Geometric Reduction for Identity Testing of Reversible Markov Chains

Geoffrey Wolfer[1(✉)] and Shun Watanabe[2]

[1] RIKEN Center for AI Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
geoffrey.wolfer@riken.jp

[2] Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, 2-24-16 Nakamachi, Koganei-shi, Tokyo 184-8588, Japan
shunwata@cc.tuat.ac.jp

Abstract. We consider the problem of testing the identity of a reversible Markov chain against a reference from a single trajectory of observations. Employing the recently introduced notion of a lumping-congruent Markov embedding, we show that, at least in a mildly restricted setting, testing identity to a reversible chain reduces to testing to a symmetric chain over a larger state space and recover state-of-the-art sample complexity for the problem.

Keywords: Irreducible Markov chains · Information geometry · Identity testing · Markov embedding · Congruent embedding · Lumpability

1 Introduction

Uniformity testing is the flagship problem of the modern distribution testing [1] research program. From n independent observations sampled from an unknown distribution μ on a finite space $\mathcal{X}$, the goal is to distinguish between the two cases where μ is uniform and μ is ε-far from being uniform with respect to some notion of distance. The complexity of this problem in total variation is known to be [12] of the order[1] of $\tilde{\Theta}(\sqrt{|\mathcal{X}|}/\varepsilon^2)$, which compares favorably with the linear dependency in $|\mathcal{X}|$ required for estimating the distribution to precision ε [17]. Interestingly, the uniform distribution can be replaced by any arbitrary reference at same statistical cost. In fact, Goldreich [7] proved that the latter problem formally reduces to the former. Inspired by his approach, we seek and obtain a reduction result in the much less understood and more challenging Markovian setting.

[1] As is customary in the property testing literature, we respectively write $\Theta, \mathcal{O}$ and Ω for tight, upper and lower bounds, and the tilda notation suppresses lower-order logarithmic factors in any parameter.

F. Nielsen and F. Barbaresco (Eds.): GSI 2023, LNCS 14071, pp. 328–337, 2023.
https://doi.org/10.1007/978-3-031-38271-0_32

Informal Markovian Problem Statement — The scientist is given the full description of a reference transition matrix $\overline{P}$ and a single Markov chain X_1^n sampled with respect to some unknown transition operator P and arbitrary initial distribution. For fixed proximity parameter $\varepsilon > 0$, the goal is to design an algorithm that distinguishes between the two cases $P = \overline{P}$ and $K(P, \overline{P}) > \varepsilon$, with high probability, where K is a contrast function[2] between stochastic matrices.

Related Work — Under the contrast function (1) described in Sect. 2, and the hypothesis that P and $\overline{P}$ are both irreducible and symmetric over a finite space $\mathcal{X}$, a first tester with sample complexity $\tilde{\mathcal{O}}(|\mathcal{X}|/\varepsilon + h)$, where h [4, Definition 3] is the hitting time of the reference chain, and a lower bound in $\Omega(|\mathcal{X}|/\varepsilon)$, were obtained in [4]. In [3], a graph partitioning algorithm delivers, under the same symmetry assumption, a testing procedure with sample complexity $\mathcal{O}(|\mathcal{X}|/\varepsilon^4)$, i.e. independent of hitting properties. More recently, [6] relaxed the symmetry requirement, replacing it with a more natural reversibility assumption. The algorithm therein has a sample complexity of $\mathcal{O}(1/(\overline{\pi}_\star \varepsilon^4))$, where $\overline{\pi}_\star$ is the minimum stationary probability of the reference $\overline{P}$, gracefully recovering [3] under symmetry. In parallel, [18] started and [2] complemented the research program of inspecting the problem under the infinity norm for matrices, and derived nearly minimax-optimal bounds.

Contribution — We show how to mostly recover [6] from [3] under additional assumptions (see Sect. 3), with a technique based on a geometry preserving embedding. We obtain a more economical proof than [6], which went through the process of re-deriving a graph partitioning algorithm for the reversible case. Furthermore, the impact of our approach, by its generality, stretches beyond the task at hand and is also applicable to related inference problems (see Remark 2).

2 Preliminaries

We let $\mathcal{X}, \mathcal{Y}$ be finite sets, and denote $\mathcal{P}(\mathcal{X})$ the set of all probability distributions over $\mathcal{X}$. All vectors are written as row vectors. For matrices A, B, $\rho(A)$ is the spectral radius of A, $A \circ B$ is the Hadamard product of A and B defined by $(A \circ B)(x, x') = A(x)B(x')$ and $A^{\circ 1/2}(x, x') = \sqrt{A(x, x')}$. For $n \in \mathbb{N}$, we use the compact notation $x_1^n = (x_1, \ldots, x_n)$. $\mathcal{W}(\mathcal{X})$ is the set of all row-stochastic matrices over the state space $\mathcal{X}$, and π is called a stationary distribution for $P \in \mathcal{W}(\mathcal{X})$ when $\pi P = \pi$.

Irreducibility and Reversibility — Let $(\mathcal{X}, \mathcal{D})$ be a digraph with vertex set $\mathcal{X}$ and edge-set $\mathcal{D} \subset \mathcal{X}^2$. When $(\mathcal{X}, \mathcal{D})$ is strongly connected, a Markov chain with connection graph $(\mathcal{X}, \mathcal{D})$ is said to be irreducible. We write $\mathcal{W}(\mathcal{X}, \mathcal{D})$ for the set

[2] General contrast functions under consideration satisfy identity of indiscernibles and non-negativity (e.g. proper metrics induced from matrix norms), and need not satisfy symmetry or the triangle inequality (e.g. information divergence rate between Markov processes).

of irreducible stochastic matrices over $(\mathcal{X}, \mathcal{D})$. When $P \in \mathcal{W}(\mathcal{X}, \mathcal{D})$, π is unique and we denote $\pi_\star \doteq \min_{x\in\mathcal{X}} \pi(x) > 0$ the minimum stationary probability. When P satisfies the detailed-balance equation $\pi(x)P(x,x') = \pi(x')P(x',x)$ for any $(x,x') \in \mathcal{D}$, we say that P is reversible.

Lumpability — In contradistinction with the distribution setting, merging symbols in a Markov chain may break the Markov property, resulting in a hidden Markov model. For $P \in \mathcal{W}(\mathcal{Y}, \mathcal{E})$ and a surjective map $\kappa\colon \mathcal{Y} \to \mathcal{X}$ merging elements of $\mathcal{Y}$ together, we say that P is κ-lumpable [10] when the output process still defines a Markov chain. Introducing $\mathcal{S}_x = \kappa^{-1}(\{x\})$ for the collection of symbols that merge into $x \in \mathcal{X}$, lumpability was characterized by [10, Theorem 6.3.2] as follows. P is κ-lumpable, when for any $x, x' \in \mathcal{X}$, and $y_1, y_2 \in \mathcal{S}_x$, it holds that

$$P(y_1, \mathcal{S}_{x'}) = P(y_2, \mathcal{S}_{x'}).$$

The lumped transition matrix $\kappa_\star P \in \mathcal{W}(\mathcal{X}, \kappa_2(\mathcal{E}))$, with connected edge set

$$\kappa_2(\mathcal{E}) \doteq \left\{(x,x') \in \mathcal{X}^2 \colon \exists (y,y') \in \mathcal{E}, (\kappa(y), \kappa(y')) = (x,x')\right\},$$

is then given by

$$\kappa_\star P(x,x') = P(y, \mathcal{S}_{x'}), \text{ for some } y \in \mathcal{S}_x.$$

Contrast Function — We consider the following notion of discrepancy between two stochastic matrices $P, P' \in \mathcal{W}(\mathcal{X})$,

$$K(P,P') \doteq 1 - \rho\left(P^{\circ 1/2} \circ P'^{\circ 1/2}\right). \tag{1}$$

Although K made its first appearance in [4] in the context of Markov chain identity testing, its inception can be traced back to Kazakos [9]. K is directly related to the Rényi entropy of order 1/2, and asymptotically connected to the Bhattacharyya/Hellinger distance between trajectories (see e.g. proof of Lemma 2). It is instructive to observe that K vanishes on chains that share an identical strongly connected component and does not satisfy the triangle inequality for reducible matrices, hence is not a proper metric on $\mathcal{W}(\mathcal{X})$ [4, p.10, footnote 13]. Some additional properties of K of possible interest are listed in [6, Section 7].

Reduction Approach for Identity Testing of Distributions — Problem reduction is ubiquitous in the property testing literature. Our work takes inspiration from [7], who introduced two so-called "stochastic filters" in order to show how in the distribution setting, identity testing was reducible to uniformity testing, thereby recovering the known complexity of $\mathcal{O}(\sqrt{|\mathcal{X}|}/\varepsilon^2)$ obtained more directly by [14]. Notable works also include [5], who reduced a collection of distribution testing problems to ℓ_2-identity testing.

3 The Restricted Identity Testing Problem

Let $\mathcal{V}_{\mathsf{test}} \subset \mathcal{W}(\mathcal{X})$ be a class of stochastic matrices of interest, and let $\overline{P} \in \mathcal{V}_{\mathsf{test}}$ be a fixed reference. The identity testing problem consists in determining, with high probability, from a single stream of observations $X_1^n = X_1, \ldots, X_n$ drawn according to a transition matrix $P \in \mathcal{V}_{\mathsf{test}}$, whether

$$P \in \mathcal{H}_0 \doteq \left\{\overline{P}\right\}, \text{ or } P \in \mathcal{H}_1(\varepsilon) \doteq \left\{P \in \mathcal{V}_{\mathsf{test}} \colon K(P, \overline{P}) > \varepsilon\right\}.$$

We note the presence of an exclusion region, and regard the problem as a Bayesian testing problem with a prior which is uniform over the two hypotheses classes $\mathcal{H}_0$ and $\mathcal{H}_1(\varepsilon)$ and vanishes on the exclusion region. Casting our problem in the minimax framework, the worst-case error probability $e_n(\phi, \varepsilon)$ of a given test $\phi \colon \mathcal{X}^n \to \{0, 1\}$ is defined as

$$2e_n(\phi, \varepsilon) \doteq \mathbb{P}_{X_1^n \sim \overline{\pi}, \overline{P}}\left(\phi(X_1^n) = 1\right) + \sup_{P \in \mathcal{H}_1(\varepsilon)} \mathbb{P}_{X_1^n \sim \pi, P}\left(\phi(X_1^n) = 0\right).$$

We subsequently define the minimax risk $\mathcal{R}_n(\varepsilon)$ as,

$$\mathcal{R}_n(\varepsilon) \doteq \min_{\phi \colon \mathcal{X}^n \to \{0,1\}} e_n(\phi, \varepsilon),$$

where the minimum is taken over all —possibly randomized— testing procedures. For a confidence parameter δ, the sample complexity is

$$n_\star(\varepsilon, \delta) \doteq \min\left\{n \in \mathbb{N} \colon \mathcal{R}_n(\varepsilon) < \delta\right\}.$$

We briefly recall the assumptions made in [6]. For $(P, \overline{P}) \in (\mathcal{V}_{\mathsf{test}}, \mathcal{H}_0)$,

(A.1) P and $\overline{P}$ are irreducible and reversible.
(A.2) P and $\overline{P}$ share the same[3] stationary distribution $\overline{\pi} = \pi$.

The following additional assumptions will make our approach readily applicable.

(B.1) $P, \overline{P}$ and share the same connection graph, $P, \overline{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D})$.
(B.2) The common stationary distribution is rational, $\overline{\pi} \in \mathbb{Q}^{\mathcal{X}}$.

Remark 1. A sufficient condition for $\overline{\pi} \in \mathbb{Q}^{\mathcal{X}}$ is $\overline{P}(x, x') \in \mathbb{Q}$ for any $x, x' \in \mathcal{X}$.

Without loss of generality, we express $\overline{\pi} = \left(p_1, p_2, \ldots, p_{|\mathcal{X}|}\right)/\Delta$, for some $\Delta \in \mathbb{N}$, and $p \in \mathbb{N}^{|\mathcal{X}|}$ where $0 < p_1 \leq p_2 \leq \cdots \leq p_{|\mathcal{X}|} < \Delta$. We henceforth denote by $\mathcal{V}_{\mathsf{test}}$ the subset of stochastic matrices that verify assumptions $(A.1), (A.2), (B.1)$ and $(B.2)$ with respect to the fixed $\overline{\pi} \in \mathcal{P}(\mathcal{X})$. Our below-stated theorem provides an upper bound on the sample complexity $n_\star(\varepsilon, \delta)$ in $\widetilde{\mathcal{O}}(1/(\overline{\pi}_\star \varepsilon))$.

[3] We note that [6] also slightly loosen the requirement of having a matching stationary distributions to being close in the sense where $\left\|\pi/\overline{\pi} - 1\right\|_\infty < \varepsilon$.

Theorem 1. *Let $\varepsilon, \delta \in (0,1)$ and let $\overline{P} \in \mathcal{V}_{\mathsf{test}} \subset \mathcal{W}(\mathcal{X}, \mathcal{D})$. There exists a randomized testing procedure $\phi\colon \mathcal{X}^n \to \{0,1\}$, with $n = \tilde{\mathcal{O}}(1/(\overline{\pi}_\star \varepsilon^4))$, such that the following holds. For any $P \in \mathcal{V}_{\mathsf{test}}$ and X_1^n sampled according to P, ϕ distinguishes between the cases $P = \overline{P}$ and $K(P, \overline{P}) > \varepsilon$ with error probability less than δ.*

Proof (sketch). Our strategy can be broken down into two steps. First, we employ a transformation on Markov chains, termed Markov embedding [20], in order to symmetrize both the reference chain (algebraically, by computing the new transition matrix) and the unknown chain (operationally, by simulating an embedded trajectory). Crucially, our transformation preserves the contrast between two chains and their embedded version (Lemma 2). Second, we invoke the known tester [3] for symmetric chains as a black box and report its output. The proof is deferred to Sect. 6.

Remark 2. Our reduction approach has applicability beyond recovery of the sample complexity of [6], for instance in the tolerant testing setting, where the two competing hypotheses are

$$K(P, \overline{P}) < \varepsilon/2 \text{ and } K(P, \overline{P}) > \varepsilon.$$

Even in the symmetric setting, this problem remains open. Our technique shows that future work can focus on solving the problem under a symmetry assumption, as we provide a natural extension to the reversible class.

4 Symmetrization of Reversible Markov Chains

Information geometry — Our construction and notation follow [11], who established a dually-flat structure

$$(\mathcal{W}(\mathcal{X}, \mathcal{D}), \mathfrak{g}, \nabla^{(e)}, \nabla^{(m)})$$

on the space of irreducible stochastic matrices, where $\mathfrak{g}$ is a Riemannian metric, and $\nabla^{(e)}, \nabla^{(m)}$ are dual affine (exponential and mixture) connections. Introducing a model $\mathcal{V} = \left\{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\right\} \subset \mathcal{W}(\mathcal{X}, \mathcal{D})$, we write $P_\theta \in \mathcal{V}$ for the transition matrix at coordinates $\theta = (\theta^1, \ldots, \theta^d)$, and where d is the manifold dimension of $\mathcal{V}$. Using the shorthand $\partial_i \cdot \doteq \partial \cdot / \partial \theta^i$, the Fisher metric is expressed [11, (9)] in the chart induced basis $(\partial_i)_{i \in [d]}$ as

$$\mathfrak{g}_{ij}(\theta) = \sum_{(x,x') \in \mathcal{D}} \pi_\theta(x) P_\theta(x, x') \partial_i \log P_\theta(x, x') \partial_j \log P_\theta(x, x'), \text{ for } i, j \in [d]. \quad (2)$$

Following this formalism, it is possible to define mixture families (m-families) and exponential families (e-families) of stochastic matrices [8,11].

Example 1. The class $\mathcal{W}_{\mathsf{rev}}(\mathcal{X}, \mathcal{D})$ of reversible Markov chains irreducible over a connection graph $(\mathcal{X}, \mathcal{D})$ forms both an e-family and an m-family of dimension

$$\dim \mathcal{W}_{\mathsf{rev}}(\mathcal{X}, \mathcal{D}) = \frac{|\mathcal{D}| + |\ell(\mathcal{D})|}{2} - 1,$$

where $\ell(\mathcal{D}) \subset \mathcal{D}$ is the set of loops present in the connection graph [19, Theorem 3,5].

Embeddings — The operation converse to lumping is embedding into a larger space of symbols. In the distribution setting, Markov morphisms were introduced by Čencov [16] as the natural operations on distributions. In the Markovian setting, [20] proposed the following notion of an embedding for stochastic matrices.

Definition 1 (Markov embedding for Markov chains [20]). *We call Markov embedding, a map* $\Lambda_\star \colon \mathcal{W}(\mathcal{X},\mathcal{D}) \to \mathcal{W}(\mathcal{Y},\mathcal{E}), P \mapsto \Lambda_\star P$, *such that for any* $(y,y') \in \mathcal{E}$,

$$\Lambda_\star P(y,y') = P(\kappa(y),\kappa(y'))\Lambda(y,y'),$$

and where κ *and* Λ *satisfy the following requirements*

(*i*) $\kappa \colon \mathcal{Y} \to \mathcal{X}$ *is a lumping function for which* $\kappa_2(\mathcal{E}) = \mathcal{D}$.
(*ii*) Λ *is a positive function over the edge set,* $\Lambda \colon \mathcal{E} \to \mathbb{R}_+$.
(*iii*) *Writing* $\bigcup_{x\in\mathcal{X}} \mathcal{S}_x = \mathcal{Y}$ *for the partition defined by* κ, Λ *is such that for any* $y \in \mathcal{Y}$ *and* $x' \in \mathcal{X}$,

$$(\kappa(y),x') \in \mathcal{D} \implies (\Lambda(y,y'))_{y'\in\mathcal{S}_{x'}} \in \mathcal{P}(\mathcal{S}_{x'}).$$

The above embeddings are characterized as the linear maps over the space of lumpable matrices that satisfy a set of monotonicity requirements and are congruent with respect to the lumping operation [20, Theorem 3.1]. When for any $y,y' \in \mathcal{Y}$, it additionally holds that $\Lambda(y,y') = \Lambda(y')\delta\left[(\kappa(y),\kappa(y')) \in \mathcal{D}\right]$, the embedding $\Lambda_\star$ is called memoryless [20, Section 3.4.2] and is e/m-geodesic affine [20, Th. 3.2, Lemma 3.6], preserving both e-families and m-families of stochastic matrices.

Given $\overline{\pi}$ and Δ as defined in Sect. 3, from [20, Corollary 3.3], there exists a lumping function $\kappa \colon [\Delta] \to \mathcal{X}$, and a memoryless embedding $\sigma_\star^{\overline{\pi}} \colon \mathcal{W}(\mathcal{X},\mathcal{D}) \to \mathcal{W}([\Delta],\mathcal{E})$ with $\mathcal{E} = \left\{(y,y') \in [\Delta]^2 \colon (\kappa(y),\kappa(y')) \in \mathcal{D}\right\}$, such that $\sigma_\star^{\overline{\pi}} \overline{P}$ is symmetric. Furthermore, identifying $\mathcal{X} \cong \{1,2,\ldots,|\mathcal{X}|\}$, its existence is constructively given by

$$\kappa(j) = \operatorname*{arg\,min}_{1\le i\le|\mathcal{X}|} \left\{ \sum_{k=1}^{i} p_k \ge j \right\}, \text{ with } \sigma^{\overline{\pi}}(j) = p^{-1}_{\kappa(j)}, \text{ for any } 1 \le j \le \Delta.$$

As a consequence, we obtain 1. and 2. below.

1. The expression of $\sigma_\star^{\overline{\pi}} \overline{P}$ following algebraic manipulations in Definition 1.
2. A randomized algorithm to memorylessly simulate trajectories from $\sigma_\star^{\overline{\pi}} P$ out of trajectories from P (see [20, Section 3.1]). Namely, there exists a stochastic mapping $\Psi^{\overline{\pi}} \colon \mathcal{X} \to \Delta$ such that,

$$X_1,\ldots,X_n \sim P \implies \Psi^{\overline{\pi}}(X_1^n) = \Psi^{\overline{\pi}}(X_1),\ldots,\Psi^{\overline{\pi}}(X_n) \sim \sigma_\star^{\overline{\pi}} P.$$

5 Contrast Preservation

It was established in [20, Lemma 3.1] that similar to their distribution counterparts, Markov embeddings in Definition 1 preserve the Fisher information metric $\mathfrak{g}$ in (2), the affine connections $\nabla^{(e)}, \nabla^{(m)}$ and the informational (Kullback-Leibler) divergence between points. In this section, we show that memoryless

embeddings, such as the symmetrizer $\sigma_\star^{\overline{\pi}}$ introduced in Sect. 4, also preserve the contrast function K. Our proof will rely on first showing that the memoryless embeddings of [20, Section 3.4.2] induce natural Markov morphisms [15] from distributions over $\mathcal{X}^n$ to $\mathcal{Y}^n$.

Lemma 1. *Let a lumping function $\kappa\colon \mathcal{Y} \to \mathcal{X}$, and*

$$L_\star\colon \mathcal{W}(\mathcal{X},\mathcal{D}) \to \mathcal{W}(\mathcal{Y},\mathcal{E})$$

be a κ-congruent memoryless Markov embedding. For $P \in \mathcal{W}(\mathcal{X},\mathcal{D})$, let $Q^n \in \mathcal{P}(\mathcal{X}^n)$ (resp. $\widetilde{Q}^n \in \mathcal{P}(\mathcal{Y}^n)$) be the unique distribution over stationary paths of length n induced from P (resp. $L_\star P$). Then there exists a Markov morphism $M_\star\colon \mathcal{P}(\mathcal{X}^n) \to \mathcal{P}(\mathcal{Y}^n)$ such that $M_\star Q^n = \widetilde{Q}^n$.

Proof. Let $\kappa_n\colon \mathcal{Y}^n \to \mathcal{X}^n$ be the lumping function on blocks induced from κ,

$$\forall y_1^n \in \mathcal{Y}^n, \kappa_n(y_1^n) = (\kappa(y_t))_{1\le t\le n} \in \mathcal{X}^n,$$

and introduce

$$\mathcal{Y}^n = \bigcup_{x_1^n\in\mathcal{X}^n} \mathcal{S}_{x_1^n}, \text{ with } \mathcal{S}_{x_1^n} = \{y_1^n \in \mathcal{Y}^n\colon \kappa_n(y_1^n) = x_1^n\},$$

the partition associated to κ_n. For any realizable path $x_1^n, Q^n(x_1^n) > 0$, we define a distribution $M^{x_1^n} \in \mathcal{P}(\mathcal{Y}^n)$ concentrated on $\mathcal{S}_{x_1^n}$, and such that for any $y_1^n \in \mathcal{S}_{x_1^n}$, $M^{x_1^n}(y_1^n) = \prod_{t=1}^n L(y_t)$. Non-negativity of $M^{x_1^n}$ is immediate, and

$$\sum_{y_1^n\in\mathcal{Y}^n} M^{x_1^n}(y_1^n) = \sum_{y_1^n\in\mathcal{Y}^n\colon \kappa_n(y_1^n)=x_1^n} M^{x_1^n}(y_1^n) = \prod_{t=1}^n \left(\sum_{y_t\in\mathcal{S}_{x_t}} L(y_t)\right) = 1,$$

thus $M^{x_1^n}$ is well-defined. Furthermore, for $y_1^n \in \mathcal{Y}^n$, it holds that

$$\begin{aligned}\widetilde{Q}^n(y_1^n) &= L_\star\pi(y_1)\prod_{t=1}^{n-1} L_\star P(y_t,y_{t+1}) \overset{(\spadesuit)}{=} \pi(\kappa(y_1))L(y_1)\prod_{t=1}^{n-1} P(\kappa(y_t),\kappa(y_{t+1}))L(y_t)\\ &= Q^n(\kappa(y_1),\dots,\kappa(y_n))\prod_{t=1}^n L(y_t) = Q^n(\kappa_n(y_1^n))\prod_{t=1}^n L(y_t)\\ &= \sum_{x_1^n\in\mathcal{X}^n} Q^n(\kappa_n(y_1^n))M^{x_1^n}(y_1^n) = M_\star Q^n(y_1^n),\end{aligned}$$

where ($\spadesuit$) stems from [20, Lemma 3.5], whence our claim holds.

Lemma 1 essentially states that the following diagram commutes

$$\begin{array}{ccc}\mathcal{W}(\mathcal{X},\mathcal{D}) & \xrightarrow{L_\star} & L_\star\mathcal{W}(\mathcal{X},\mathcal{D})\\ \downarrow & & \downarrow\\ \mathcal{Q}^n_{\mathcal{W}(\mathcal{X},\mathcal{D})} & \xrightarrow{M_\star} & \mathcal{Q}^n_{L_\star\mathcal{W}(\mathcal{X},\mathcal{D})},\end{array}$$

for the Markov morphism $M_\star$ induced by $L_\star$, and where we denoted $\mathcal{Q}^n_{\mathcal{W}(\mathcal{X},\mathcal{D})} \subset \mathcal{P}(\mathcal{X}^n)$ for the set of all distributions over paths of length n induced from the family $\mathcal{W}(\mathcal{X},\mathcal{D})$. As a consequence, we can unambiguously write $L_\star Q^n \in \mathcal{Q}^n_{L_\star \mathcal{W}(\mathcal{X},\mathcal{D})}$ for the distribution over stationary paths of length n that pertains to $L_\star P$.

Lemma 2. *Let $L_\star \colon \mathcal{W}(\mathcal{X},\mathcal{D}) \to \mathcal{W}(\mathcal{Y},\mathcal{E})$ be a memoryless embedding,*

$$K(L_\star P, L_\star \overline{P}) = K(P, \overline{P}).$$

Proof. We recall for two distributions $\mu, \nu \in \mathcal{P}(\mathcal{X})$ the definition of $R_{1/2}$ the Rényi entropy of order $1/2$,

$$R_{1/2}(\mu \| \nu) \doteq -2 \log \left(\sum_{x \in \mathcal{X}} \sqrt{\mu(x)\nu(x)} \right),$$

and note that $R_{1/2}$ is closely related to the Hellinger distance between μ and ν. This definition extends to the notion of a divergence rate between stochastic processes $(X_t)_{t\in\mathbb{N}}, (X'_t)_{t\in\mathbb{N}}$ on $\mathcal{X}$ as follows

$$R_{1/2}\left((X_t)_{t\in\mathbb{N}} \| (X'_t)_{t\in\mathbb{N}}\right) = \lim_{n\to\infty} \frac{1}{n} R_{1/2}\left(X_1^n \| X_1'^n\right),$$

and in the irreducible time-homogeneous Markovian setting where $(X_t)_{t\in\mathbb{N}}$, $(X'_t)_{t\in\mathbb{N}}$ evolve according to transition matrices P and P', the above reduces [13] to

$$R_{1/2}\left((X_t)_{t\in\mathbb{N}} \| (X'_t)_{t\in\mathbb{N}}\right) = -2 \log \rho(P^{\circ 1/2} \circ P'^{\circ 1/2}) = -2 \log(1 - K(P, P')).$$

Reorganizing terms and plugging for the embedded stochastic matrices,

$$K(L_\star P, L_\star \overline{P}) = 1 - \exp\left(-\frac{1}{2} \lim_{n\to\infty} \frac{1}{n} R_{1/2}\left(L_\star Q^n \| L_\star \overline{Q}^n \right) \right),$$

where $L_\star \overline{Q}^n$ is the distribution over stationary paths of length n induced by the embedded $L_\star \overline{P}$. For any $n \in \mathbb{N}$, from Lemma 1 and information monotonicity of the Rényi divergence, $R_{1/2}\left(L_\star Q^n \| L_\star \overline{Q}^n \right) = R_{1/2}\left(Q^n \| \overline{Q}^n \right)$, hence our claim.

6 Proof of Theorem 1

We assume that P and $\overline{P}$ are in $\mathcal{V}_{\mathsf{test}}$. We reduce the problem as follows. We construct $\sigma^{\overline{\pi}}_\star$, the symmetrizer[4] defined in Sect. 4. We proceed to embed both the reference chain (using Definition 1) and the unknown trajectory (using the operational definition in [20, Section 3.1]). We invoke the tester of [3] as a black box, and report its answer.

[4] If we wish to test for the identity of multiple chains against the same reference, we only need to perform this step once.

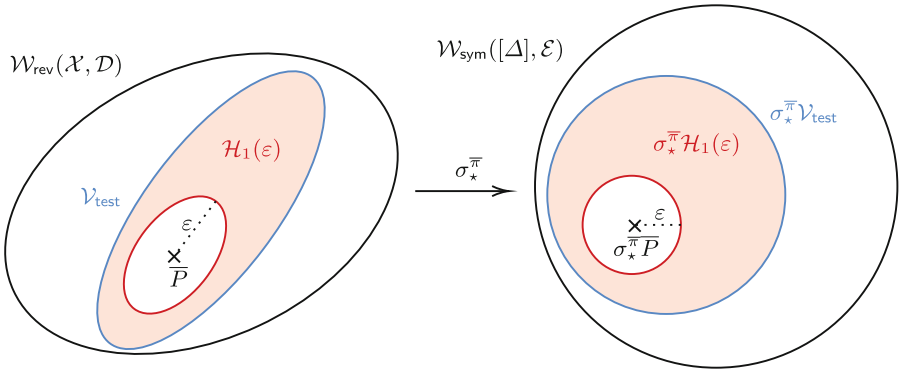

Fig. 1. Reduction of the testing problem by isometric embedding.

Completeness case. It is immediate that $P = \overline{P} \implies \sigma_\star^{\overline{\pi}} P = \sigma_\star^{\overline{\pi}} \overline{P}$.

Soundness case. From Lemma 2, $K(P, \overline{P}) > \varepsilon \implies K(\sigma_\star^{\overline{\pi}} P, \sigma_\star^{\overline{\pi}} \overline{P}) > \varepsilon$.

As a consequence of [3, Theorem 10], the sample complexity of testing is upper bounded by $\mathcal{O}(\Delta/\varepsilon^4)$. With $\overline{\pi}_\star = p_1/\Delta$ and treating p_1 as a small constant, we recover the known sample complexity.

Acknowledgements. GW is supported by the Special Postdoctoral Researcher Program (SPDR) of RIKEN and by the Japan Society for the Promotion of Science KAKENHI under Grant 23K13024. SW is supported in part by the Japan Society for the Promotion of Science KAKENHI under Grant 20H02144.

References

1. Canonne, C.L., et al.: Topics and techniques in distribution testing: A biased but representative sample. Found. Trends Commun. Inf. Theor. **19**(6), 1032–1198 (2022)
2. Chan, S.O., Ding, Q., Li, S.H.: Learning and testing irreducible Markov chains via the k-cover time. In: Algorithmic Learning Theory. pp. 458–480. PMLR (2021)
3. Cherapanamjeri, Y., Bartlett, P.L.: Testing symmetric Markov chains without hitting. In: Proceedings of the Thirty-Second Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 99, pp. 758–785. PMLR (2019)
4. Daskalakis, C., Dikkala, N., Gravin, N.: Testing symmetric Markov chains from a single trajectory. In: Conference On Learning Theory. pp. 385–409. PMLR (2018)
5. Diakonikolas, I., Kane, D.M.: A new approach for testing properties of discrete distributions. In: 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS). pp. 685–694. IEEE (2016)
6. Fried, S., Wolfer, G.: Identity testing of reversible Markov chains. In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 151, pp. 798–817. PMLR (2022)
7. Goldreich, O.: The uniform distribution is complete with respect to testing identity to a fixed distribution. In: Electron. Colloquium Comput. Complex. vol. 23, p. 15 (2016)

8. Hayashi, M., Watanabe, S.: Information geometry approach to parameter estimation in Markov chains. Ann. Stat. **44**(4), 1495–1535 (2016)
9. Kazakos, D.: The Bhattacharyya distance and detection between Markov chains. IEEE Trans. Inf. Theor. **24**(6), 747–754 (1978)
10. Kemeny, J.G., Snell, J.L.: Finite Markov chains: with a new appendix Generalization of a fundamental matrix. Springer (1983)
11. Nagaoka, H.: The exponential family of Markov chains and its information geometry. In: The proceedings of the Symposium on Information Theory and Its Applications. vol. 28(2), pp. 601–604 (2005)
12. Paninski, L.: A coincidence-based test for uniformity given very sparsely sampled discrete data. IEEE Trans. Inf. Theor. **54**(10), 4750–4755 (2008)
13. Rached, Z., Alajaji, F., Campbell, L.L.: Rényi's divergence and entropy rates for finite alphabet Markov sources. IEEE Trans. Inf. Theor. **47**(4), 1553–1561 (2001)
14. Valiant, G., Valiant, P.: An automatic inequality prover and instance optimal identity testing. SIAM J. Comput. **46**(1), 429–455 (2017)
15. Čencov, N.N.: Algebraic foundation of mathematical statistics. Series Stat. **9**(2), 267–276 (1978)
16. Čencov, N.N.: Statistical decision rules and optimal inference, transl. math. monographs, vol. 53. Amer. Math. Soc., Providence-RI (1981)
17. Waggoner, B.: l_p testing and learning of discrete distributions. In: Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science. pp. 347–356 (2015)
18. Wolfer, G., Kontorovich, A.: Minimax testing of identity to a reference ergodic Markov chain. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. vol. 108, pp. 191–201. PMLR (2020)
19. Wolfer, G., Watanabe, S.: Information geometry of reversible Markov chains. Inf. Geom. **4**(2), 393–433 (12 2021)
20. Wolfer, G., Watanabe, S.: Geometric aspects of data-processing of Markov chains (2022), arXiv:2203.04575