A distance function for stochastic matrices

Antony R. Lee* and Peter Tiňo
School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT

Iain B. Styles
School of Electronics, Electrical Engineering, and Computer Science, Queen's University Belfast, University Road, Belfast, BT7 1NN

(Dated: May 13, 2026)

The ability to quantify the similarity of stochastic processes is important in a wide range of areas including communications, simulation, and operational research. Such processes are commonly modelled as Markov Chains and so a natural way to compare their similarity is to compute distances between pairs of Markov Chains. We propose a novel distance function on the space of stochastic matrices that draws on ideas from information geometry. We first show that the Bhattacharyya angle is an appropriate measure of distance between Markov chain sequences, drawing on examples taken from healthcare processes and deriving bounds on the convergence of the distance and mixing times. We then extend these ideas to derive a novel distance measure which enables direct comparison of the transition matrices themselves. Our result is a true metric which has a closed form and is efficient to implement for numerical evaluation. In the case of ergodic Markov chains, it is shown that considering either the Bhattacharyya angle on Markov sequences or the new stochastic matrix distance leads to the same distance between models.

I. INTRODUCTION

Markov chains are a well studied tool in statistical physics with numerous applications in communications [1, 2], simulation [3], queuing [4], and many other areas. Of particular interest to us is the application of techniques originating in physical systems that use Markov chain models to describe operational processes [5, 6] and other complex phenomena in the healthcare domain including disease transmission [7–10], general healthcare services [11–13], mental health [14], community care [15], and cost-effectiveness of interventions [16]. These model can influence both healthare outcomes and inform decisions made by both healthcare professionals and policymakers. In healthcare common questions about operational processes could include: What are the typical sequence of events when someone enters a hospital in an emergency? Can we redesign an operational process to be more efficient? How can we compare operational processes to see which are similar or different? In particular, the UK's National Institute for Health Excellence (NICE) regularly publishes guidance and recommendations based on estimating transition probabilities which model healthcare processes as a Markov chain [17], with an explicit example being the use of clinical test results to define a transition matrix with respect to the role of Peginterferon beta-1a for treating relapsing–remitting multiple sclerosis [18]. Although applicable to any physical system described by a discrete Markov chain, it is this focus on comparing descriptions of health services which motivates our work here [19–22].

Within the domain of discrete time Markov chains, much work has been done on quantifying the similarity between two chains. The most widely used approach to comparing Markov chains is to consider the probabilities induced on the sequence space they generate, i.e comparing fixed length sequences of two models to distinguish different behaviour. Work in this area ranges from using the Wasserstein distances [23], vectorisation of (hidden) Markov chain transitions compared with simple functions (e.g. dot products, cosine angle and the Frobenius distance) [24]. Alternatively, a general approach to L^p norm–induced distances and the Kullback-Leibler divergence has also been studied [25], extending them to infinite time runs while exploring their properties of continuity. With for some specific cases the ability to estimate the distance statistically also studied [26].

Although they have proven useful, the L^p induced distances considered do not have a natural probabilistic interpretation, and similarities based on the Kullback-Leibler divergence might be infinite or very sensitive to small changes to a stochastic matrix. More formally Kullback-Leibler–type divergences are not true distances because they sacrifice properties such as symmetry (quasimetric), identity of indiscernibles (pseudometric) or the triangle inequality (semimetric). Specifically, the KL divergence $D_{\mathrm{KL}}(p\|q) = \sum_i p_i \log(p_i/q_i)$ diverges to infinity whenever $q_i \to 0$ while $p_i > 0$, making it undefined or infinite for distributions with differing support. This motivates the identification of a distance function for comparing the trace spaces of Markov chains which: 1) is a true distance, 2) is efficient to evaluate, 3) incorporates the probabilistic nature of traces. To address these issues, we propose the use of the well known distance function, the Bhattacharyya angle, motivated and derived from first principles via information geometry by associating a Markov chain with a categorical random variable [27, 28]. In contrast, the Bhattacharyya coefficient $\sum_i \sqrt{p_i q_i}$ and its derived angle remain finite and

* arl290@student.bham.ac.uk

continuous even as probability mass approaches zero.

Although a step in a good direction, as previously noted some similarity functions used for Markov sequences can in some scenarios still be overly sensitive to initial distributions or small changes in the stochastic matrix which represents a Markov chain. For example a simple model of a biased coin toss, where outcome bias is parameterised by some value $p \in [0, 1]$, can generate completely difference trace sequences if comparing two different initial distributions. This is obvious: given two different starting conditions we end up at different places. However the underlying generative process, as specified by the stochastic matrix, is the same. It could be that in a given scenario, considering starting different chains with different initial distributions might give the quickest way to distinguish behaviour. However, if those distributions are not already known or computable then the choice of one initial distribution over another might be difficult to justify.

In healthcare, another example of where the initial state is of lesser importance is in hospital emergency department patient attendances or other specific disease pathways. As the process typically starts in the same initial state, e.g. "presented at hospital", every comparison of processes across different organisations will use the same initial state distribution for a Markov chain, but the actual patient pathways will depend on the stochastic matrix representing the process.

It therefore seems reasonable to separate the initial conditions from the Markov chain's structural properties. This amounts to equipping the space of stochastic matrices with a distance function. This suggests an approach to comparing Markov chains via the structure of the stochastic matrices themselves, rather than traces generated by short or long term runs.

The Markov theory literature for direct comparisons of stochastic matrices seems less formalised than its sequence space counterpart. Early work on comparing the structural influence of stochastic matrices can be found in the field of channel capacities [29]. In more recent work, Markov mixture models were created in [30] which use the log-likelihood function and BIC criteria to cluster a person's sequence of daily activity via estimated stochastic matrices. The investigation of the internal block structure of a special class of Markov chains in [31] used L^p norms and the Kullback-Leibler divergence. They also appealed to Markov chain mixing times to show estimated values converge in a short enough time to be accurate. In the context of analysing protein structure models, [32] propose a similarity score between the τ step stochastic matrices of two different proteins. Mixing long term and short term properties of a Markov chain, [33] propose modelling recommendations for a user's next choice given previous choices as an order k Markov chain.

Learning a stochastic matrix, given observation data, is also important in the fields of sensor-based human activity recognition [34] and designing statistically consistent algorithms for noisy data labelling [35]. Underlying these estimation techniques is the choice of objective function one might want to optimise. Typical objective functions used are: the direct calculation of the Kullback-Leibler divergence [36, 37] which is the usual divergence calculated row-wise and summed for a pair of stochastic matrices, some form of log-likelihood function as proposed for denoising in image analysis [38], the related cross entropy for video analysis [39], cross entropy hypothesis testing an estimated stochastic matrix against a reference [40] or even Bayesian based test approaches [41, 42]. In the majority of these settings short term runs are more important as recommendations for the next action are all that is needed, or the long time run of a Markov chain gives no more benefit after a finite number of steps, and the information contained within the stochastic matrix is all that is needed. Our observation is that a well motivated comparison between Markov chain stochastic matrices would be beneficial, and could make more sense in certain scenarios than comparing sequences directly. This leads us to identify a distance function between stochastic matrices which mirrors desired properties of distances on sequence spaces.

In summary, our ideas are to further develop the use of information geometry in Markov chain theory as a principled way to explore a true distance measure in two scenarios:

1. Sequence space: Comparing the probability distributions induced by two Markov chains in terms of their τ step sequences
2. Stochastic matrix space: Comparing the structural, i.e. local state to state jump, difference between two Markov chains, independently of initial distributions

To address these ideas, our contributions in this work are:

1. Result 1: Motivation from first principles of the use of the Bhattacharyya angle as an appropriate distance measure for comparing Markov chain sequences.
2. Result 2: Derive a long term rate comparison of Markov chain sequences.
3. Result 3: Extending these principles to formulate a novel distance measure which compares Markov chain stochastic matrices directly.
4. Result 4: Derive a long term rate comparison of stochastic matrices.

Throughout, we denote a time homogeneous discrete time Markov chain with M which is defined over a state space $X = \{x_1, \ldots, x_k\}$. In other words we consider sequences of random variables $(Y_1, Y_2, \ldots)$ which satisfy the standard Markov property. Initial Markov chain distributions are written $\boldsymbol{\pi}$, with components $\pi(x_i)$ representing the probability of being in state x_j, and the associated model's (row) stochastic matrix $\boldsymbol{P}$, with components $P(x_i, x_j) = \Pr(Y_{t+1} = x_j | Y_t = x_i)$. Sequences of

length τ form a space X^τ and the space of infinite length sequences is written $X^{(\omega)}$. Over all elements of either X^τ, or alternatively $X^{(\omega)}$, we can identify an induced probability vector $\boldsymbol{p}$, with components denoted $p(w)$, assigned for each sequence $w = (x_1, \ldots, x_\tau) \in X^\tau$, according to initial distribution $\boldsymbol{\pi}$ and stochastic matrix $\boldsymbol{P}$. The Iverson bracket is written as $[A]$ for some boolean statement A, and the collection of eigenvalues of a matrix $\boldsymbol{R}$ is denoted $\sigma(\boldsymbol{R})$. Finally we make use of the Hadamard product (i.e. element wise product) between two elements, denoted as $\circ$, and the Hadamard square root (i.e. element wise square root) denoted as $\sqrt{\cdot}$, unless stated otherwise.

Remark 1. *Throughout, we use the superscript notation $d^{(\tau)}$ and $d^{(\omega)}$ for distances on sequence spaces (finite and infinite respectively), while subscript notation d_Ω denotes distances on the space of stochastic matrices.*

II. SEQUENCE DISTANCES

Information geometry is the study of parametric families of probability distributions using the language of differential geometry [28, 43]. An extremely well developed and diverse field, we will only present the topics we need and briefly. Other than our stated results, all other concepts can be taken as standard literature.

We treat the parameter space of a given statistical model as a manifold. The fundamental object in both information and differential geometry is the metric tensor. Manifolds allow us to discuss concepts such as continuity and differentiation. A metric tensor allows us to define a notion of distance between points on a manifold, which means we can equip the parameter space of a statistical model with a distance function.

There are many ways to define a metric tensor for a manifold and historically in information geometry this has been via the Fisher information metric tensor. Consider a parametric family of stochastic models whose distribution function is denoted $h(x; \boldsymbol{p})$. Here x is an element of the distribution's support space X and $\boldsymbol{p} \in \mathbb{R}^{|X|}$ is a vector of the parameters for the family, along which a standard result is it can be treated as a probability itself. Up to a constant scalar, the Fisher information metric at a point $\boldsymbol{p}$ is defined as an $|X| \times |X|$ matrix with components,

$$g_{ij}(\boldsymbol{p}) = -\mathbb{E}_h \left[\frac{\partial^2 \log h(x; \boldsymbol{p})}{\partial p_i \partial p_j} \right] \tag{1}$$

where $\mathbb{E}_h$ indicates the expectation value is taken with respect to the probability distribution $h(x; \boldsymbol{p})$ over the support of x. The g_{ij} are known as metric tensor components, and define the geometry of the parameter space, which is a statistical manifold. We want to discuss the concept of geodesic curves, i.e. curves connecting two points of the parameter space which respect the underlying geometry of the space. Geodesics are derived using the geodesic equation which in turn is derived from the metric tensor. Intuitively, the solutions to the geodesic equations are the "shortest" path curves with respect to the geometry of the space under consideration. Conventionally, geodesic curves are parameterised by some real variable s, which encodes the start point and end point of the curve. Ultimately we want to find a closed form for the induced distance function, minimised over all possible geodesic curves $\boldsymbol{c}(s)$. In the case of a parametric family of probability distributions these points are denoted $\boldsymbol{c}(s_1) = \boldsymbol{p}_1$ and $\boldsymbol{c}(s_2) = \boldsymbol{p}_2$. In general, there can be multiple geodesic curves connecting two points. For example the two ways round a great circle on a sphere. Thus it is standard to define the distance between two points using the geodesic curve which has the smallest length,

$$\delta(\boldsymbol{p}_1, \boldsymbol{p}_2) = \min_{\boldsymbol{c}(s)} \int_{s_1}^{s_2} \mathrm{d}s \sqrt{\sum_{i,j} g_{ij}(\boldsymbol{c}(s)) \frac{\mathrm{d}c_i(s)}{\mathrm{d}s} \frac{\mathrm{d}c_j(s)}{\mathrm{d}s}} \tag{2}$$

$\delta(\boldsymbol{p}_1, \boldsymbol{p}_2)$ would then be our desired distance function on a statistical manifold. It is well known that Markov chains induce a probability distribution over sequences of fixed length [44]. These distributions correspond to categorical distributions and are defined as having the probability mass function,

$$h(x; \boldsymbol{p}) = \prod_{i=1}^{|X|} p(x_i)^{[x = x_i]} \tag{3}$$

with domain of input events $X = \{x_1, \ldots, x_{|X|}\}$ and associated probabilities $\boldsymbol{p} = (p(x_1), \ldots, p(x_{|X|}))$, with additional notational shorthand $p_i = p(x_i)$ used when considering points on a manifold. In other words the probability of observing each possible event x_i is $p(x_i)$, akin to the probability distribution of rolling an $|X|$-sided dice where each side x_i has probability $p(x_i)$. For a fixed state space of $|X|$ elements, the space of all probabilities which can parametrise a categorical distribution is known as the simplex space $\Delta = \{\boldsymbol{p} : \|\boldsymbol{p}\|_1 = 1\}$. Thus from here we can compute the metric tensor components in Eq. 1 and work through until we arrive at a solution for Eq. 2.

It is well known the metric tensor components we require are, up to an overall positive scaling constant,

$$g_{ij}(\boldsymbol{p}) = \frac{\delta_{ij}}{p_i} \tag{4}$$

where δ_{ij} is a Kronecker delta. In differential geometry a metric tensor can be used to define geodesic equations, i.e. equations which define in some sense the "straightest" path. Typically solving these equations is difficult, however in this case we can sidestep the usual challenge. The coordinate transformation $\boldsymbol{y} = \sqrt{\boldsymbol{p}}$ allows us to show the metric tensor components become $g_{ij}(\boldsymbol{y}) = \delta_{ij}$. This transformation shows the components are the standard Euclidean metric in Euclidean space, with the condition that points are constrained to a positive quadrant of unit

hypersphere centered at the origin ($\boldsymbol{y} \cdot \boldsymbol{y} = 1$). Hence, the associated geodesic curves we want are arcs following the great circles defined by points $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. As a result, a closed form for the distance between two categorical distributions with $|X|$ states and parameterised with $\boldsymbol{p}_1, \boldsymbol{p}_2 \in \Delta$ is,

$$\delta^{(|X|)}(\boldsymbol{p}_1, \boldsymbol{p}_2) = 2 \arccos \mathrm{BC}(\boldsymbol{p}_1, \boldsymbol{p}_2) \tag{5}$$

This distance is known under various names such as the Rao distance [45], categorical Fisher-Rao distance [46], but we will refer to it as the Bhattacharyya angle [47]. Here we have made clear the link of this distance to the Bhattacharyya coefficient,

$$\mathrm{BC}(\boldsymbol{p}_1, \boldsymbol{p}_2) = \sum_{i=1}^{|X|} \sqrt{p_1(x_i) p_2(x_i)} \tag{6}$$

The Bhattacharyya coefficient has been studied in relation to comparing discrete probability distributions already [48, 49]. However the Bhattacharyya angle's use as a true distance measure does not yet seem to have been pursued for the purpose of comparing Markov chains.

The final link back to Markov chains comes from realising that the collection of all probabilities of observing a τ step Markov chain sequence forms a categorical distribution over X^τ.

Result 1. *Consider two Markov chains M_1 and M_2 over the same state space X, with initial distribution π_k and transition probabilities $P_k(x_i, x_j)$, collected in stochastic matrices $\boldsymbol{P}_k$, $k = 1, 2$. The probability of observing a sequence of events $w = (x_1, x_2, \ldots, x_\tau)$ $\tau \geq 2$, under the Markov chain M_k is*

$$p_k(w) = \pi_k(x_1) P_k(x_1, x_2) \cdots P_k(x_{\tau-1}, x_\tau) \tag{7}$$

Fixing the sequence length to $\tau \geq 2$, the probabilities assigned to all length-τ sequences $w \in X^\tau$ by the two MArkov chains are collected (assuming an order on X^τ) in categorical probability vectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$. We thus arrive at a distance for Markov chains induced by the sequence length τ, $\delta^{(|X|^\tau)}(\boldsymbol{p}_1, \boldsymbol{p}_2)$, which we will denote by $d^{(\tau)}(M_1, M_2)$:

$$d^{(\tau)}(M_1, M_2) = 2 \arccos \sum_{w \in X^\tau} \sqrt{p_1(w) p_2(w)} \tag{8}$$

A realisation of a Markov chain is written as a sequence of events $w = (x_1, x_2, \ldots, x_\tau)$ for a chain M and $\tau \geq 2$. The probability of observing the sequence w given an initial distribution $\boldsymbol{\pi}$ for a Markov chain with stochastic matrix $\boldsymbol{P}$ is

$$p(w) = \pi(x_1) P(x_1, x_2) \cdots P(x_{\tau-1}, x_\tau) \tag{9}$$

Making the identification $M_1 \to \boldsymbol{p}_1$, $M_2 \to \boldsymbol{p}_2$ we arrive at a distance for Markov chains,

$$d^{(\tau)}(M_1, M_2) = 2 \arccos \sum_{w \in X^\tau} \sqrt{p_1(w) p_2(w)} \tag{10}$$

The key concepts here are: 1) using information geometry, a notion of distance can be naturally motivated for Markov chains, 2) the distance satisfies all properties of a true distance function and 3) the distance is constructed to respect the basic probability properties of a Markov chain. Following [50], for convenience we introduce the following definitions,

$$\boldsymbol{r} = \sqrt{\boldsymbol{\pi}_1} \circ \sqrt{\boldsymbol{\pi}_2} \tag{11a}$$

$$\boldsymbol{R} = \sqrt{\boldsymbol{P}_1} \circ \sqrt{\boldsymbol{P}_2} \tag{11b}$$

Considering the product structure for each $p_k(w)$ and denoting the vector of all ones as $\mathbf{1}$, we can compactly write the sum over length τ words,

$$d^{(\tau)}(M_1, M_2) = 2 \arccos \left(\boldsymbol{r}^T \boldsymbol{R}^{\tau-1} \mathbf{1} \right) \tag{12}$$

A proof of the equivalence between Eq. 10 and Eq. 12 is in Apx. A. The Bhattacharyya angle in Eq. 12 serves as a true distance measure for comparing Markov chain induced probabilities. It is numerically very appealing as calculations can be reduced to matrix and vector multiplication.

Turning to long run sequences, we can compute asymptotic quantities akin to the Kullback-Leibler divergence rate. Such quantities are important in the analysis of systems which have become stationary, i.e. do not change any more, and are linked to the information content of a probability distribution. Typically they compare infinite step traces for two Markov chains, averaged by the number of steps. However we immediately see for any pair of models $d^{(\tau)}(M_1, M_2)/\tau \to 0$ as $\tau \to \infty$. This is similar to most "distances" considered in Markov theory, with the notable exception of the Kullback-Leibler divergence rate [51]. Hence we are motivated to study the non-regularised distance limit. In the case of the Bhattacharyya angle the quantity we are interested in is,

$$\begin{aligned} d^{(\omega)}(M_1, M_2) &= \lim_{\tau \to \infty} d^{(\tau)}(M_1, M_2) \\ &= 2 \arccos \sum_{w \in X^{(\omega)}} \sqrt{p_1(w) p_2(w)} \end{aligned} \tag{13}$$

Note this is equivalent to the matrix product form of $d^{(\omega)}(M_1, M_2) = \lim_{\tau \to \infty} 2 \arccos \left(\boldsymbol{r}^T \boldsymbol{R}^{\tau-1} \mathbf{1} \right)$, using the derivation in Apx. A. We can see the Bhattacharyya angle always exists in the limit $\tau \to \infty$ by observing (via the inequality of arithmetic and geometric means) $\sqrt{p_1(w) p_2(w)} \leq \frac{p_1(w) + p_2(w)}{2}$ and using the direct comparison test for series convergence [52]. As $\sum_{w \in X^{(\omega)}} \frac{p_1(w) + p_2(w)}{2} = 1$ (i.e. converges) the direct comparison test proves $d^{(\omega)}(M_1, M_2)$ converges. Convergence of the Bhattacharyya angle is then ensured by the continuity of arccos on $[0, 1]$.

A simple characterisation of when $d^{(\omega)}(M_1, M_2)$ is guaranteed to attain its maximum can be found from the inequality $\rho(\sqrt{\boldsymbol{P}_1} \circ \sqrt{\boldsymbol{P}_2}) \leq \rho(\boldsymbol{P}_1) \rho(\boldsymbol{P}_2) = 1$ (see Huang et. al. for details [53]). In other words when the spectral radius of $\sqrt{\boldsymbol{P}_1} \circ \sqrt{\boldsymbol{P}_2}$ is strictly less than unity the

distance converges to the numerical constant π in the limit $\tau \to \infty$. This means the long term run of two such Markov chains are statistically distinguishable according to the Bhattacharyya angle. However we can go further and obtain a closed form for the long term rate for any pair of stochastic matrices.

Result 2. *Let X be the state space for a Markov chain with $|X|$ states. Let $\boldsymbol{r} \in [0,1]^{|X|}$ be a sub-stochastic vector (row sums ≤ 1) and $\boldsymbol{R} \in [0,1]^{|X|\times|X|}$ be a sub-stochastic matrix (row sums ≤ 1). There exists a partitioning of X such that Q states form closed recurrent classes and T states are transient. Denote by $\mathbf{1}_Q$ the vector of ones of dimension Q, by $\boldsymbol{T}$ the sub-matrix of transition probabilities among transient states, and by $\boldsymbol{C}$ the sub-matrix of transition probabilities from transient states to recurrent states. When the spectral radius of $\boldsymbol{T}$ is strictly less than unity, the fundamental matrix $\boldsymbol{N} = (\boldsymbol{I} - \boldsymbol{T})^{-1}$ exists, and the absorption probability vector is $\boldsymbol{a} = \boldsymbol{N}\,\boldsymbol{C}\,\mathbf{1}_Q$. The Bhattacharyya rate then satisfies:*

$$d^{(\omega)}(M_1, M_2) = 2 \arccos\left[\boldsymbol{r}^\top \begin{pmatrix} \mathbf{1}_Q \\ \boldsymbol{a} \end{pmatrix}\right]. \tag{14}$$

The proof of Result 2 is contained in Apx. B. The closed form nature of the Bhattacharyya angle is such that numerical analysis can be efficiently implemented, which suggests we look at a position between short term runs and comparing Markov chains asymptotically through the field of mixing times [54, 55].

Markov chain mixing times are concerned with estimating a finite time step $\tau_\star$ such that the distance between some reference stationary distribution and the $\tau_\star$ step distribution of some Markov chain is "close". For a state space X with $|X|$ elements, denote an initial Markov chain distribution for a given state $x_j \in X$ as $\boldsymbol{e}_j$ (i.e. the standard basis in $\mathbb{R}^{|X|}$), we define the Bhattacharyya angle induced mixing time of a Markov chain $\boldsymbol{P}$ (with dimensions $|X| \times |X|$) starting in state $\boldsymbol{e}_j$, with unique stationary distribution $\bar{\boldsymbol{\pi}}_\star$, as

$$\tau_\star(x) = \min\left\{\tau > 0 : d^{(|X|)}\left(\bar{\boldsymbol{\pi}}_\star, \boldsymbol{e}_j^T \boldsymbol{P}^\tau\right) \leq \epsilon\right\} \tag{15}$$

We can operationally call two Markov chains "similar" if after some time $\tau_\star$ their state distributions are "close enough" for the problem at hand. Another way to view this is that if we have a obtained a mixing time for one chain, how can we use it to bound the mixing time of another chain? In some sense one can then approximate one Markov chain with another after a finite number of steps, when comparing the marginal state distributions of two Markov chains.

A well studied class of discrete Markov chains are both ergodic and and reversible. A well known result is that ergodic chains have a unique stationary distribution $\bar{\boldsymbol{\pi}}$. For an ergodic chain, reversible means the associated unique stationary distribution and stochastic matrix satisfy,

$$\operatorname{diag}(\bar{\boldsymbol{\pi}})\,\boldsymbol{P} = \boldsymbol{P}^T \operatorname{diag}(\bar{\boldsymbol{\pi}}) \tag{16}$$

which are known as the detailed balance equations. Under these assumptions, bounds on $\tau_\star$ can be placed as we can take advantage of the fact that the τ-step transition matrix components can be written as [55],

$$P^\tau(x_j, x_k) = \pi_\star(x_k) + \sqrt{\frac{\pi_\star(x_k)}{\pi_\star(x_j)}} \sum_{i=2}^{|X|} \lambda_i^\tau e_j^{(i)} e_k^{(i)} \tag{17}$$

Here λ_i are the eigenvalues of $\boldsymbol{P}$, the $\boldsymbol{e}^{(i)}$ are an orthonormal basis $\boldsymbol{e}^{(i)} \cdot \boldsymbol{e}^{(j)} = \delta_{ij}$, with $\boldsymbol{e}^{(1)} = \sqrt{\bar{\boldsymbol{\pi}}_\star}$. Finally the assumption $\bar{\boldsymbol{\pi}}_\star > \mathbf{0}$ for the unique stationary distribution of $\boldsymbol{P}$ is imposed. Usefully, the eigenvalues of ergodic and reversible Markov chains are real and satisfy $\lambda_1 = 1 > \lambda_2 \geq \ldots \geq \lambda_{|X|} > -1$. This allows us to take the limit $\tau \to \infty$ and know with certainty the power of all eigenvalues, except λ_1, will eventually vanish. With the expression in Eq. 17, and using the Bhattacharyya angle's Taylor series, with associated bounds on remainder terms, we can give two conditions needed to be sufficient to bound the mixing time,

$$\tau_- \geq \frac{1}{\log \lambda_2} \log\left[\gamma\left(1 + \frac{1}{\pi_-}\right)^{-1}\right] \tag{18a}$$

$$\tau_+ \geq \frac{1}{2\log \lambda_2} \log\left[4(1 - \cos\epsilon/2)\left(1 + \frac{1}{\pi_-}\right)^{-1}\right] \tag{18b}$$

with $\pi_- = \min \bar{\boldsymbol{\pi}}_\star$, and the constant $\gamma = 1 - (5/16)^{2/7} \approx 0.282$. The constant γ is chosen to ensure the remainder of the Taylor series expansion of the mixing time can be replaced safely with a known lower bound, which entails a minimum requirement on $\tau_\star$ given by τ_-. The second inequality using τ_+ gives a sufficient condition which guarantees the desired distance accuracy ϵ has been achieved. Consequently, the true mixing time is known to fall within the range $[\tau_-, \tau_+]$. Given the number of uses, and easily obtained results afforded by the Bhattacharyya angle we promote it as a robust tool to investigate both short and long time Markov chain runs.

III. STOCHASTIC MATRIX DISTANCES

Our idea is to, again appealing to information geometry, elevate the Bhattacharyya angle from a space of probability distributions to matrices. To this end, note that any stochastic matrix is a collection of individual state-conditional next-state probability distributions [56, 57]. This can be interpreted as a stochastic matrix being an element of a Cartesian product space of underlying categorical distributions (as explained in the trace distance section). Thus we identify $\boldsymbol{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{|X|}) \in \Delta \times \cdots \times \Delta$, where Δ is the $|X|$-dimensional simplex space, i.e. $\boldsymbol{P}$ is an $|X| \times |X|$ matrix. Such a decomposition induces what is known as a product metric in differential geometry [46]. In other words, we can build a distance function on the space

of stochastic matrices by using the Bhattacharyya angle inherited from the underlying simplex space Δ. Using a product metric structure here is actually very natural. The work of [56, 57] characterise the metric tensor components for stochastic matrices in terms of invariance under linear transformations called Markov maps. The result is choosing a product metric structure or asking for invariance under Markov maps gives the same distance on stochastic matrices.

As a demonstration using two manifolds, consider the product metric tensor induced via the Cartesian product $(\Delta, g) \times (\Delta, g)$ where g is the metric tensor defined via Eq. 4. The product metric tensor components in this case are defined as,

$$g_{ij}^{\text{prod}}(\boldsymbol{p}_1, \boldsymbol{p}_2) = g_{ij}(\boldsymbol{p}_1) \oplus g_{ij}(\boldsymbol{p}_2) \tag{19}$$

with $\oplus$ denoting the direct sum of the two independent manifolds. Further, we can construct the product curve which connects pairs of points on $\Delta \times \Delta$ as $\boldsymbol{c}(s) = (\boldsymbol{c}_1(s), \boldsymbol{c}_2(s))$ where $\boldsymbol{c}_1(s)$ and $\boldsymbol{c}_2(s)$ are the geodesic curves on the two manifolds respectively. A well known result in differential geometry is that product metrics induce an associated distance measure on the product manifold,

$$d_{\text{prod}}(\boldsymbol{p}_1 \oplus \boldsymbol{q}_1, \boldsymbol{p}_2 \oplus \boldsymbol{q}_2) = \sqrt{d_1^2(\boldsymbol{p}_1, \boldsymbol{p}_2) + d_2^2(\boldsymbol{q}_1, \boldsymbol{q}_2)} \tag{20}$$

More generally, for $|X|$ simplexes we have the general form for a distance measure as $d_{\text{prod}} = \sqrt{\sum_{i=1}^{|X|} d_i^2}$ and, consequently, we can define a distance function on the space of stochastic matrices.

Result 3 (Stochastic Matrix Distance)**.** *Let $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ be square (row) stochastic matrices. The minimum length geodesic distance between $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ is*

$$d(\boldsymbol{P}_1, \boldsymbol{P}_2) = 2\sqrt{\sum_{x_i \in X} \arccos^2 \sum_{x_j \in X} \sqrt{P_1(x_i, x_j) P_2(x_i, x_j)}} \tag{21}$$

The proof of Result 3 can be found in Apx. C. As well as enjoying many of the previously defined properties of the Bhattacharyya angle, it is also valid for any pair of stochastic matrices. There are no other requirements to compute the distance on the space of stochastic matrices. This appears to be a genuinely new distance function derived via information geometry.

If initial distributions are of primary importance to a practitioner, the sequence distance $d^{(\tau)}$ or its infinite-step form $d^{(\omega)}$ (Section II) would be more appropriate. The stochastic matrix distance is best suited for applications where structural properties of the transition process are the focus, independent of starting conditions.

For a stochastic matrix $\boldsymbol{A}$ and integer $L \geq 1$, we define its L-Cesàro projection as

$$Z_L(\boldsymbol{A}) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} (\boldsymbol{A}^L)^n, \tag{22}$$

which exists for any stochastic matrix by the Perron–Frobenius theorem and represents the matrix contribution of the eigenspace associated with the eigenvalues for which $\lambda^L = 1$ of $\boldsymbol{A}^L$. Note for $L = 1$ this definition is just the classic Cesàro projection.

Result 4. *Let $\boldsymbol{P}_1, \boldsymbol{P}_2 \in [0,1]^{|X| \times |X|}$ be stochastic matrices. Define the long time Cesàro regularised stochastic matrix distance as,*

$$d_\Omega(\boldsymbol{P}_1, \boldsymbol{P}_2) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} d(\boldsymbol{P}_1^n, \boldsymbol{P}_2^n) \tag{23}$$

Further, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ have spectral decompositions $\boldsymbol{P}_1^n = \boldsymbol{A}^n + \boldsymbol{D}_1^n$ and $\boldsymbol{P}_2^n = \boldsymbol{B}^n + \boldsymbol{D}_2^n$, where $\boldsymbol{A}^n, \boldsymbol{B}^n$ are the projections onto the unimodular eigenspaces, and $\boldsymbol{D}_1^n, \boldsymbol{D}_2^n$ any decaying terms. Let $g : [0,1] \to \mathbb{R}$ be a continuous function, and let $\sqrt{\cdot}$ denote the Hadamard square root. Let L be the joint period such that $\boldsymbol{A}^{n+L} = \boldsymbol{A}^n$ and $\boldsymbol{B}^{n+L} = \boldsymbol{B}^n$ for all n. Then our long time distance satisfies:

$$d_\Omega(\boldsymbol{P}_1, \boldsymbol{P}_2) = 2\sqrt{\frac{1}{L} \sum_{k=0}^{L-1} \mathbf{1}^\top \arccos^2 \left[(\sqrt{\boldsymbol{A}^k} \circ \sqrt{\boldsymbol{B}^k}) \mathbf{1} \right]}. \tag{24}$$

The stochastic matrix rate can equivalently be written as

$$d_\Omega(\boldsymbol{P}_1, \boldsymbol{P}_2) = \sqrt{\frac{1}{L} \text{tr} \arccos^2 \left[\sqrt{\Gamma(\boldsymbol{P}_1)} \sqrt{\Gamma(\boldsymbol{P}_2)}^\top \right]}, \tag{25}$$

where $\Gamma(\boldsymbol{A}) = \text{diag}(\boldsymbol{I}_{|X|}, \boldsymbol{A}, \ldots, \boldsymbol{A}^{L-1})(\boldsymbol{I}_L \otimes Z_L(\boldsymbol{A}))$.

A proof of Result 4 is in Apx. D. As noted in the proof, considering ergodic transition matrices in Eq. 24 reduces numerically to considering the sequence distance Eq. 14 if considering just their stationary distributions. I.e. if $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are ergodic with respective assoicated stationary distributions $\bar{\boldsymbol{\pi}}_1, \bar{\boldsymbol{\pi}}_2$ then $L = 1$ and $d_\Omega(\boldsymbol{P}_1, \boldsymbol{P}_2) = 2\sqrt{|X|} \arccos \text{BC}(\bar{\boldsymbol{\pi}}_1, \bar{\boldsymbol{\pi}}_2)$. Note also that by normalising the stochastic matrix distance for ergodic chains by the size of the state space $|X|$ we obtain a distance function which is finite for even countably infinite state spaces.

Intuitively this makes sense as the long term run of an ergodic Markov chain converges to its stationary distribution. Therefore any differences between two Markov chains should only be accounted for by their respective stationary distributions, and not initial conditions. As noted in the mixing times section, the stochastic matrix distance for ergodic chains is also identical to considering the distance between the asymptotic distributions from starting in any arbitrary initial distribution, i.e. $\delta^{(|X|)}(\boldsymbol{e}_x^T \boldsymbol{P}_1^\tau, \boldsymbol{e}_y^T \boldsymbol{P}_2^\tau) \to d_\Omega(\mathbf{1}\bar{\boldsymbol{\pi}}_1^\top, \mathbf{1}\bar{\boldsymbol{\pi}}_2^\top)$ as $\tau \to \infty$. This adds further weight to using the Bhattacharyya angle and stochastic matrix distance as they have a natural correspondence on ergodic chains.

An alternative interpretation is that the stochastic matrix rate equals the expectation value of the Bhattacharyya angle over the rows of the stochastic matrices,

when each row is weighted uniformly. This perspective arises from viewing each row as defining a probability measure over the next state, and averaging the pairwise Bhattacharyya angles across all initial states. Therefore it can be considered as the expectation of the overlap of two processes under the assumption of having equal probability of starting in any state.

IV. NUMERICAL SIMULATION

Clustering is a fundamental data analysis technique that partitions datasets into groups where similar elements are grouped together. This approach has broad applications across domains: identifying user behaviour patterns on websites, segmenting customers into risk profiles for insurance companies, and modelling disease progression pathways (e.g., healthy → diseased → recovered → death) across different patient populations. Common to all scenarios is that individual behaviours are often modelled via stochastic matrices representing Markov chains, making effective distance metrics for comparing such matrices essential for meaningful cluster analysis.

To demonstrate the effectiveness of our proposed distance function for stochastic matrices, we design a simulation that validates its performance against established distance functions in a clustering context. The simulation generates four distinct groups of stochastic matrices, mimicking real-world scenarios where practitioners need to identify and distinguish between groups of transition matrices with known underlying similarities, such as comparing patient cohorts with different disease progression rates, or distinguishing user behaviour patterns across different demographic segments.

We employ 3×3 matrices for their intuitive visual interpretation, facilitating clear understanding of the underlying data structures. The simulation framework proceeds as follows: we establish a reference cluster parameterized by $\boldsymbol{\alpha}_0$, then generate three additional clusters with distinct parameters $\boldsymbol{\alpha}_i$. Each stochastic matrix within cluster k is constructed by sampling each row from a Dirichlet distribution parameterized by $\boldsymbol{\alpha}_i$. To assess clustering performance across varying degrees of separation, we interpolate between clusters using the transformation $\boldsymbol{\alpha}_i(t) = (1-t)\boldsymbol{\alpha}_i + t\boldsymbol{\alpha}_0$ for $t \in [0,1]$, systematically varying the similarity between clusters while measuring the accuracy of different distance functions in correctly identifying cluster membership.

This experimental design directly addresses the practical challenge of distinguishing between groups of stochastic processes when their underlying parameters vary continuously, a common scenario in applications ranging from behavioural analytics to epidemiological modelling.

We choose the adjusted Rand index as the figure of merit for the clustering accuracy and average the index over many runs to mitigate observing a single extreme result [58]. Intuitively, the Rand index measures agreement, across two given clusterings, by checking how many pairs of clustered data elements are in the same cluster, and which pairs of clustered data elements are not in the same cluster. Ideally, all pairings according to one clustering match all parings in the other cluster, giving complete agreement. In practice it is better to adjust the Rand index under the null hypothesis assumption the clusterings were randomly assigned with equal probability by permuting the data elements, hence giving the adjusted Rand index as a measure of agreement.

Formally, given N data points $Y = \{y_i | i \in [N]\}$, for two clusterings $U = \{U_1, \ldots, U_R\}$ and $V = \{V_1, \ldots, V_C\}$ we can quantify the two clusterings contingency as $n_{ij} = |U_i \cap V_j|$. The collection of integers n_{ij} count the overlap between different clusters. In consequence, we define the auxiliary quantities,

$$n_{i+} = \sum_{j=1}^{C} n_{ij}, \quad n_{+j} = \sum_{i=1}^{R} n_{ij} \tag{26a}$$

$$\bar{t} = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2, \quad \bar{r} = \sum_{i=1}^{R} n_{i+}^2, \quad \bar{c} = \sum_{j=1}^{C} n_{+j}^2 \tag{26b}$$

$$\begin{gathered} a = \frac{\bar{t} - N}{2}, \quad b = \frac{\bar{r} - \bar{t}}{2}, \quad c = \frac{\bar{c} - \bar{t}}{2}, \\ d = \frac{\bar{t} - \bar{r} - \bar{c} + N^2}{2} \end{gathered} \tag{26c}$$

Taken together, these auxiliary quantities represent the agreement between two different clusterings of the data Y. We can collect the above into a matrix $\boldsymbol{M}$ and express the adjusted Rand index in a compact manner as,

$$\boldsymbol{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \boldsymbol{\Xi} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{27}$$

$$\mathrm{Rand}_{\mathrm{adjusted}} = \frac{\binom{N}{2}\mathrm{tr}\boldsymbol{M} - \mathrm{tr}\boldsymbol{\Xi}\boldsymbol{M}^2}{\binom{N}{2}^2 - \mathrm{tr}\boldsymbol{\Xi}\boldsymbol{M}^2} \tag{28}$$

Where $\boldsymbol{\Xi}$ is known as the 2×2 exchange matrix (i.e. it exchanges rows and columns). We compared the symmetrised Kullback-Leibler divergence rate, matrix distances induced by the L^1 and L^2 norms, and our new stochastic matrix distances. We use cluster Dirichlet parameters, visualised in Fig. 1, $\boldsymbol{\alpha}_0 = (8,8,8)$, $\boldsymbol{\alpha}_1 = (8,2,2)$, $\boldsymbol{\alpha}_2 = (2,8,2)$, $\boldsymbol{\alpha}_3 = (2,2,8)$.

The results of the simulation are shown in Fig. 2, where a value of one on the vertical axis indicates clusters have been assigned correctly, and lower values mean cluster memberships are incorrect. We observe that the metrics all have a similar behaviour. Moreover, the stochastic matrix distance performance is comparable to the other functions. All functions approach a mean adjusted Rand score of zero as the distributions for clusters converge i.e. the clusters are indistinguishable and hence cluster

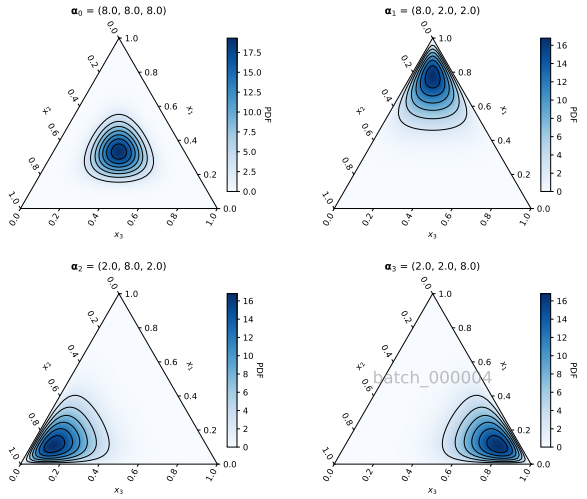

FIG. 1. Ternary plot of the four base Dirichlet distributions used to generate clusters of stochastic matrices. The top-left density plot serves as a reference distribution. The parameters $\boldsymbol{\alpha}_i(t)$ approach $\boldsymbol{\alpha}_0$ linearly such that all distributions eventually coincide, and thus cannot be distinguished.

membership can not be determined. We stress that the purpose of this simulation is not to pick a "best" distance to perform clustering, as this is too subjective, but to show our new distance function indeed behaves sensibly in a controlled setting.

Applications to real world data and settings will be a better test of the new stochastic matrix distance. We do emphasise though the stochastic matrix distance is a true distance measure, purposefully motivated through information geometry and so seems a natural distance measure to pick in the sense of probability theory. In terms of machine learning, both the sequence space and stochastic matrix space distances put forward here seem good candidates for loss functions in their applicable domains.

V. CONCLUSION

We have developed the use of the Bhattacharyya angle as a distance function on discrete Markov chain sequences which is well motivated and respects the intrinsic probability properties of Markov chains. Previous work in discrete Markov theory had shown the use of various functions to compare discrete chains, but this work derives a true distance function which is computable (both analytically and numerically) and valid for any pair of chains.

However these distances explicitly require the initial distributions of a chain and can be extremely sensitive to small changes. The Bhattacharyya rate and stochastic matrix rate vary continuously with the matrix entries. This is in contrast to many divergence-based

Algorithm 1: Steps to generate simulated data for testing different stochastic matrix distance functions. The adjusted Rand index is a measure of how well data has been clustered according to the true cluster label (which is the cluster distribution parameter $\boldsymbol{\alpha}_k$).

Require: Cluster distribution parameters $\{\boldsymbol{\alpha}_k\}_{k=0}^{K}$ (which serve as the true labels for generated data), T (time steps), R (repetitions), N (cluster size), d (distance function)

Ensure : For function d, $K+1 \geq 2$ clusters with N members each, output an array of pairs $(t, \bar{S})$ for step t, mean adjusted Rand index $\bar{S}$ over R clustering repetitions

$t \leftarrow 0$
while $t \leq 1$ **do**
  Update cluster parameters for time step t
  $\alpha_k(t) \leftarrow (1-t)\alpha_k + t\alpha_0$
  $r \leftarrow 0$
  while $r < R$ **do**
    Generate $K \cdot N$ stochastic matrices based on parameters $\{\alpha_k(t)\}_{k=0}^{K}$
    Using function d, calculate the distance matrix for all pairs of the generated matrices
    Cluster the $K \cdot N$ matrices into K labels using the distance matrix
    Calculate the adjusted Rand index $S(r)$ comparing the clustering labels to the known generative clusters
    $r \leftarrow r + 1$
  end
  Calculate the mean adjusted Rand index for time t: $\bar{S}(t) \leftarrow \frac{1}{R}\sum_{r=0}^{R-1} S(r)$
  $t \leftarrow t + 1/T$
end

functions, of which the Kullback–Leibler divergence is a member, which may exhibit discontinuous behaviour under smooth perturbations to distributions. Specifically, the Kullback–Leibler divergence $D_{\mathrm{KL}}(p\|q) = \sum_i p_i \log(p_i/q_i)$ diverges to infinity whenever $q_i \to 0$ while $p_i > 0$, making it undefined or infinite for distributions with differing support. In contrast, the Bhattacharyya coefficient $\sum_i \sqrt{p_i q_i}$ and its derived angle remain finite and continuous even as probability mass approaches zero. In this sense our results might give a more robust measure of similarity, but still contingent to the problem at hand.

Moreover, it is valid for any pair of stochastic matrices, is computable and through Result 4 in the case of ergodic chains the asymptotic step distance is the same as considering distances on the sequence space. We were also able to find compact analytical results for very general settings, only needing a Cesàro type regularisation at most. Using underlying local structures (i.e. state to state jumps) to understand Markov chain similarity in this sense is more akin to comparing Hamiltonians or Lagrangians in physics, rather than looking at the whole

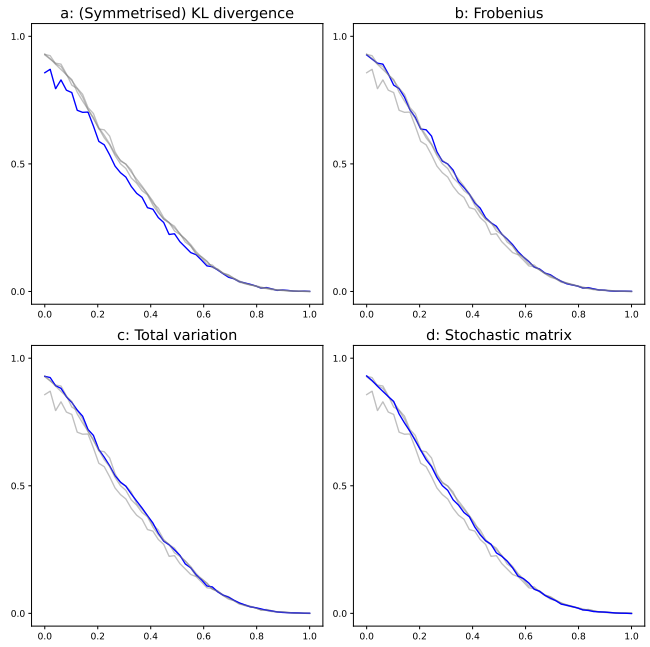

FIG. 2. Comparison of the four selected distance functions' performance over a simulated time series. Progression is from the initial cluster distributions at $t = 0$ (Fig.1) to their final, overlapping state at $t = 1$ (i.e. all distributions look like Fig.1a). Vertical axis: adjusted Rand index (value of 1 is best). Horizontal axis: time parameter $t \in [0, 1]$.

possible trace space dependent on initial distributions.

Returning to our initial motivation, we wanted to find a way of comparing Markov chains for use in healthcare services. For example, if the emergency department of a hospital is described by a Markov chain, we can use our methods of comparing two hospitals. Moreover, each initial distribution of the Markov chain corresponds to a starting event for a patient. As there are many different patients, comparing the sequence space generated by all different patients across two hospitals doesn't have great meaning. What does have more meaning is the direct comparison of the stochastic matrices which represent the generative processes of the hospitals. Therefore our new stochastic matrix distance should have much applicability to comparing health services in any setting represented by Markov chains. Of course this is true for any other field in general.

Despite the number of results, there is still much more worthwhile future work. The exploration of mixing times and the application of the Bhattacharyya angle to general mixing time bounds has only received an elementary review here. Tighter bounds are surely possible. Applications of the Bhattacharyya angle, and its stochastic matrix counterpart, also seem readily available to machine learning in cases where the problem requires the evaluation of a model which outputs categorical probabilities. As both distances are both analytically and numerically approachable, it would be pleasant to see it incorporated in to computational frameworks, such as for gradient boosting [59].

ACKNOWLEDGMENTS

The authors would like to thank Sam Power, George Deligiannidis, Christopher Yau for their constructive and supportive comments for this paper. ARL would like to thank Sara Oriana Gomes Tavares for their very elegant argument on the convergence of the Bhattacharyya angle, which much improved upon his own.

ARL acknowledges the receipt of studentship awards from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme in Health Data Science (Grant Ref: 218529/Z/19/Z). P. Tiňo was supported by the EPSRC Prosperity Partnerships grant ARCANE, EP/X025454/1.

Appendix A: Substochastic R matrix representation

Let X be a finite state space, and let π_1, π_2 be initial distributions over X. Define $\boldsymbol{r} \in [0,1]^{|X|}$ and $\boldsymbol{R} \in [0,1]^{|X|\times|X|}$ (sub-stochastic, i.e., row sums ≤ 1) by:

$$r(x) := \sqrt{\pi_1(x)\pi_2(x)}, \qquad R(x,y) := \sqrt{P_1(x,y)P_2(x,y)},$$

and adopt the convention $\boldsymbol{R}^0 = \boldsymbol{I}$ (equivalently, an empty product equals 1). For $\tau \geq 1$ we have

$$S_\tau = \sum_{(x_1,\ldots,x_\tau)\in X^\tau} r(x_1) \prod_{k=1}^{\tau-1} R(x_k, x_{k+1}) = \sum_{x_1\in X} r(x_1) \sum_{x_2,\ldots,x_\tau \in X} \prod_{k=1}^{\tau-1} R(x_k, x_{k+1}).$$

By the index expansion of matrix powers,

$$(\boldsymbol{R}^{\tau-1}\mathbf{1})(x_1) = \sum_{x_2,\ldots,x_\tau\in X} R(x_1,x_2)\cdots R(x_{\tau-1},x_\tau)\,\mathbf{1}(x_\tau) = \sum_{x_2,\ldots,x_\tau\in X} \prod_{k=1}^{\tau-1} R(x_k,x_{k+1}),$$

since $\mathbf{1}(x_\tau) = 1$ for all $x_\tau \in X$. Substituting gives

$$S_\tau = \sum_{x_1 \in X} r(x_1)(\boldsymbol{R}^{\tau-1}\mathbf{1})(x_1) = \boldsymbol{r}^\top \boldsymbol{R}^{\tau-1}\mathbf{1}.$$

Appendix B: Convergence of Bhattacharyya rate

Theorem 1. *Let X be the state space for a Markov chain with $|X|$ states. Let $\boldsymbol{r} \in [0,1]^{|X|}$ be a non-negative vector and $\boldsymbol{R} \in [0,1]^{|X|\times|X|}$ be sub-stochastic (row sums ≤ 1). There exists a partitioning of X such that Q states form closed recurrent classes and T states are transient. Denote by $\mathbf{1}_Q$ the vector of ones of dimension Q, by $\boldsymbol{T}$ the sub-matrix of transition probabilities among transient states, and by $\boldsymbol{C}$ the sub-matrix of transition probabilities from transient states to recurrent states. When the spectral radius of $\boldsymbol{T}$ is strictly less than unity, the fundamental matrix $\boldsymbol{N} = (\boldsymbol{I} - \boldsymbol{T})^{-1}$ exists, and the absorption probability vector is $\boldsymbol{a} = \boldsymbol{N}\,\boldsymbol{C}\,\mathbf{1}_Q$. The Bhattacharyya rate then satisfies:*

$$\lim_{\tau\to\infty} d^{(\tau)}(M_1, M_2) = 2\arccos\left[\boldsymbol{r}^\top \begin{pmatrix}\mathbf{1}_Q \\ \boldsymbol{a}\end{pmatrix}\right]. \tag{B1}$$

Proof. Let X be a finite state space with $|X|$ states, $\boldsymbol{r} \in [0,1]^{|X|}$, and $\boldsymbol{R} \in [0,1]^{|X|\times|X|}$ be sub-stochastic (row sums ≤ 1). We can write the argument of the sequence distance arccos function as

$$S_\tau \;=\; \boldsymbol{r}^\top \boldsymbol{R}^{\tau-1}\mathbf{1}, \qquad \tau \geq 1,$$

with $\boldsymbol{r} \geq 0$ and $\mathbf{1} \in \mathbb{R}^{|X|}$ the all-ones vector. Permute states so that $\boldsymbol{R}$ is in canonical form for sub-stochastic matrices (recurrent states first, transient states last):

$$\boldsymbol{R} \;=\; \begin{pmatrix}\boldsymbol{Q} & \mathbf{0} \\ \boldsymbol{C} & \boldsymbol{T}\end{pmatrix}, \qquad \mathbf{1} = \begin{pmatrix}\mathbf{1}_Q \\ \mathbf{1}_T\end{pmatrix}, \qquad \boldsymbol{r} = \begin{pmatrix}\boldsymbol{r}_Q \\ \boldsymbol{r}_T\end{pmatrix},$$

where $\boldsymbol{Q}$ is block diagonal with one or more stochastic irreducible blocks (the closed classes); hence $\rho(\boldsymbol{Q}) = 1$, and $\boldsymbol{T}$ is the transient sub-stochastic block with $\rho(\boldsymbol{T}) < 1$. Then for all $\tau \geq 1$,

$$\boldsymbol{R}^{\tau-1} = \begin{pmatrix}\boldsymbol{Q}^{\tau-1} & \mathbf{0} \\ \sum_{k=0}^{\tau-2} \boldsymbol{T}^k \boldsymbol{C} \boldsymbol{Q}^{\tau-2-k} & \boldsymbol{T}^{\tau-1}\end{pmatrix}.$$

Since each block of $\boldsymbol{Q}$ is stochastic and satisfies $\boldsymbol{Q}\mathbf{1}_Q = \mathbf{1}_Q$, we have $\boldsymbol{Q}^m\mathbf{1}_Q = \mathbf{1}_Q$ for all $m \geq 0$. Since $\rho(\boldsymbol{T}) < 1$, we have $\boldsymbol{T}^{\tau-1}\mathbf{1}_T \to \mathbf{0}$ and $\sum_{k=0}^{\tau-2} \boldsymbol{T}^k \to (\boldsymbol{I} - \boldsymbol{T})^{-1}$ as $\tau \to \infty$. Therefore,

$$\boldsymbol{R}^{\tau-1}\mathbf{1} = \begin{pmatrix}\mathbf{1}_Q \\ \sum_{k=0}^{\tau-2} \boldsymbol{T}^k \boldsymbol{C}\mathbf{1}_Q + \boldsymbol{T}^{\tau-1}\mathbf{1}_T\end{pmatrix} \;\longrightarrow\; \begin{pmatrix}\mathbf{1}_Q \\ (\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{C}\mathbf{1}_Q\end{pmatrix} = \begin{pmatrix}\mathbf{1}_Q \\ \boldsymbol{a}\end{pmatrix},$$

where the fundamental matrix of the transient block is $\boldsymbol{N} := (\boldsymbol{I} - \boldsymbol{T})^{-1}$ and

$$\boldsymbol{a} := \boldsymbol{N}\,\boldsymbol{C}\,\mathbf{1}_Q \in \mathbb{R}^T.$$

Taking the inner product with $\boldsymbol{r}^\top$ gives the limit

$$S_\tau \;=\; \boldsymbol{r}^\top \boldsymbol{R}^{\tau-1}\mathbf{1} \;\longrightarrow\; \boldsymbol{r}_Q^\top \mathbf{1}_Q + \boldsymbol{r}_T^\top \boldsymbol{a} \;=\; \boldsymbol{r}^\top \begin{pmatrix}\mathbf{1}_Q \\ \boldsymbol{a}\end{pmatrix}.$$

The result follows from the continuity of the arccos function applied to the limit of S_τ. □

Remark 2. *In the special case where $\boldsymbol{Q}$ is 0×0 (i.e., the entire process is transient), the convergence $\boldsymbol{T}^{\tau-1} \to \mathbf{0}$ gives $S_\tau \to 0$, and thus the Bhattacharyya rate equals π. Conversely, when $\boldsymbol{T}$ is 0×0 (i.e., no transient states), the stochasticity of $\boldsymbol{Q}$ gives $S_\tau \to \|\boldsymbol{r}\|_1$.*

Appendix C: Proof of stochastic matrix distance

Proof. Using the product structure of stochastic matrices, we can decompose two stochastic matrices as

$$\begin{aligned} \boldsymbol{P}_1 &= \left(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{|X|}\right) \\ \boldsymbol{P}_2 &= \left(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{|X|}\right) \end{aligned} \tag{C1}$$

Each $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ are probability vectors in their own right, with dimension $|X|$, and each matrix are therefore elements of the product space $\Delta \times \cdots \times \Delta$. This means we can construct the induced minimum length geodesic curve distance from each underlying simplex space. In other words,

$$d\left(\left(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{|X|}\right), \left(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{|X|}\right)\right) = \sqrt{d^2(\boldsymbol{p}_1, \boldsymbol{q}_1) + \ldots + d^2(\boldsymbol{p}_{|X|}, \boldsymbol{q}_{|X|})} \tag{C2}$$

As the underlying simplex spaces can be equipped with the Bhattacharyya angle, we can write the induced distance as

$$d\left(\left(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{|X|}\right), \left(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{|X|}\right)\right) = 2\sqrt{\sum_{i=1}^{|X|} \arccos^2 \mathrm{BC}(\boldsymbol{p}_i, \boldsymbol{q}_i)} \tag{C3}$$

Noting $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{i|X|})$, $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{i|X|})$, where $p_{ij} = P_1(x_i, x_j)$ and $q_{ij} = P_2(x_i, x_j)$, it is easy to see the Bhattacharyya coefficient expressions can be written as $\mathrm{BC}(\boldsymbol{p}_i, \boldsymbol{q}_i) = \sum_{j=1}^{|X|} \sqrt{p_{ij} q_{ij}}$ which allows us to write

$$d\left(\left(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{|X|}\right), \left(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{|X|}\right)\right) = 2\sqrt{\sum_{x_i \in X} \arccos^2 \sum_{x_j \in X} \sqrt{P_1(x_i, x_j) P_2(x_i, x_j)}} \tag{C4}$$

□

Appendix D: Convergence of Cesàro-regularised stochastic matrix rate

We first state a classical result that enables a self-contained proof.

Lemma 1 (Perron–Frobenius)**.** *For a non-negative stochastic matrix $\boldsymbol{M}$: (i) all eigenvalues satisfy $|\lambda| \leq 1$; (ii) unimodular eigenvalues are semisimple roots of unity; (iii) spectral projections onto unimodular eigenspaces preserve non-negativity [60].*

Recall the definition of the Cesàro projection from Eq. 22: for a stochastic matrix $\boldsymbol{A}$,

$$Z_L(\boldsymbol{A}) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} (\boldsymbol{A}^L)^n,$$

which exists for any stochastic matrix by the Perron–Frobenius theorem and represents the projection onto the invariant subspace of $\boldsymbol{A}^L$.

Theorem 2. *Let $\boldsymbol{P}_1, \boldsymbol{P}_2 \in [0,1]^{|X| \times |X|}$ be stochastic matrices with spectral decompositions $\boldsymbol{P}_1^n = \boldsymbol{A}^n + \boldsymbol{D}_1^n$ and $\boldsymbol{P}_2^n = \boldsymbol{B}^n + \boldsymbol{D}_2^n$, where $\boldsymbol{A}^n, \boldsymbol{B}^n$ are the projections onto the unimodular eigenspaces (i.e., those with $|\lambda| = 1$), and the decaying terms satisfy $\|\boldsymbol{D}_1^n\|, \|\boldsymbol{D}_2^n\| \leq C n^p \gamma^n$ for some $0 \leq \gamma < 1$ and constants $C, p \geq 0$. Let $g : [0,1] \to \mathbb{R}$ be a continuous function, and let $\sqrt{\cdot}$ denote the Hadamard (element-wise) square root. Let the sequence $\{g[(\sqrt{\boldsymbol{A}^n} \circ \sqrt{\boldsymbol{B}^n})\mathbf{1}]\}_{n \in \mathbb{N}}$ have a joint period L, such that $\boldsymbol{A}^{n+L} = \boldsymbol{A}^n$ and $\boldsymbol{B}^{n+L} = \boldsymbol{B}^n$ for all n. Then the Cesàro mean of the transformed Hadamard product satisfies:*

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{1}^\top g\left[(\sqrt{\boldsymbol{P}_1^n} \circ \sqrt{\boldsymbol{P}_2^n})\mathbf{1}\right] = \frac{1}{L} \sum_{k=0}^{L-1} \mathbf{1}^\top g\left[(\sqrt{\boldsymbol{A}^k} \circ \sqrt{\boldsymbol{B}^k})\mathbf{1}\right]. \tag{D1}$$

Proof. The proof follows from the decomposition of the stochastic powers and the properties of Cesàro summability.

1. **Perron–Frobenius theorem:** The Perron–Frobenius theorem guarantees that all eigenvalues of a stochastic matrix $\boldsymbol{M}$ satisfy $|\lambda| \leq 1$. Moreover, unimodular eigenvalues are semisimple roots of unity, meaning they have a fixed integer period. Thus we can always choose $L = \mathrm{lcm}(L_1, L_2)$, where L_1 and L_2 are the periods of the unimodular eigenvalues of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ respectively, to obtain a joint period for all unimodular eigenvalues. In the case both matrices have no oscillatory parts we define $L = 1$.

2. **Spectral decomposition:** For the spectral decomposition, let $\gamma < 1$ be the largest modulus among eigenvalues of $\boldsymbol{P}_1$ or $\boldsymbol{P}_2$ that are strictly inside the unit circle, and let p be the maximum size of all Jordan blocks for $\boldsymbol{P}_1$ or $\boldsymbol{P}_2$. A standard application of the triangle inequality in any matrix norm gives $\|\boldsymbol{D}_1^n\|, \|\boldsymbol{D}_2^n\| \leq Cn^p\gamma^n$ for all $n \in \mathbb{N}$. Thus the spectral decomposition of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ always has the required joint period L and geometrically decaying remainder term, since $\gamma^n \to 0$ dominates the polynomial factor.

3. **Convergence of error term:** Since g is continuous on the compact interval $[0, 1]$, it is uniformly continuous. Given $\boldsymbol{P}_1^n \to \boldsymbol{A}^n$ and $\boldsymbol{P}_2^n \to \boldsymbol{B}^n$ at a geometric rate, we define the error term associated with the transient components:

$$\mathcal{E}(n) = g\left[(\sqrt{\boldsymbol{P}_1^n} \circ \sqrt{\boldsymbol{P}_2^n})\mathbf{1}\right] - g\left[(\sqrt{\boldsymbol{A}^n} \circ \sqrt{\boldsymbol{B}^n})\mathbf{1}\right],$$

   where $\lim_{n\to\infty} \mathcal{E}(n) = \mathbf{0}$ as a consequence of the continuous mapping theorem applied to $\sqrt{\cdot}$ and g.

4. **Vanishing of transient terms:** By the properties of Cesàro means, if a sequence $\mathcal{E}(n)$ converges to a limit $\boldsymbol{F}$, its average also converges to $\boldsymbol{F}$. Since $\mathcal{E}(n) \to \mathbf{0}$, we have:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathcal{E}(n) = \mathbf{0}.$$

5. **Periodic mean convergence:** The term $g\left[(\sqrt{\boldsymbol{A}^n} \circ \sqrt{\boldsymbol{B}^n})\mathbf{1}\right]$ is purely periodic with period L. The Cesàro mean of any periodic sequence with period L converges to the arithmetic mean over one period:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} g\left[(\sqrt{\boldsymbol{A}^n} \circ \sqrt{\boldsymbol{B}^n})\mathbf{1}\right] = \frac{1}{L} \sum_{k=0}^{L-1} g\left[(\sqrt{\boldsymbol{A}^k} \circ \sqrt{\boldsymbol{B}^k})\mathbf{1}\right].$$

6. **Conclusion:** Combining the vanishing of the error term ($\mathcal{E}(n) \to \mathbf{0}$) with the Cesàro mean of the periodic sequence $g\left[(\sqrt{\boldsymbol{A}^n} \circ \sqrt{\boldsymbol{B}^n})\mathbf{1}\right]$, and taking the inner product with $\mathbf{1}^\top$, yields the result.

□

Remark 3. *The limit matrices $\boldsymbol{A}^k$ and $\boldsymbol{B}^k$ are uniquely determined by the stationary projections of the L-th powers of the transition matrices. Specifically, let $Z_L(\boldsymbol{P}_1)$ denote the Cesàro projection of $\boldsymbol{P}_1^L$, and let $Z_L(\boldsymbol{P}_2)$ denote the Cesàro projection of $\boldsymbol{P}_2^L$. Then for any $k \in \{0, \ldots, L-1\}$:*

$$\boldsymbol{A}^k = \boldsymbol{P}_1^k Z_L(\boldsymbol{P}_1) \quad \textit{and} \quad \boldsymbol{B}^k = \boldsymbol{P}_2^k Z_L(\boldsymbol{P}_2).$$

Substituting these into the Cesàro mean yields:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} g\left[(\sqrt{\boldsymbol{P}_1^n} \circ \sqrt{\boldsymbol{P}_2^n})\mathbf{1}\right] = \frac{1}{L} \sum_{k=0}^{L-1} g\left[\left(\sqrt{\boldsymbol{P}_1^k Z_L(\boldsymbol{P}_1)} \circ \sqrt{\boldsymbol{P}_2^k Z_L(\boldsymbol{P}_2)}\right)\mathbf{1}\right]. \tag{D2}$$

Remark 4. *The individual long-term matrices can be collected into a single map defined as:*

$$\Gamma(\boldsymbol{A}) = \mathrm{diag}(\boldsymbol{I}_{|X|}, \boldsymbol{A}, \ldots, \boldsymbol{A}^{L-1})(\boldsymbol{I}_L \otimes Z_L(\boldsymbol{A})),$$

where $\boldsymbol{I}_{|X|}$ is the identity matrix of size $|X| \times |X|$ and $\boldsymbol{I}_L$ is the identity matrix of size $L \times L$. Using the component-wise nature of g and the identity $\mathbf{1}^\top(\boldsymbol{A} \circ \boldsymbol{B})\mathbf{1} = \mathrm{tr}[\boldsymbol{A}\boldsymbol{B}^\top]$, we can rewrite the Cesàro mean as:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{1}^\top g\left[(\sqrt{\boldsymbol{P}_1^n} \circ \sqrt{\boldsymbol{P}_2^n})\mathbf{1}\right] = \frac{1}{L} \mathrm{tr}\, g\left[\sqrt{\Gamma(\boldsymbol{P}_1)}\sqrt{\Gamma(\boldsymbol{P}_2)}^\top\right]. \tag{D3}$$

Remark 5. *The special case* $g = \arccos^2$ *yields the stochastic matrix rate between stochastic matrices* $\boldsymbol{P}_1$ *and* $\boldsymbol{P}_2$*:*

$$d_\Omega(\boldsymbol{P}_1, \boldsymbol{P}_2) = \sqrt{\frac{1}{L}\mathrm{tr}\,\arccos^2\left[\sqrt{\Gamma(\boldsymbol{P}_1)}\sqrt{\Gamma(\boldsymbol{P}_2)}^\top\right]}. \tag{D4}$$

This quantity admits an interpretation as the root-mean-square Bhattacharyya angle over the rows of the stochastic matrices, averaged uniformly across all initial states and over one period of the joint dynamics. This uniform weighting ensures that the distance reflects the intrinsic structure of the transition matrices rather than any particular choice of initial distribution.

[1] A. S. Omer and D. H. Woldegebreal, in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, Vol. 443 LNICST (2022).

[2] M. Abu Alsheikh, D. T. Hoang, D. Niyato, H. P. Tan, and S. Lin, IEEE Communications Surveys and Tutorials **17**, 10.1109/COMST.2015.2420686 (2015).

[3] V. Roy, Convergence diagnostics for markov chain monte carlo (2020).

[4] M. M. Hamdi, H. F. Mahdi, M. S. Abood, R. Q. Mohammed, A. D. Abbas, and A. H. Mohammed, IOP Conference Series: Materials Science and Engineering **1076**, 10.1088/1757-899x/1076/1/012034 (2021).

[5] B. Mor, S. Garhwal, and A. Kumar, Archives of Computational Methods in Engineering **28**, 10.1007/s11831-020-09422-4 (2021).

[6] S. J. van Zelst, F. Mannhardt, M. de Leoni, and A. Koschmider, Granular Computing **6**, 10.1007/s41066-020-00226-2 (2021).

[7] L. D. Valdez, L. Vassallo, and L. A. Braunstein, Physical Review E **107**, 054304 (2023).

[8] Y. Guo, L. Tu, H. Shen, and L. Chai, Physical Review E **106**, 034307 (2022).

[9] D. H. Silva, F. A. Rodrigues, and S. C. Ferreira, Physical Review E **110**, 014302 (2024).

[10] M. A. Achterberg, B. Prasse, and P. Van Mieghem, Physical Review E **105**, 054305 (2022).

[11] N. Martin, G. Van Houdt, and G. Janssenswillen, Expert Systems with Applications **191**, 116274 (2022).

[12] E. De Roock and N. Martin, Journal of biomedical informatics **127**, 103995 (2022).

[13] J. Munoz-Gama, N. Martin, C. Fernandez-Llatas, O. A. Johnson, M. Sepúlveda, E. Helm, V. Galvez-Yanjari, E. Rojas, A. Martinez-Millana, D. Aloini, I. A. Amantea, R. Andrews, M. Arias, I. Beerepoot, E. Benevento, A. Burattin, D. Capurro, J. Carmona, M. Comuzzi, B. Dalmas, R. de la Fuente, C. Di Francescomarino, C. Di Ciccio, R. Gatta, C. Ghidini, F. Gonzalez-Lopez, G. Ibanez-Sanchez, H. B. Klasky, A. Prima Kurniati, X. Lu, F. Mannhardt, R. Mans, M. Marcos, R. Medeiros de Carvalho, M. Pegoraro, S. K. Poon, L. Pufahl, H. A. Reijers, S. Remy, S. Rinderle-Ma, L. Sacchi, F. Seoane, M. Song, A. Stefanini, E. Sulis, A. H. ter Hofstede, P. J. Toussaint, V. Traver, Z. Valero-Ramon, I. v. d. Weerd, W. M. van der Aalst, R. Vanwersch, M. Weske, M. T. Wynn, and F. Zerbato, Journal of Biomedical Informatics **127**, 103994 (2022).

[14] D. Claudio, S. Moyce, T. Albano, E. Ibe, N. Miller, and M. O'Leary, International Journal of Environmental Research and Public Health **20**, 10.3390/ijerph20043525 (2023).

[15] S. McClean, M. Faddy, and P. Millard, in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, Vol. 2006 (2006).

[16] B. A. Craig and P. P. Sendi, Health Economics **11**, 10.1002/hec.654 (2002).

[17] T. Srivastava, N. R. Latimer, and P. Tappenden, PharmacoEconomics **39**, 10.1007/s40273-021-01034-5 (2021).

[18] *Peginterferon beta-1a for treating relapsing–remitting multiple sclerosis*, Tech. Rep. (NICE, 2020).

[19] R. Gatta, J. Lenkowicz, M. Vallati, E. Rojas, A. Damiani, L. Sacchi, B. De Bari, A. Dagliati, C. Fernandez-Llatas, M. Montesi, A. Marchetti, M. Castellano, and V. Valentini, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **10259 LNAI**, 351 (2017).

[20] F. Marazza, F. A. Bukhsh, O. Vijlbrief, J. Geerdink, S. Pathak, M. van Keulen, and C. Seifert, Lecture Notes in Business Information Processing **362 LNBIP**, 496 (2019).

[21] A. F. Ghahfarokhi, A. Berti, and W. M. P. van der Aalst, International Journal of Mechanical and Industrial Engineering **14** (2021).

[22] M. Vallati, S. Orini, M. Lorusso, M. Savino, R. Gatta, and M. Filosto, The International FLAIRS Conference Proceedings **36**, 10.32473/FLAIRS.36.133049 (2023).

[23] V. M. Panaretos and Y. Zemel, Annual Review of Statistics and Its Application **6**, 405 (2019).

[24] R. B. Lyngsø, C. N. Pedersen, and H. Nielsen, in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology, ISMB 1999* (1999).

[25] M. Jaeger, H. Mao, K. Guldstrand Larsen, and R. Mardare, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 8657 LNCS (Springer International Publishing, 2014) pp. 297–312.

[26] P. Daca, T. A. Henzinger, J. Křetínský, and T. Petrov, in *Leibniz International Proceedings in Informatics, LIPIcs*, Vol. 59 (2016).

[27] S. I. Amari, IEEE Transactions on Information Theory **47**, 1701 (2001).

[28] S. i. Amari, Japanese Journal of Mathematics **16**, 1 (2021).

[29] J. E. Cohen, J. H. B. Kemperman, and G. Zbăganu, *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population Sciences*, 1st ed. (Birkhäuser Boston, MA, 1998).

[30] Y. Zhou, Q. Yuan, C. Yang, and Y. Wang, Travel

Behaviour and Society **24**, 10.1016/j.tbs.2021.03.005 (2021).
[31] J. Sanders, A. Proutière, and S.-Y. Yun, The Annals of Statistics **48**, 10.1214/19-AOS1939 (2020).
[32] T. Kawabata and K. Nishikawa, Proteins: Structure, Function and Genetics **41**, 10.1002/1097-0134(20001001)41:1¡108::AID-PROT130¿3.0.CO;2-S (2000).
[33] R. He and J. McAuley, in *Proceedings - IEEE International Conference on Data Mining, ICDM*, Vol. 0 (2016).
[34] C. Wang, B. Wang, H. Liang, J. Zhang, W. Huang, and W. Zhang, IEEE Access **8**, 10.1109/ACCESS.2020.2984456 (2020).
[35] Y. Yao, T. Liu, B. Han, M. Gong, J. Deng, G. Niu, and M. Sugiyama, in *Advances in Neural Information Processing Systems*, Vol. 2020-December (2020).
[36] Y. Zou, T. Liu, D. Liu, and F. Sun, Applied Energy **171**, 10.1016/j.apenergy.2016.03.082 (2016).
[37] G. Du, Y. Zou, X. Zhang, Z. Kong, J. Wu, and D. He, Applied Energy **251**, 10.1016/j.apenergy.2019.113388 (2019).
[38] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, in *Advances in Neural Information Processing Systems*, Vol. 22 (2021).
[39] A. A. Jabri, A. Owens, and A. A. Efros, in *Advances in Neural Information Processing Systems*, Vol. 2020-December (2020).
[40] I. Nevat, D. M. Divakaran, S. G. Nagarajan, P. Zhang, L. Su, L. L. Ko, and V. L. Thing, IEEE/ACM Transactions on Networking **26**, 10.1109/TNET.2017.2765719 (2018).
[41] S. Bacallado, J. D. Chodera, and V. Pande, Journal of Chemical Physics **131**, 10.1063/1.3192309 (2009).
[42] C. C. Strelioff, J. P. Crutchfield, and A. W. Hübler, Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **76**, 10.1103/PhysRevE.76.011106 (2007).
[43] G. Wolfer and S. Watanabe, Frontiers in Physics **11**, 1195562 (2023).
[44] S. Kiefer, in *Leibniz International Proceedings in Informatics, LIPIcs*, Vol. 107 (2018).
[45] F. Nielsen, Entropy 2020, Vol. 22, Page 1100 **22**, 1100 (2020).
[46] H. K. Miyamoto, F. C. C. Meneghetti, J. Pinele, and S. I. R. Costa, Information Geometry 10.1007/s41884-024-00143-2 (2024).
[47] A. Bhattacharyya, The Indian Journal of Statistics (1933-1960) **7** (1946).
[48] F. J. Aherne, N. A. Thacker, and P. I. Rockett, Kybernetika **34** (1998).
[49] S. Bi, M. Broggi, and M. Beer, Mechanical Systems and Signal Processing **117**, 10.1016/j.ymssp.2018.08.017 (2019).
[50] D. Kazakos, IEEE Transactions on Information Theory **24**, 747 (1978).
[51] Z. Rached, F. Alajaji, and L. Campbell, IEEE Transactions on Information Theory **50**, 917 (2004).
[52] M. A. Munem and D. J. Foulis, *Calculus 2ed Munem*, 2nd ed. (Worth, New York, NY, 1984) p. 1048.
[53] Z. Huang, Linear Algebra and Its Applications **434**, 10.1016/j.laa.2010.08.038 (2011).
[54] R. Montenegro and P. Tetali, Mathematical aspects of mixing times in Markov Chains (2006).
[55] M. Dyer, L. A. Goldberg, M. Jerrum, and R. Martin, Probability Surveys **3**, 10.1214/154957806000000041 (2006).
[56] G. Lebanon, IEEE Transactions on Information Theory **51**, 1283 (2005).
[57] G. Montúfar, J. Rauh, and N. Ay, Entropy 2014, Vol. 16, Pages 3207-3233 **16**, 3207 (2014).
[58] D. Steinley, Psychological Methods **9**, 386 (2004).
[59] T. Chen and C. Guestrin, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 13-17-August-2016 (2016).
[60] D. Dembélé, Numerical Linear Algebra with Applications **28**, 10.1002/nla.2340 (2021).