S. Sundaram, "A new approach to distributed hypothesis testing and non-bayesian learning: Improved learning rate and byzantine resilience," *IEEE Transactions on Automatic Control*, vol. 66, pp. 4084–4100, 10 2020.

[32] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-bayesian social learning," *Games and Economic Behavior*, vol. 76, pp. 210–225, 9 2012.

[33] A. Jadbabaie, P. Molavi, and A. Tahbaz-Salehi, "Information heterogeneity and the speed of learning in social networks," *SSRN Electronic Journal*, 5 2013. [Online]. Available: https://papers.ssrn.com/abstract=2266979

[34] A. Lalitha, T. Javidi, and A. D. Sarwate, "Social learning and distributed hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, pp. 6161–6179, 9 2018.

[35] A. Nedic, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-bayesian learning," *IEEE Transactions on Automatic Control*, vol. 62, pp. 5538–5553, 11 2017.

[36] Y. Inan, M. Kayaalp, E. Telatar, and A. H. Sayed, "Social learning under randomized collaborations," vol. 2022-June. Institute of Electrical and Electronics Engineers Inc., 1 2022, pp. 115–120. [Online]. Available: https://arxiv.org/abs/2201.10957v2

[37] V. Bordignon, V. Matta, and A. H. Sayed, "Adaptive social learning," *IEEE Transactions on Information Theory*, vol. 67, pp. 6053–6081, 9 2021.

[38] S. French, "Aggregating expert judgement," *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales - Serie A: Matematicas*, vol. 105, pp. 181–206, Feb. 2011.