arXiv:2107.11858v2 [math.ST] 10 Dec 2021

Estimation of Stationary Optimal Transport Plans

Kevin O'Connor
UNC-Chapel Hill

Kevin McGoff
UNC-Charlotte

Andrew B Nobel
UNC-Chapel Hill

December 13, 2021

Abstract

We study optimal transport for stationary stochastic processes taking values in finite spaces. In order to reflect the stationarity of the underlying processes, we restrict attention to stationary couplings, also known as joinings. The resulting optimal joining problem captures differences in the long run average behavior of the processes of interest. We introduce estimators of both optimal joinings and the optimal joining cost, and we establish consistency of the estimators under mild conditions. Furthermore, under stronger mixing assumptions we establish finite-sample error rates for the estimated optimal joining cost that extend the best known results in the iid case. Finally, we extend the consistency and rate analysis to an entropy-penalized version of the optimal joining problem.

1 Introduction

The application and theory of optimal transport has recently received a great deal of attention in statistics and machine learning. This work has resulted in novel approaches to statistical estimation [11, 40, 30, 3, 6], deep generative modeling [2, 81], clustering [37, 46, 56], cell modeling [71], and other applications [13, 12, 28, 27, 86]. In most of this work the objects under study (for example, images, text documents, graphs, and point clouds) are regarded as static and do not evolve over time. Accordingly, the distributions and cost functions appearing in the optimal transport problem capture the behavior of these objects at a fixed point in time. In this static setting the statistical properties of the optimal transport problem, such as definition of estimators, consistency, and rates of convergence, have been well-studied [65, 67, 79, 76, 68, 39, 20, 66, 72].

In contrast with the static situation, we are interested in optimal transport problems in settings where the objects of interest are processes that evolve dynamically over time. Examples include the alignment or generative modeling of text sequences or musical scores, or hybrid settings involving dynamic text and images. Other examples include transportation of goods between a set of manufacturers and a set of retailers when supply and demand vary over time in a stochastic fashion, or the comparison of brain networks observed at multiple points in time. Problems involving sequential data or objects varying over time are pervasive in statistics and machine learning. While standard optimal transport techniques can be applied to such problems, they do not account for the structure of the underlying measures, which reflect dynamic processes rather than static quantities.

In this paper we investigate optimal transport for finite alphabet stationary ergodic processes, together with cost functions that measure differences at a single time point (or finitely many time points). Optimal transport for stationary processes is a special case of the ordinary optimal transport problem in which the distributions of interest are shift invariant measures on infinite product spaces (the sequence spaces associated with the given processes). As such, existing methods and theory apply. However, it is easy to show that a coupling of two stationary processes need not be stationary, and the same is true of optimal transport plans (see Examples 1 and 2 below). To address these, and other, issues arising in the general setting, we restrict our attention to *stationary* couplings of stationary processes. This seemingly mild restriction has far reaching consequences.

In restricting to stationary couplings, also known as *joinings*, we draw connections between optimal transport and ergodic theory. In particular, properties and theoretical applications of joinings have been studied in ergodic theory for many years. Given two stationary processes, the *optimal joining problem* is the problem of finding a joining of the processes having minimal expected cost. The primary focus of this paper is estimating an optimal joining, and the associated optimal joining cost, of two finite alphabet stationary ergodic processes using n observations from each process. Roughly speaking, we use the available observations to estimate the k-dimensional distribution of each process, find an optimal coupling of these k-dimensional estimates, and then use this coupling to construct a joint process that is stationary. In order to ensure that the constructed process converges to an optimal joining, it is necessary to balance estimating the k-dimensional distribution by letting the sample size n grow and learning the dependence structure of the optimal joining by letting k grow. Thus the task of choosing an appropriate sequence $\{k(n)\}$ of block sizes indexed by sample size is critical for consistently estimating an optimal joining.

Under the stated assumptions, we show that there exists a sequence $\{k(n)\}$ for which the corresponding joint processes will converge to an optimal joining of the marginal processes, and the expected cost of the joint processes will converge to the cost of the optimal joining (Theorem 3). This result generalizes existing results regarding the estimation of optimal joining costs and, to our knowledge, provides the first consistency result for estimating optimal joinings. Under additional mixing assumptions on the observed processes, we identify an explicit growth rate for k and obtain rates of convergence for estimates of the optimal joining cost (Theorem 5). To the best of our knowledge, these are the first finite-sample bounds for estimation of an optimal joining cost. In the iid case, optimal joining and optimal transport coincide, and we recover existing, state-of-the-art bounds for estimation of the optimal transport cost. As special cases of our results we obtain new, finite-sample bounds for estimation of the $\overline{d}$- and $\overline{\rho}$-distances between stationary ergodic processes.

In recent years, there has been a substantial amount of work on regularized optimal transport, in which the regularization is obtained by adding an entropic penalty to the usual optimal transport cost. As with (unregularized) optimal transport, the regularized problem has been primarily studied in static settings. In this work, we bring these ideas to a dynamic setting. More specifically, we propose and analyze a regularized form of the optimal joining problem, which is obtained by adding a penalty based on the entropy rate of a process to the expected cost. We then extend our estimation scheme from the standard optimal joining problem to the regularized problem, and we establish both consistency and rate results for the resulting estimates (Theorems 9 and 10). Existing algorithms for computing regularized optimal transport plans may be applied to compute the proposed estimates in the regularized case more efficiently compared to the unregularized case.

Organization of the Paper. The rest of the paper is organized as follows. Background on optimal transport, optimal joinings, some initial results, and related work are presented in the next section. In Section 3, we detail the proposed estimation scheme for an optimal joining and its expected cost and state our main consistency result. Statement of our finite sample error bound under mixing assumptions, and a corollary, are presented in Section 4. In Section 5, we introduce the entropic optimal joining problem and discuss how our estimation scheme, consistency result, and error bound extend to this problem. We close with a discussion of our results in Section 6. Proofs of the main results are presented in Section 7.

2 Preliminaries and First Results

We begin by defining the optimal transport problem. Let $\mathcal{U}$ and $\mathcal{V}$ be metric spaces and let $\mathcal{M}(\mathcal{U})$ and $\mathcal{M}(\mathcal{V})$ denote the set of Borel probability measures on these spaces. Recall that a Borel probability measure $\pi \in \mathcal{M}(\mathcal{U} \times \mathcal{V})$ on the product space $\mathcal{U} \times \mathcal{V}$ is said to be a *coupling* of $\mu \in \mathcal{M}(\mathcal{U})$ and $\nu \in \mathcal{M}(\mathcal{V})$ if for every measurable $A \subset \mathcal{U}$ and $B \subset \mathcal{V}$, $\pi(A \times \mathcal{V}) = \mu(A)$ and $\pi(\mathcal{U} \times B) = \nu(B)$, that is, the $\mathcal{U}$-marginal of π is μ and the $\mathcal{V}$-marginal of π is ν. Given a non-negative cost function $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$, the optimal transport problem is to minimize the expectation of c over the set of

couplings $\Pi(\mu,\nu)$:

$$T_c(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int c\, d\pi. \tag{2.1}$$

Any optimal solution to (2.1) is referred to as an optimal transport plan (or coupling), and the quantity $T_c(\mu,\nu)$ is referred to as the optimal transport cost of μ and ν with respect to c. For an introduction to couplings and the optimal transport problem, we refer the reader to [83].

As noted in Section 1, the optimal transport problem is general, and does not incorporate information about about the structure of the sets $\mathcal{U}$ and $\mathcal{V}$, or the measures μ and ν. Existing work has considered different choices of $\mathcal{U}$ and $\mathcal{V}$, including finite dimensional Euclidean spaces [81, 2], graphs [88, 89, 80], trees [85], finite sets [76, 57], and sequence spaces [60, 45]. In what follows, we consider the latter case in which μ and ν are shift-invariant probability measures on sequence spaces, equivalently the distributions of stationary processes.

2.1 Couplings of Stationary Processes

We briefly cover some notation and background on stationary probability measures and processes. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets with their discrete topology, and let $\mathcal{U} = \mathcal{X}^{\mathbb{N}}$ and $\mathcal{V} = \mathcal{Y}^{\mathbb{N}}$ be associated sequence spaces. Elements $\mathbf{x} = (x_1, x_2, \ldots) \in \mathcal{U}$ and $\mathbf{y} = (y_1, y_2, \ldots) \in \mathcal{V}$ are infinite sequence with entries in $\mathcal{X}$ and $\mathcal{Y}$, respectively. For a Borel measure $\mu \in \mathcal{M}(\mathcal{X}^{\mathbb{N}})$, let $\mu_k \in \mathcal{M}(\mathcal{X}^k)$ denote the distribution of the first k coordinates of $\mathbf{x} = (x_1, x_2, \ldots)$ under μ, that is, $\mu_k(A) = \mu(A \times \mathcal{X} \times \mathcal{X} \times \cdots)$ for each $A \subseteq \mathcal{X}^k$. Let $\sigma : \mathcal{X}^{\mathbb{N}} \to \mathcal{X}^{\mathbb{N}}$ be the left-shift map on $\mathcal{X}^{\mathbb{N}}$ defined by $\sigma(x_1, x_2, ...) = x_2, x_3, \ldots$; note that σ is continuous under the usual product topology on $\mathcal{X}^{\mathbb{N}}$. A Borel measure $\mu \in \mathcal{M}(\mathcal{X}^{\mathbb{N}})$ is said to be *stationary* if $\mu \circ \sigma^{-1} = \mu$. A stationary measure μ is said to be *ergodic* if $\mu(A) \in \{0, 1\}$ for any measurable set $A \subset \mathcal{X}^{\mathbb{N}}$ such that $\sigma^{-1}(A) = A$. Let $\mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ denote the set of stationary Borel measures on $\mathcal{X}^{\mathbb{N}}$. In the same way we may define the left-shift $\tau : \mathcal{Y}^{\mathbb{N}} \to \mathcal{Y}^{\mathbb{N}}$ and the corresponding set of stationary measures $\mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ on $\mathcal{Y}^{\mathbb{N}}$. For the remainder of the paper, we fix the finite spaces $\mathcal{X}$ and $\mathcal{Y}$, and consider stationary ergodic measures $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$.

It is helpful to recall the simple equivalence between stationary measures and stationary processes. The measure $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ corresponds to a stationary process $X = X_1, X_2, \ldots$ with $X_i \in \mathcal{X}$ via the relation $\mathbb{P}(X_1^k \in A) = \mu_k(A)$ for all $A \subseteq \mathcal{X}^k$ and all $k \geq 1$. Ergodicity of μ is equivalent to ergodicity of the process X. In the same way, the measure $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ corresponds to a stationary process $Y = Y_1, Y_2, \ldots$ with values in $\mathcal{Y}$. Each coupling $\pi \in \Pi(\mu,\nu)$ corresponds to a joint process $(\tilde{X}, \tilde{Y}) = (\tilde{X}_1, \tilde{Y}_1), (\tilde{X}_2, \tilde{Y}_2), \ldots$ such that $\tilde{X} \overset{d}{=} X$ and $\tilde{Y} \overset{d}{=} Y$. With a slight abuse of notation, we will use $\Pi(X, Y)$ to refer to the set of couplings of stationary processes X and Y. Importantly, the definition of coupling does *not* ensure that the joint process $(\tilde{X}, \tilde{Y})$ is stationary, even when X and Y are (see Examples 1 and 2 below).

In what follows we will consider a non-negative, single letter cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ that is defined on pairs of elements from $\mathcal{X}$ and $\mathcal{Y}$. Note that c is necessarily bounded as $\mathcal{X}$ and $\mathcal{Y}$ are finite. By considering sliding blocks, our results may be readily extended to cost functions depending on a finite number of letters. Single (and finite) letter cost functions are the norm in information theory, and are natural when making inferences about processes that are only partially observed. Note also that any cost function $c' : \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to \mathbb{R}_+$ that is continuous in the usual product topology can be uniformly approximated by a finite letter cost function.

Any single letter cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ can be extended to a cost function $c_0 : \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to \mathbb{R}_+$ on infinite sequences by defining $c_0(\mathbf{x}, \mathbf{y}) = c(x_1, y_1)$. In this case, the optimal transport problem for stationary measures $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ can be written as

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c_0\, d\pi \;=\; \inf_{\pi\in\Pi(\mu,\nu)} \int c\, d\pi_1 \;=\; \inf_{(\tilde{X},\tilde{Y})\in\Pi(X,Y)} \mathbb{E}c(\tilde{X}_1, \tilde{Y}_1) \tag{2.2}$$

where X and Y are the stationary processes associated with μ and ν respectively, and π_1 is the one-dimensional marginal distribution of π. In this case we will, with a slight abuse of notation, write $T_c(\mu,\nu)$ instead of $T_{c_0}(\mu,\nu)$.

In order to motivate the consideration of stationary couplings, we present two elementary examples. The first shows that couplings of stationary processes can have very different properties than the processes being coupled.

Example 1 ([60]). Let X and Y be Bernoulli($1/2$) processes, which need not be defined on the same probability space. Let $Z = Z_1, Z_2, \ldots$ be iid Bernoulli($1/2$) random variables defined on a common probability space. Define new processes $\tilde{X} = \tilde{X}_1, \tilde{X}_2, \ldots$ and $\tilde{Y} = \tilde{Y}_1, \tilde{Y}_2, \ldots$ in terms of Z as follows:

$$\tilde{X}_i = Z_{2i} \quad \text{and} \quad \tilde{Y}_i = \begin{cases} Z_{2i+1} & \text{if } i \neq 2^k \\ Z_{2i} & \text{if } \exists k,\ i = 2^k. \end{cases}$$

Clearly $\tilde{X} \stackrel{d}{=} X$ and $\tilde{Y} \stackrel{d}{=} Y$, and therefore the joint process $(\tilde{X}, \tilde{Y}) = (\tilde{X}_1, \tilde{Y}_1), (\tilde{X}_2, \tilde{Y}_2), \ldots$ is a coupling of X and Y. However, it is easy to see that joint process $(\tilde{X}, \tilde{Y})$ is not stationary and that it exhibits long-range dependence.

The next example shows that the optimal transport distance between two stationary processes can be zero even when the processes are distinct.

Example 2. Suppose that $\mathcal{X} = \mathcal{Y}$ and that $c(x, y) = \mathbb{1}(x \neq y)$ is the 0-1 cost function. Let X and Y be two stationary processes that are distinct but share the same one-dimensional distribution. (For example, X might be a Bernoulli($1/2$) process and Y a stationary Markov chain that cycles between 0 and 1.) We may couple X and Y as follows. First, let $\tilde{X}_1 = \tilde{Y}_1$. For $k \geq 2$ generate $\tilde{X}_k$ according to the conditional distribution of X_k given X_1^{k-1} and independently generate $\tilde{Y}_k$ according to the conditional distribution of Y_k given Y_1^{k-1}. The resulting process $(\tilde{X}, \tilde{Y})$ is a non-stationary coupling of X and Y. Moreover $\mathbb{E}c(\tilde{X}_1, \tilde{Y}_1) = 0$ by design, and therefore the optimal transport cost between X and Y is zero in this case, despite the fact that X and Y are distinct processes.

2.2 Joinings of Stationary Processes

Example 2 illustrates an important feature of the optimal transport problem for stationary processes: for single or finite letter cost functions, an optimal coupling $(\tilde{X}, \tilde{Y})$ of processes X and Y need only align the initial components of $\tilde{X}$ and $\tilde{Y}$. If, for example, X and Y represent discrete-time audio and video sequences, an optimal coupling will seek to align the initial sequence of audio and video, but will be insensitive to differences at subsequent time points. While the use of infinite-letter cost functions can address some of these issues, infinite or long-range costs can be problematic in practice. More importantly, as the examples above illustrate, optimal couplings of stationary processes need not be stationary. From a theoretical and practical perspective, it is natural to consider stationary couplings of stationary processes, which are also known as joinings.

Definition 1. *A probability measure λ is a* joining *of $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ if λ is a coupling of μ and ν and is itself stationary, that is, $\lambda \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$. The set of joinings of μ and ν will be denoted by $\mathcal{J}(\mu, \nu)$.*

Equivalently, a joining of two stationary processes X and Y is a coupling $(\tilde{X}, \tilde{Y})$ that is itself stationary. This set of processes will be referred to as $\mathcal{J}(X, Y)$. Note that $\mathcal{J}(\mu, \nu)$ is non-empty, as the independent coupling $\mu \otimes \nu$ is always stationary. Joinings were introduced by Furstenberg [31], and have been studied extensively in the ergodic theory literature since that time; an overview and more details can be found in [19, 34].

Restricting the optimal transport problem (2.2) (with a single letter cost) to the set of stationary couplings leads to the *optimal joining problem*: given measures $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ find

$$S_c(\mu, \nu) = \inf_{\lambda \in \mathcal{J}(\mu,\nu)} \int c \, d\lambda_1 = \inf_{(\tilde{X},\tilde{Y}) \in \mathcal{J}(X,Y)} \mathbb{E}c(\tilde{X}_1, \tilde{Y}_1) \tag{2.3}$$

where X and Y are the stationary processes associated with μ and ν respectively. We will denote the set of joinings attaining the infimum in (2.3) by $\mathcal{J}_{\min}(\mu,\nu)$. Elements of $\mathcal{J}_{\min}(\mu,\nu)$ will be called optimal joinings, and $S_c(\mu,\nu)$ will be called the optimal joining cost. Note that when there is no risk of confusion, we will omit the cost c in our notation for the optimal joining cost, writing $S(\mu,\nu)$ instead of $S_c(\mu,\nu)$. The following proposition collects some standard properties of $\mathcal{J}(\mu,\nu)$ and $\mathcal{J}_{\min}(\mu,\nu)$; for more details see [74] or [54] and the references therein.

Proposition 1. *Under the stated assumptions, the set $\mathcal{J}(\mu,\nu)$ is non-empty, convex, and compact in the weak topology, and its extreme points coincide with ergodic joinings of μ and ν. Moreover, the set $\mathcal{J}_{min}(\mu,\nu)$ of optimal joinings of μ and ν is non-empty, convex, and compact in the weak topology, and its extreme points coincide with the set of ergodic optimal joinings.*

It turns out that there are close connections between the optimal joining problem and the optimal transport problem using long run average cost. For $k \geq 1$ let $c_k : \mathcal{X}^k \times \mathcal{Y}^k \to \mathbb{R}_+$ be the k-step cumulative cost defined by $c_k(x_1^k, y_1^k) = \sum_{\ell=1}^k c(x_\ell, y_\ell)$, and let $\overline{c}(\mathbf{x},\mathbf{y}) = \limsup_{k\to\infty} k^{-1} c_k(x_1^k, y_1^k)$.

Proposition 2. *If $\mathcal{X}$ and $\mathcal{Y}$ are finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ are ergodic then*

$$S(\mu,\nu) = \lim_{k\to\infty} \frac{1}{k} T_{c_k}(\mu_k,\nu_k) = T_{\overline{c}}(\mu,\nu).$$

The first equality in Proposition 2 was proven in [35] in the special case where $\mathcal{X} = \mathcal{Y}$ and c is a metric; a straightforward extension of their arguments establishes the general case above. The second equality is proven in Section 7 using cyclical monotonicity of optimal couplings. We remark that the arguments in the proof of Proposition 2 do not rely on the finiteness of $\mathcal{X}$ and $\mathcal{Y}$ and may be adapted to the case of Polish spaces and continuous and bounded cost.

Proposition 2 shows that the optimal joining cost may be obtained as a limit of k-step optimal transport costs, and that this limit is equal to the optimal transport cost under the long term average cost function. In this sense, the optimal joining problem seeks couplings that have good long-run average behavior, relative to the single letter cost, over the complete history of the joint process. This is a natural objective when considering optimal transport for stationary processes. Whereas the limiting average cost $\overline{c}$ may be highly irregular as a function, the set of joinings $\mathcal{J}(\mu,\nu)$ has a relatively simple structure (e.g., compactness and convexity) and leads to an optimization problem that is easier to study.

Remark 1. Proposition 2 implies that the optimal joining problem $S(\mu,\nu)$ satisfies Kantorovich duality with respect to $\overline{c}$. In particular,

$$S(\mu,\nu) = \sup_{f,g} \left\{ \int f\, d\mu + \int g\, d\nu : f \oplus g \leq \overline{c} \right\}, \tag{2.4}$$

where the supremum is taken over μ-integrable $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$ and ν-integrable $g : \mathcal{Y}^{\mathbb{N}} \to \mathbb{R}$. The only difference between (2.4) and the standard Kantorovich dual problem is the appearance of $\overline{c}$ instead of c. More details on Kantorovich duality can be found in [83] and [70]. While we do not make use of the dual optimal joining problem in this paper, we expect that it may be of use in future analyses.

Related Work. A special case of the optimal joining cost first appeared in the work of Ornstein under the name $\overline{d}$-distance [62]. For finite alphabets, the $\overline{d}$-distance is equivalent to the optimal joining cost with respect to the discrete metric. Subsequently, [35] generalized the $\overline{d}$-distance to Polish alphabets and metric costs. More recently, [69] considered the same problem for a specific class of examples in which the optimal joining could be determined exactly and extended the problem to random fields. Subsequent work has considered various aspects of stationary optimal transport in the context of dynamical systems such as duality [91], decompositions [90], and the existence of optimal transport maps [45]. Another line of work [42, 43], referred to as ergodic optimization, studies optimization of linear functionals over the set of invariant measures. One may regard the optimal joining problem as a constrained ergodic

optimization problem. Recent work in statistical inference [53, 54, 52] has shown that stationary optimal transport problems arise naturally in the context of fitting dynamical models. In another direction, [60] studied computational aspects of a constrained form of the optimal joining problem for Markov chains. This constrained optimal joining problem was applied to the comparison and alignment of graphs in [61].

3 Estimation of an Optimal Joining and its Expected Cost

In the rest of the paper, we consider the optimal joining problem in a statistical setting, focusing on the problem of estimating an optimal joining and the optimal joining cost of two stationary, ergodic processes from finite sequences of observations. Fix finite sets $\mathcal{X}$ and $\mathcal{Y}$, a single-letter cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, and stationary, ergodic process measures $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ as in the previous section. Let $X = X_1, X_2, ...$ and $Y = Y_1, Y_2, ...$ be two processes, possibly defined on different probability spaces, with distributions μ and ν, respectively, and suppose that we observe $X_1, ..., X_n$ and $Y_1, ..., Y_n$. The task of interest is to estimate an optimal joining $\lambda \in \mathcal{J}_{\min}(\mu, \nu)$ and the optimal joining cost $S(\mu, \nu)$ from the observed sequences. In this section, we define the proposed estimates and state our main result regarding their consistency in the limit $n \to \infty$.

Before describing our proposed estimation scheme in detail, we first provide some intuition. Proposition 2 ensures that the optimal transport cost between the k-dimensional distributions of μ and ν converges to the optimal joining cost as k tends to infinity. Thus when k is large we expect that a good estimate of $k^{-1}T_{c_k}(\mu_k, \nu_k)$ will approximate the optimal joining cost $S(\mu, \nu)$. Furthermore, a stationary coupling achieving the k-step optimal joining cost should be close to the set of optimal joinings $\mathcal{J}_{\min}(\mu, \nu)$. In the setting of interest to us, we do not have access to the finite dimensional distributions of the observed processes; instead we estimate these from the available observations. In what follows, let k be a fixed integer between 1 and n. The choice of k is discussed below in Section 3.2.

We propose an estimation scheme that is comprised of three steps. First, we construct k-block empirical measures from the available observations of each process. The resulting probability measures act as empirical estimates of μ_k and ν_k. Second, we select an optimal transport plan between these empirical k-block measures with respect to c_k. The expected cost of this coupling acts as an estimate of $T_{c_k}(\mu_k, \nu_k)$. Finally, we construct a stationary process measure from the coupling in the second step. This is done via a k-block process construction, described formally below in Definition 2.

Step 1: (Empirical k-block measure) Define the probability measure $\hat{\mu}_{k,n} := \hat{\mu}_k[X_1^n] \in \mathcal{M}(\mathcal{X}^k)$ by

$$\hat{\mu}_{k,n}(x_1^k) = \frac{1}{n-k+1} \sum_{\ell=0}^{n-k} \mathbb{1}(x_1^k = X_{\ell+1}^{\ell+k}),$$

for $x_1^k \in \mathcal{X}^k$. This is referred to as the *k-block empirical measure* constructed from observations X_1^n. Let $\hat{\nu}_{k,n}$ be the k-block empirical measure constructed from Y_1^n in the same manner. Note that $\hat{\mu}_{k,n}$ is a probability measure on $\mathcal{X}^k$, and is an estimate of μ_k, the k-dimensional distribution of μ; analogous remarks apply to $\hat{\nu}_{k,n}$.

Step 2: (Optimal coupling of k-block measures) Find an optimal coupling of $\hat{\mu}_{k,n}$ and $\hat{\nu}_{k,n}$ with respect to c_k. Formally, let $\pi_k \in \Pi(\hat{\mu}_{k,n}, \hat{\nu}_{k,n})$ be any coupling such that $\int c_k \, d\pi_k \leq \int c_k \, d\pi'_k$ for all $\pi'_k \in \Pi(\hat{\mu}_{k,n}, \hat{\nu}_{k,n})$. Thus $\int c_k \, d\pi_k = T_{c_k}(\hat{\mu}_{k,n}, \hat{\nu}_{k,n})$. To simplify notation in what follows, define

$$\hat{S}_{k,n} \;=\; \hat{S}_k(X_1^n, Y_1^n) \;=\; k^{-1}T_{c_k}(\hat{\mu}_{k,n}, \hat{\nu}_{k,n}). \tag{3.1}$$

We regard the number $\hat{S}_{k,n}$ defined in (3.1) as an estimate of the optimal joining cost of the observed processes X and Y. We estimate an optimal joining of these processes by constructing a stationary process measure from the optimal coupling π_k appearing in Step 2 above. In order to do this, we leverage a well-known construction described in Definition 2 below.

Definition 2 (Block process construction)**.** *Let $\mathcal{U}$ be finite and $k \geq 1$. Define $\tilde{\Lambda}^k : \mathcal{M}(\mathcal{U}^k) \to \mathcal{M}(\mathcal{U}^{\mathbb{N}})$ to be the map that takes a probability measure $\gamma \in \mathcal{M}(\mathcal{U}^k)$ to the unique probability measure on $\mathcal{U}^{\mathbb{N}}$ obtained by independently concatenating γ with itself infinitely many times. Formally, for any ℓk-dimensional cylinder set $C = C_1 \times \cdots \times C_{\ell k} \times \mathcal{U} \times \cdots \subset \mathcal{U}^{\mathbb{N}}$,*

$$\begin{aligned}\tilde{\Lambda}^k[\gamma](C) &= \gamma(C_1 \times \cdots \times C_k)\gamma(C_{k+1} \times \cdots \times C_{2k}) \cdots \gamma(C_{(\ell-1)k+1} \times \cdots \times C_{\ell k}) \\ &= \prod_{i=0}^{\ell-1} \gamma(C_{ik+1}^{ik+k}).\end{aligned}$$

Moreover, define $\Lambda^k : \mathcal{M}(\mathcal{U}^k) \to \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ to be the map defined by randomizing the start of the output of $\tilde{\Lambda}^k$ over the first k coordinates. Formally, for any set $A \subset \mathcal{U}^{\mathbb{N}}$,

$$\Lambda^k[\gamma](A) = \frac{1}{k} \sum_{\ell=0}^{k-1} \tilde{\Lambda}^k[\gamma](\mathcal{U}^\ell \times A).$$

We will refer to $\tilde{\Lambda}^k[\gamma]$ as the independent k-block process induced by γ *and $\Lambda^k[\gamma]$ as the* stationary k-block process induced by γ.

The block process construction described above is standard in ergodic theory [64, 74], and ensures that the process $\Lambda^k[\gamma]$ is stationary.

Step 3: (Construct stationary process) Given the k-dimensional measure π_k obtained in Step 2, define the stationary joint measure $\hat{\lambda}^{k,n} = \Lambda^k[\pi_k] \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$.

We regard $\hat{\lambda}^{k,n}$ as an estimate of an optimal joining of the observed processes. We establish in Appendix B that $\hat{\lambda}^{k,n}$ is a joining of empirical estimates $\Lambda^k[\hat{\mu}_{k,n}]$ and $\Lambda^k[\hat{\nu}_{k,n}]$ of μ and ν with expected cost equal to $\hat{S}_{k,n}$.

3.1 Consistency

Having detailed the proposed estimators $\hat{\lambda}^{k,n}$ and $\hat{S}_{k,n}$, we now consider their behavior as the length n of the observed sequences goes to infinity. Intuitively, for fixed k we expect that when n is large, $\hat{S}_{k,n}$ will be close to $k^{-1}T_{c_k}(\mu_k, \nu_k)$. However, as Proposition 2 suggests, this quantity will only be close to the optimal joining cost when k is large. Thus in order for our estimates to converge to the desired targets, we must let k grow with n. In particular, we consider sequences of estimates $\{\hat{\lambda}^{k(n),n}\}_{n\geq 1}$ and $\{\hat{S}_{k(n),n}\}_{n\geq 1}$ for some sequence $k(n)$ such that $k(n) \to \infty$ and ask whether the two sequences converge to an optimal joining and the optimal joining cost, respectively. We show in Theorem 3 that under the stated assumptions, such a sequence $k(n)$ necessarily exists. We will say that a sequence of Borel probability measures $\gamma^1, \gamma^2, \ldots \in \mathcal{M}(\mathcal{U})$ converges weakly to a set $\Gamma \subset \mathcal{M}(\mathcal{U})$, written as $\gamma^n \Rightarrow \Gamma$, if every subsequence of γ^n contains a further subsequence that converges weakly to an element of Γ. Moreover, to simplify notation going forward, we will occasionally write k for $k(n)$ when there is no risk of confusion.

Theorem 3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ be ergodic. Then there exists a sequence $k = k(n)$ with $k(n) \to \infty$ such that $\hat{S}_{k,n} \to S(\mu, \nu)$ and $\hat{\lambda}^{k,n} \Rightarrow \mathcal{J}_{min}(\mu, \nu)$ almost surely as $n \to \infty$.*

The problem of estimating optimal joinings appears to have not been considered explicitly in the literature. However, the special case of estimating the $\overline{d}$-distance between two ergodic processes from finite observations has been considered. The focus of this line of work has been in finding universal estimation schemes, including choices of the sequence $k(n)$, such that the desired convergence holds uniformly over all pairs μ and ν from some set. The estimate $\hat{S}_{k,n}$ for the optimal joining cost that we

propose is an extension of that proposed in [64] for the $\overline{d}$-distance. In the case that $\mathcal{X} = \mathcal{Y}$ and c is the 0-1 cost, it was shown that $\hat{S}_{k,n}$ with $k = \mathcal{O}(\log n)$ converges to the optimal joining cost of μ and ν whenever μ is a stationary coding of an iid process. Later work studied the limits of this estimation scheme [51] and the properties of processes for which the scheme is consistent [63]. In this context, our consistency results allow for relatively weak assumptions on μ and ν (ergodicity) at the expense of loss of control over the sequence $k(n)$.

Remark 2. The arguments underlying the proof of Theorem 3 may be adapted in a straightforward way to show that the proposed estimates are consistent more generally whenever $\mathcal{X}$ and $\mathcal{Y}$ are compact and c is continuous.

3.2 Choice of $k(n)$

Theorem 3 raises the question of how the sequence $k(n)$ depends on the processes X and Y. In the proof, we find that the choice of the sequence $k(n)$ is related to a notion of admissibility, which we now describe.

Definition 3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a cost function. Then the $\mathcal{X}$-adapted cost $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is defined by*

$$c_{\mathcal{X}}(x, x') = \sup_{y \in \mathcal{Y}} |c(x, y) - c(x', y)|,$$

with the $\mathcal{Y}$-adapted cost $c_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ defined in the analogous way.

The adapted cost arises naturally when studying the Lipschitz properties of the optimal transport and optimal joining costs (see Lemmas 13 and 15). Define $c_{\mathcal{X},k} : \mathcal{X}^k \times \mathcal{X}^k \to \mathbb{R}_+$ to be the sum of the $\mathcal{X}$-adapted cost over k coordinates, and define $c_{\mathcal{Y},k}$ similarly. Note that $(\mathcal{X}^k, c_{\mathcal{X},k})$ and $(\mathcal{Y}^k, c_{\mathcal{Y},k})$ are well-defined pseudometric spaces for every $k \geq 1$.

Definition 4. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a cost function. We will say that a nondecreasing sequence $k = k(n)$ with $k \to \infty$ is* c-admissible *for $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ if*

$$\mu\left(\mathbf{x} \in \mathcal{X}^{\mathbb{N}} : \lim_{n\to\infty} \frac{1}{k} T_{c_{\mathcal{X},k}}\left(\hat{\mu}_{k,n}, \mu_k\right) = 0\right) = 1.$$

We define c-admissibility for $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ in the analogous way.

The c-admissibility of a sequence $k = k(n)$ ensures that the average transport distance between the k-dimensional distributions of μ and their empirical counterparts tends to zero almost surely as the number of observations n increases. The property c-admissibility weakens the notions of admissibility (with respect to total variation distance) and $\overline{d}$-admissibility discussed in [74]. We note that certain sequences growing like $\mathcal{O}(\log n)$ are known to be admissible (and thus c-admissible) for aperiodic Markov chains [51].

Proposition 4. *Under the hypotheses of Theorem 3, if a sequence $k = k(n)$ is c-admissible for both μ and ν, then with $\mathbb{P}$-probability one, $\hat{S}_{k,n} \to S(\mu, \nu)$ and $\hat{\lambda}^{k,n} \Rightarrow \mathcal{J}_{min}(\mu, \nu)$ as $n \to \infty$.*

Example 3. Let μ and ν be aperiodic Markov chains with entropy rates $h(\mu)$ and $h(\nu)$, respectively (defined in Section 5). Then for any $\varepsilon > 0$, the sequence $k(n) = (h(\mu) \vee h(\nu) + \varepsilon)^{-1} \log n$ is admissible (and thus c-admissible) for both μ and ν [51]. Thus by Proposition 4, $\hat{S}_{k,n} \to S(\mu, \nu)$ and $\hat{\lambda}^{k,n} \Rightarrow \mathcal{J}_{\text{min}}(\mu, \nu)$ as $n \to \infty$.

Remark 3. Other stationary process constructions may be used in Step 3 of the proposed estimation scheme. For example, one may construct a finite-order, stationary Markov process from π_k. However, the expected cost of a stationary Markov chain constructed from π_k will generally not be equal to $\hat{S}_{k,n}$ and may require more care to control. On the other hand, the approach detailed in Step 3 enables us to control the expected cost of the constructed process $\hat{\lambda}^{k,n}$, namely $\hat{S}_{k,n}$, which we show converges to the optimal joining cost.

Remark 4. The optimal transport distance between k-block measures was proposed as an extension of optimal transport to stationary time series in [58]. However, that work did not consider the relationship of this approach to the optimal joining problem or the consistency of the proposed distance.

4 Finite-Sample Error Bound

In this section, we provide an upper bound on the expected error of the proposed estimate of the optimal joining cost when the observed processes satisfy suitable mixing conditions. Our bounds are derived from bounds on the optimal transport cost between a measure and an estimate of that measure based on a finite number of samples detailed in [8]. A substantial body of work has considered this problem for iid processes from both asymptotic and finite-sample perspectives [24, 8, 29, 87, 55, 33, 44]. Other work has focused on rates of convergence and central limit theorems for the 1-Wasserstein distance (T_d for a metric d) when samples are drawn from a stationary process satisfying a certain mixing condition [21, 8, 22, 7]. In order to obtain an error bound for the estimated optimal joining cost, we consider the case that the marginal processes μ and ν are ϕ-mixing.

Definition 5. *Let $\mathcal{U}$ be finite. A process measure $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ has ϕ-mixing coefficients $\phi_\gamma : \mathbb{N}_0 \to \mathbb{R}_+$ where $\phi_\gamma(0) = 1$, and for any $g \geq 0$,*

$$\phi_\gamma(g+1) = \sup\left\{|\gamma(\mathcal{U}^g \times B|A) - \gamma(B)| : A \subset \mathcal{U}^\ell, \ell \geq 1, B \subset \mathcal{U}^{\mathbb{N}}\right\},$$

*where $\gamma(\mathcal{U}^g \times B|A) = \gamma(A \times \mathcal{U}^g \times B)/\gamma(A)$. The measure γ is called ϕ-*mixing *if $\lim_{g\to\infty} \phi_\gamma(g) = 0$.*

The ϕ-mixing condition is a standard strong mixing condition in the study of stochastic processes. Existing work on estimation of optimal transport costs under dependence has generally focused on weaker mixing conditions such as α-mixing. We find ϕ-mixing to be particularly suited to our arguments in proving an error bound for the estimated optimal joining cost. For more details on ϕ-mixing and its relationship to other strong mixing conditions, we refer the reader to [9].

For a pseudometric space $(\mathcal{U}, d)$, let $\mathcal{N}(\mathcal{U}, d, \varepsilon)$ denote the ε-covering number of $\mathcal{U}$ with respect to the pseudometric d. We now present our finite sample error bound.

Theorem 5. *Let μ and ν have ϕ-mixing coefficients ϕ_μ and ϕ_ν, respectively. Then there exists a constant $C < \infty$ such that for every $n \geq 1$, $k \in \{1, ..., n\}$, $g \geq 0$ and $t \in (0, \frac{1}{4}\|c\|_\infty]$,*

$$\mathbb{E}\left|\hat{S}_{k,n} - S(\mu,\nu)\right| \leq \|c\|_\infty \left(\frac{k(\phi_\mu(g+1) + \phi_\nu(g+1))}{k+g} + \frac{3g}{k}\right) + C\left(t + u_t(k,n) + v_t(k,n)\right),$$

where

$$u_t(k,n) = \left(\frac{1}{n^2}\sum_{\ell=0}^{n}(n-\ell+1)\phi_\mu^{1/2}(\ell)\right)^{1/2} \int_t^{\frac{1}{4}\|c\|_\infty} \mathcal{N}\left(\mathcal{X}^k, \frac{1}{k}c_{\mathcal{X},k}, \varepsilon\right)^{1/2} d\varepsilon,$$

and $v_t(k,n)$ is defined similarly in terms of ϕ_ν and $\mathcal{Y}$.

Theorem 5 gives a general upper bound on the expected error of the estimate of the optimal joining cost in terms of the ϕ-mixing coefficients of μ and ν and the covering numbers of the product spaces $\mathcal{X}^k$ and $\mathcal{Y}^k$ with respect to $c_{\mathcal{X},k}$ and $c_{\mathcal{Y},k}$. When the cost c is less variable the covering numbers under $c_\mathcal{X}$ and $c_\mathcal{Y}$ will be smaller, and the upper bound of the theorem will be smaller as well.

Corollary 6. *Let μ and ν have ϕ-mixing coefficients ϕ_μ and ϕ_ν, respectively, satisfying*

$$\sum_{\ell=0}^{n}(n-\ell)\phi_\mu^{1/2}(\ell) = \mathcal{O}(n^p) \qquad \textit{and} \qquad \sum_{\ell=0}^{n}(n-\ell)\phi_\nu^{1/2}(\ell) = \mathcal{O}(n^p)$$

for some $p \in [1,2)$. Then there exists a constant $C < \infty$ depending only on ϕ_μ and ϕ_ν such that for every $k \geq 1$, $g \geq 0$, and n large enough,

$$\mathbb{E}\left|\hat{S}_{k,n} - S(\mu,\nu)\right| \leq \|c\|_\infty \left(\frac{k(\phi_\mu(g+1)+\phi_\nu(g+1))}{k+g} + \frac{3g}{k} + \frac{C(|\mathcal{X}|^{k/2}+|\mathcal{Y}|^{k/2})}{n^{1-p/2}}\right). \tag{4.1}$$

In particular, if $k(n) < \frac{(2-p)\log n}{\log(|\mathcal{X}|\vee|\mathcal{Y}|)\vee 1}$ and $g(n) = o(k(n))$ with $k(n), g(n) \to \infty$, then the upper bound converges to zero as $n \to \infty$.

Corollary 6 provides finite-sample control of the mean error in the estimated optimal joining cost. In particular, it sheds some light on how the choice of block size k interacts with the amount of dependence of the marginal processes (as quantified by their ϕ-mixing coefficients) and the sample size n. Previous work [15, 78, 77, 32, 10] has established error bounds and rates of convergence for Markov approximations to ergodic processes. However, it appears that no previous work has established such results for the k-block process estimate. We remark that Theorem 5 and Corollary 6 include the special case of Ornstein's $\overline{d}$-distance and thus provide some additional insight into the estimation scheme for this distance proposed in [64]. To provide further context for Corollary 6, we consider two examples below.

Example 4 (IID Processes). If μ and ν are iid processes, then $S(\mu,\nu) = T_c(\mu_1,\nu_1)$. Moreover, $\phi_\mu(g) = \phi_\nu(g) = 0$ for every $g \geq 1$, and so we may let $k = 1$, $g = 0$, and $p = 1$. Then by Corollary 6, we see that

$$\mathbb{E}\left|\hat{S}_{1,n} - S(\mu,\nu)\right| = \mathcal{O}\left(n^{-1/2}\right).$$

The rate above is consistent with known rates for the estimation of the 1-Wasserstein distance on finite spaces [8].

For iid processes, the optimal joining problem reduces to the optimal coupling problem of their 1-dimensional marginal measures. However, when at least one of the measures is not iid, the optimal joining need not be Markov of any order [26], and one must let k tend to infinity in order to estimate the full behavior of an optimal joining. As such, one expects to find slower rates outside the iid setting.

Example 5 (Markov Processes). If μ and ν are aperiodic irreducible Markov chains, then there exist constants $C < \infty$ and $\rho \in (0,1)$ such that $\phi_\mu(g) \leq C\rho^g$ and $\phi_\nu(g) \leq C\rho^g$ [18, 9] for each $g \geq 1$. Thus the summability conditions in Corollary 6 are satisfied with $p = 1$. Applying Corollary 6 with $k(n) = \left\lfloor \frac{\alpha\log(n)}{\log(|\mathcal{X}|\vee|\mathcal{Y}|)\vee 1}\right\rfloor$ and $g(n) = \left\lfloor \frac{\log(\alpha\log(n))}{\log(1/\rho)}\right\rfloor$ for any $\alpha \in (0,1)$ and n large enough, we find

$$\mathbb{E}\left|\hat{S}_{k,n} - S(\mu,\nu)\right| = \mathcal{O}\left(\frac{\log(\log(n))}{\log(n)}\right). \tag{4.2}$$

It was established in [51] that any sequence $k(n)$ growing faster than $\log n$ is inadmissible for every ergodic process with positive entropy rate. Thus the sequence $k(n)$ in Example 5 is the best achievable rate (up to constant factors) for estimating the finite-dimensional distributions of aperiodic irreducible Markov chains in general. However, the rate (4.2) is substantially slower than the polynomial rate typically observed when estimating marginal distributions of Markov chains in, for example, total variation distance [8]. This disparity is due to the g/k term in (4.1), which arises when joining approximations of the processes μ and ν in the proof of Theorem 5.

5 The Entropic Optimal Joining Problem

A large body of recent work in optimal transport has focused on studying the computational and statistical properties of regularized versions of the optimal transport problem. Entropic regularization in particular has attracted a great deal of interest from the machine learning and statistics communities as a

means of smoothing the optimal transport problem and enabling more efficient computation of solutions. For any $\eta > 0$, the *entropic optimal transport problem* is obtained by subtracting the Shannon entropy $H(\pi) = -\sum_{u,v} \pi(u,v) \log \pi(u,v)$ from the optimal transport objective:

$$T_c^\eta(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int c \, d\pi - \eta H(\pi) \right\}.$$

Cuturi [16] showed that solutions to this problem have a special form and can be obtained via a matrix scaling method known as the Sinkhorn-Knopp algorithm [75]. Subsequent work [1, 25, 47, 48, 36] has analyzed the complexity of this and other algorithms for entropic optimal transport in detail, showing that approximations of the optimal transport cost may be obtained in time that is nearly-linear in the dimension of the couplings under consideration. Other work [44, 33, 55, 38] has studied the entropic problem from a statistical perspective, proving estimation error bounds and central limit theorems for the empirical entropic optimal transport cost. In some cases, these estimates exhibit better sample complexity and faster rates of convergence than the best achievable quantities for the unregularized problem, which are known to suffer from the "curse of dimensionality". Recent work has found closed form solutions to the entropic problem between Gaussian distributions [23, 82, 41] and examined the convergence of solutions to solutions of the unregularized problem as the regularization coefficient η converges to zero [67, 5, 59]. For additional details on entropic optimal transport, we refer the reader to [67].

In this section, we consider entropic regularization of the optimal joining problem. We identify a natural penalty term for the optimal joining problem by viewing the regularized problem as a limit of entropic optimal transport problems with increasing dimension. We observe that these problems converge to a regularized optimal joining cost with the *entropy rate* as the penalty term. Entropy rates have been studied in the context of stochastic processes and information theory for many years, dating back to Shannon [73].

Definition 6. *Let $\mathcal{U}$ be finite and let $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ be a stationary measure. For $k \geq 1$ define $H(\gamma_k) = -\sum_{u_1^k} \gamma_k(u_1^k) \log \gamma_k(u_1^k)$. The* entropy rate *of γ is defined by $h(\gamma) := \lim_{k\to\infty} \frac{1}{k} H(\gamma_k)$.*

In other words, the entropy rate is the limiting joint entropy per symbol of the finite dimensional distributions of the process. By subadditivity, the limit in the definition exists and is equal to $\inf_{k\geq 1} \frac{1}{k} H(\gamma_k)$. An iid process with one dimensional distribution p has entropy rate equal to $H(p)$. A stationary, aperiodic, irreducible Markov chain with stationary distribution p and transition matrix P has entropy rate given by $-\sum_{ij} p_i P_{ij} \log P_{ij}$. Occasionally, we will use $H_k(\cdot)$ instead of $H(\cdot)$ to emphasize the dependence on the dimension k. We remark that the entropy rate is known to be weakly upper semicontinuous on finite-alphabet sequence spaces [84]. We make use of this fact in Section 7.5, for example, when establishing the consistency of our entropically regularized estimates, defined below.

Remark 5. The entropy rate of certain processes is preserved under randomization, as in Definition 2. In particular, if a process $\tilde{\gamma}$ is k-stationary for any $k \geq 1$ and γ is the stationary process obtained by randomizing the start of $\tilde{\gamma}$, then $h(\tilde{\gamma}) = h(\gamma)$. A formal statement of this fact along with a proof may be found in Appendix A. As our proposed estimate of an optimal joining is constructed by randomizing the start of a k-stationary process, this fact is needed when considering the entropy rate of the estimate.

For $\eta > 0$, we define the *entropic optimal joining problem* by

$$S_c^\eta(\mu,\nu) \;=\; \inf_{\lambda \in \mathcal{J}(\mu,\nu)} \left\{ \int c \, d\lambda_1 - \eta h(\lambda) \right\}. \tag{5.1}$$

As in the unregularized problem, one can show under the stated assumptions that the infimum in (5.1) is attained. We include a proof of this fact in Appendix C. From now on, we will denote the set of joinings achieving the infimum in (5.1) by $\mathcal{J}^\eta_{\min}(\mu,\nu)$. Moreover, we will drop the cost c in our notation, writing $S^\eta(\cdot,\cdot)$ for $S_c^\eta(\cdot,\cdot)$ when there is no risk of confusion. Note that (5.1) is still well-defined when $\eta = 0$ but in that case, we recover the standard optimal joining problem and thus refer to it by that name.

Proposition 2 shows that the optimal joining cost may be obtained as a limit of k-step optimal transport costs. The next proposition extends this result to the entropic optimal joining problem.

Proposition 7. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$. Then for any $\eta \geq 0$,*

$$\lim_{k\to\infty} \frac{1}{k} T^{\eta}_{c_k}(\mu_k, \nu_k) \;=\; S^{\eta}(\mu, \nu).$$

In other words, the entropic optimal joining cost can be obtained as a limit of the average entropic optimal transport costs. Thus the entropy rate emerges as a natural regularizer for the optimal joining problem. Moreover, the proposition suggests that the approach to estimating an optimal joinings may extend to the regularized setting. In particular, when k is large, a good estimate of $k^{-1}T^{\eta}_{c_k}(\mu_k, \nu_k)$ should be a good estimate of $S^{\eta}(\mu, \nu)$, and existing algorithms for efficient computation of $k^{-1}T^{\eta}_{c_k}(\mu_k, \nu_k)$ will translate to faster estimates of $S^{\eta}(\mu, \nu)$.

Entropic penalization has appeared in some related work. In [52], a regularized optimal joining problem with fiber entropy as the penalty term arises naturally in the context of Bayesian estimation of dynamical models. Another line of work [50, 49] has investigated the role of entropic regularization in the thermodynamic formalism, which consists of an optimization of a linear functional over the set of invariant measures. More recently, [60] proposed an extension of entropic techniques to a constrained optimal transport problem specifically for Markov chains. In that work, the entropy constraint implies a constraint on the entropy rate of the set of joinings in question and leads to improved computational efficiency. Existing work has yet to propose a principled regularization scheme for the optimal joining problem considered here.

Before moving on to the proposed estimation scheme, we consider the stability of the entropic optimal joining cost in η, and in particular, the limiting behavior of $S^{\eta}(\mu, \nu)$ when η tends to zero.

Proposition 8. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$. Then the entropic optimal joining cost satisfies*

$$\lim_{\eta\to 0} S^{\eta}(\mu, \nu) = S(\mu, \nu).$$

Thus the entropic optimal joining cost converges to the unregularized optimal joining cost as the penalty parameter shrinks. This behavior is consistent with analogous results in optimal transport [67].

5.1 Extension of the Estimation Procedure

The proposed estimation scheme (described in Section 3) may be easily extended to the entropic optimal joining problem. One need only consider a modification of Step 2 in which one solves an entropic optimal transport problem for $\eta > 0$:

Step 2': (Entropic optimal coupling) Find an entropic optimal coupling of $\hat{\mu}_{k,n}$ and $\hat{\nu}_{k,n}$. Formally, let $\pi_k \in \Pi(\hat{\mu}_{k,n}, \hat{\nu}_{k,n})$ be any coupling such that $\int c_k \, d\pi_k - \eta H(\pi_k) \leq \int c_k \, d\pi'_k - \eta H(\pi'_k)$ for all $\pi'_k \in \Pi(\hat{\mu}_{k,n}, \hat{\nu}_{k,n})$. Thus, π_k has expected entropic k-step cost equal to $T^{\eta}_{c_k}(\hat{\mu}_{k,n}, \hat{\nu}_{k,n})$. To simplify notation in what follows, define

$$\tilde{S}^{\eta}_{k,n} \;=\; \tilde{S}^{\eta}_{k}(X^n_1, Y^n_1) \;=\; k^{-1}T^{\eta}_{c_k}(\hat{\mu}_{k,n}, \hat{\nu}_{k,n}). \tag{5.2}$$

One may then construct a stationary process from π_k via the block process construction detailed in Definition 2. In particular, define $\tilde{\lambda}^{\eta,k,n} = \Lambda^k[\pi_k] \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$. We propose $\tilde{\lambda}^{\eta,k,n}$ as an estimate of an entropic optimal joining of μ and ν and $\tilde{S}^{\eta}_{k,n}$ as an estimate of the entropic optimal joining cost. Note that we will regard η as a fixed quantity in all but Proposition 8 and thus omit it in our notation, writing $\tilde{\lambda}^{k,n}$ for $\tilde{\lambda}^{\eta,k,n}$ and $\tilde{S}_{k,n}$ for $\tilde{S}^{\eta}_{k,n}$ whenever there is no risk of confusion.

As in the unregularized case, we establish in Appendix B that the estimate $\tilde{\lambda}^{k,n}$ is a joining of empirical estimates of μ and ν with the desired expected entropic cost $\tilde{S}_{k,n}$.

5.2 Consistency and Error Bound

If one is interested in consistency of the proposed estimates as n goes to infinity, reasoning similar to that in Section 3 suggests that on consider sequences $\{\tilde{\lambda}^{k,n}\}_{n\geq 1}$ and $\{\tilde{S}_{k,n}\}_{n\geq 1}$ for appropriate sequences $k(n) \to \infty$. The following result extends Theorems 3 and 5 to the regularized setting.

Theorem 9. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ be ergodic. Then for any $\eta > 0$, there exists a sequence $k = k(n)$ with $k \to \infty$ such that $\tilde{S}_{k,n} \to S^{\eta}(\mu,\nu)$ and $\tilde{\lambda}^{k,n} \Rightarrow \mathcal{J}^{\eta}_{min}(\mu,\nu)$ almost surely as $n \to \infty$.*

The proof of Theorem 9 is similar to that of Theorem 3: we construct a c-admissible sequence $k(n)$ for μ and ν and apply a Lipschitz property of the entropic optimal transport cost to obtain the desired convergence. An analog of Proposition 4 holds for the regularized estimation scheme above. The proofs of Theorem 5 and Corollary 6 can be extended to obtain an error bound for the estimated entropic optimal joining cost.

Theorem 10. *Let μ and ν have ϕ-mixing coefficients ϕ_μ and ϕ_ν, respectively, satisfying*

$$\sum_{\ell=0}^{n}(n-\ell)\phi_\mu^{1/2}(\ell) = \mathcal{O}(n^p) \qquad \text{and} \qquad \sum_{\ell=0}^{n}(n-\ell)\phi_\nu^{1/2}(\ell) = \mathcal{O}(n^p)$$

for some $p \in [1,2)$. Then there exists a constant $C < \infty$ depending only on ϕ_μ and ϕ_ν such that for every $\eta > 0$, $k \geq 1$, $g \geq 0$, and large enough n,

$$\begin{aligned}\mathbb{E}\left|\tilde{S}_{k,n} - S^{\eta}(\mu,\nu)\right| \;\leq\; & \|c\|_\infty(\phi_\mu(g+1)+\phi_\nu(g+1))\frac{k}{k+g} + (3\|c\|_\infty + 2\eta(\log|\mathcal{X}| + \log|\mathcal{Y}|))\frac{g}{k} \\ & + u(k,n)\left(\frac{\|c\|_\infty}{2} + \frac{\eta}{k}\log\left(\frac{|\mathcal{X}|^{3k}}{u(k,n)}\right)\right) + v(k,n)\left(\frac{\|c\|_\infty}{2} + \frac{\eta}{k}\log\left(\frac{|\mathcal{Y}|^{3k}}{v(k,n)}\right)\right),\end{aligned}$$

where $u(k,n) = C|\mathcal{X}|^{k/2}n^{p/2-1}$ and $v(k,n) = C|\mathcal{Y}|^{k/2}n^{p/2-1}$. In particular, if $k(n) < \frac{(2-p)\log n}{\log(|\mathcal{X}|\vee|\mathcal{Y}|)\vee 1}$ and $g(n) = o(k(n))$ with $k(n), g(n) \to \infty$, then the upper bound converges to zero as $n \to \infty$.

The introduction of entropy rate regularization into the optimal joining problem results in an error bound that is strictly worse than that for the unregularized optimal joining problem. This reflects the increased difficulty of estimating an entropic optimal joining, which entails simultaneously learning the finite-dimensional distributions as well as the entropies of the marginal processes from observations. Despite the worse error bound, we find that the sequence $k(n)$ may be chosen in the same way as in the unregularized case. Moreover, for non iid processes (for which $k, g \to \infty$), one finds that the bound is of the same order as in the unregularized case. For example, in the setting of Example 5, we have $\mathbb{E}|\tilde{S}_{k,n} - S^{\eta}(\mu,\nu)| = \mathcal{O}\left(\frac{\log(\log(n))}{\log(n)}\right)$. Thus from an asymptotic perspective, there is no additional price paid for using entropic regularization when estimating the optimal joining cost.

6 Discussion

The extension of optimal transport techniques to stochastic processes is an important problem in statistics and machine learning. In this paper, we presented a step in this direction, considering the case of finite-alphabet, stationary and ergodic processes. We argued that, in this setting, one should consider a constrained form of the optimal transport problem, referred to as the optimal joining problem, in order to account for the long-term dynamics of the processes of interest. Given finite sequences of observations, we proposed estimates of an optimal joining and the optimal joining cost, and we proved that these estimates are consistent in the large sample limit. We presented an upper bound on the expected error of the estimated optimal joining cost in terms of the mixing coefficients of the two processes of interest.

Finally, building upon recent work in optimal transport, we also proposed a regularized problem, the entropic optimal joining problem, and extended the proposed estimation scheme, consistency result, and error bound to this new problem.

This work enables the principled application of optimal transport techniques to data arising as observations from stationary processes. Future work may investigate additional properties and uses of the entropic optimal joining problem. For example, are there conditions under which the entropic optimal joining cost exhibits a faster rate of convergence compared to the unregularized optimal joining cost? Other work may extend our results to the setting of Polish spaces. It was noted in Section 3 that the arguments in the proof of Theorem 3 may be adapted to the case when $\mathcal{X}$ and $\mathcal{Y}$ are compact and c is continuous. However, it is not clear whether the entropic optimal joining is always well-defined in that setting. Moreover, the arguments in the proof of Theorem 5 do not extend easily to continuous spaces, and so further consideration is necessary.

7 Proofs

In this section, we prove our stated results. We begin by proving Proposition 2 from Section 2. Next, we state and prove an inequality for the estimated optimal joining cost that will be used throughout the rest of this section. Then, we prove our main results from Sections 3, 4, and 5 in the order that they appear in the text. A small selection of auxiliary results are proven in the appendix and will be referred to throughout this section.

7.1 Proofs from Section 2

Here we establish the second equality in Proposition 2, which states that the optimal joining cost is equal to the optimal transport cost with respect to the averaged cost $\overline{c}$. We remind the reader that the first equality in Proposition 2 was established in [35]. We begin by showing that solutions to the optimal joining problem are characterized by a *cyclical monotonicity* property.

Definition 7. *For two sets $\mathcal{U}$ and $\mathcal{V}$ and a cost function $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$, a set $C \subset \mathcal{U} \times \mathcal{V}$ is called* c-cyclically monotone *if for every $N \geq 1$ and every sequence $(u^1, v^1), ..., (u^N, v^N) \in C$,*

$$\sum_{\ell=1}^{N} c(u^\ell, v^\ell) \leq \sum_{\ell=1}^{N} c(u^\ell, v^{\ell+1}),$$

with the convention that $v^{N+1} = v^1$. A probability measure γ on $\mathcal{U} \times \mathcal{V}$ is called c-cyclically monotone *if there exists a c-cyclically monotone set $C \subset \mathcal{U} \times \mathcal{V}$ such that $\gamma(C) = 1$.*

The characterization of optimal couplings in terms of cyclical monotonicity has been studied in the optimal transport literature. We require the following result.

Theorem A ([4])**.** *Let $\mathcal{U}$ and $\mathcal{V}$ be Polish, $\mu \in \mathcal{M}(\mathcal{U})$, $\nu \in \mathcal{M}(\mathcal{V})$, and $c : \mathcal{U} \times \mathcal{V} \to [0, \infty)$ be measurable. Then any c-cyclically monotone coupling $\pi \in \Pi(\mu, \nu)$ satisfying $\int c\, d\pi < \infty$ is a solution to $T_c(\mu, \nu)$.*

Under stronger assumptions on c (lower semicontinuity and integrability), one may also establish the reverse implication, namely that any optimal coupling is c-cyclically monotone (see [83]). An analogous result holds for the optimal joining problem.

Lemma 11. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ be ergodic. Then an ergodic joining $\lambda \in \mathcal{J}(\mu, \nu)$ is a solution to $S(\mu, \nu)$ if and only if it is $\overline{c}$-cyclically monotone.*

Proof. The limiting average cost $\overline{c}$ is invariant under the joint left-shift map $\sigma \times \tau$, and therefore by the pointwise ergodic theorem $\int c\, d\lambda_1 = \int \overline{c}\, d\lambda$ for every $\lambda \in \mathcal{J}(\mu, \nu)$. Taking infima, we find that

$S(\mu,\nu) = S_{\overline{c}}(\mu,\nu)$. Let $\lambda \in \mathcal{J}(\mu,\nu)$ be an ergodic $\overline{c}$-cyclically monotone joining. Then λ is a solution to $T_{\overline{c}}(\mu,\nu)$ by Theorem A, and therefore

$$\int c\, d\lambda_1 = \int \overline{c}\, d\lambda = T_{\overline{c}}(\mu,\nu) \leq S_{\overline{c}}(\mu,\nu) = S(\mu,\nu)$$

and it follows that λ is necessarily a solution to $S(\mu,\nu)$.

We now show that any ergodic optimal joining is $\overline{c}$-cyclically monotone. Let $\lambda \in \mathcal{J}_{\min}(\mu,\nu)$ be ergodic. As λ, μ, and ν are ergodic, the pointwise ergodic theorem ensures that

1. There exists a set $D \subset \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ with $\lambda(D) = 1$ on which $\overline{c}$ is constant and equal to $\int c\, d\lambda_1$.
2. There exist sets $E \subset \mathcal{X}^{\mathbb{N}}$ and $F \subset \mathcal{Y}^{\mathbb{N}}$ such that $\mu(E) = \nu(F) = 1$, and for any $\mathbf{x} \in E$, $\mathbf{y} \in F$, the probability measures $\mu^n_{\mathbf{x}} := \frac{1}{n}\sum_{\ell=0}^{n-1} \delta_{\sigma^\ell \mathbf{x}}$ and $\nu^n_{\mathbf{y}} := \frac{1}{n}\sum_{\ell=0}^{n-1} \delta_{\tau^\ell \mathbf{y}}$ satisfy $\mu^n_{\mathbf{x}} \Rightarrow \mu$ and $\nu^n_{\mathbf{y}} \Rightarrow \nu$.

Let $C = D \cap (E \times F)$. Then $\lambda(C) = 1$, so we need only show that C is $\overline{c}$-cyclically monotone. Let $N \geq 1$ and $(\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^N, \mathbf{y}^N) \in C$, and suppose by way of contradiction that

$$\sum_{\ell=1}^{N} \overline{c}(\mathbf{x}^\ell, \mathbf{y}^\ell) > \sum_{\ell=1}^{N} \overline{c}(\mathbf{x}^\ell, \mathbf{y}^{\ell+1}),$$

where we use the convention $\mathbf{y}^{N+1} = \mathbf{y}^1$. Define a sequence of probability measures λ^n on $\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ as follows

$$\lambda^n := \frac{1}{nN} \sum_{\ell=1}^{N} \sum_{k=0}^{n-1} \delta_{(\sigma^k \mathbf{x}^\ell, \tau^k \mathbf{y}^{\ell+1})},$$

For each n the measure λ^n is a coupling of $\overline{\mu}^n = \frac{1}{N}\sum_{\ell=1}^{N} \mu^n_{\mathbf{x}^\ell}$ and $\overline{\nu}^n = \frac{1}{N}\sum_{\ell=1}^{N} \nu^n_{\mathbf{y}^\ell}$ where $\mu^n_{\mathbf{x}^\ell}$ and $\nu^n_{\mathbf{y}^\ell}$ are defined as in the definitions of the events E and F above. The definition of C ensures that $\overline{\mu}^n \Rightarrow \mu$ and $\overline{\nu}^n \Rightarrow \nu$ as $n \to \infty$. Applying Lemma 23 in Appendix E we find that there is a subsequence λ^{n_k} converging weakly to some $\tilde{\lambda} \in \mathcal{J}(\mu,\nu)$. To simplify notation, we drop the subscript and refer to this subsequence as λ^n. Using the fact that $\overline{c}$ is constant on C and that c is continuous and bounded, we have

$$\begin{aligned}
\int c\, d\tilde{\lambda}_1 &= \lim_{n\to\infty} \int c\, d\lambda^n_1 \\
&= \limsup_{n\to\infty} \frac{1}{nN} \sum_{\ell=1}^{N} \sum_{k=1}^{n} c(x^\ell_k, y^{\ell+1}_k) \\
&\leq \frac{1}{N} \sum_{\ell=1}^{N} \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} c(x^\ell_k, y^{\ell+1}_k) \\
&= \frac{1}{N} \sum_{\ell=1}^{N} \overline{c}(\mathbf{x}^\ell, \mathbf{y}^{\ell+1}) \\
&< \frac{1}{N} \sum_{\ell=1}^{N} \overline{c}(\mathbf{x}^\ell, \mathbf{y}^{\ell}) \\
&= \int c\, d\lambda_1 \\
&= S(\mu,\nu),
\end{aligned}$$

a contradiction. Thus C is $\overline{c}$-cyclically monotone and the result follows. □

Proposition 2. *If $\mathcal{X}$ and $\mathcal{Y}$ are finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ are ergodic then*

$$S(\mu, \nu) = \lim_{k\to\infty} \frac{1}{k} T_{c_k}(\mu_k, \nu_k) = T_{\overline{c}}(\mu, \nu).$$

Proof. Under the stated conditions, it is known [74] that there exists an ergodic joining $\lambda \in \mathcal{J}_{\min}(\mu, \nu)$. Lemma 11 ensures that λ is $\overline{c}$-cyclically monotone, and therefore, by Theorem A, λ is a solution to $T_{\overline{c}}(\mu, \nu)$. It follows that $S(\mu, \nu) = \int c\, d\lambda_1 = T_{\overline{c}}(\mu, \nu)$. □

7.2 Preliminary Results

Before proving the main results from Sections 3-5, we establish a bound on the discrepancy between two entropic optimal transport costs in Lemma 13 below. Let $\mathcal{U}$ and $\mathcal{V}$ be finite, $\alpha \in \mathcal{M}(\mathcal{U})$ and $\beta \in \mathcal{M}(\mathcal{V})$, $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$, and $\eta > 0$. Recall that the optimal transport cost satisfies

$$T_c(\alpha, \beta) = \max_{\substack{f:\mathcal{U}\to\mathbb{R} \\ g:\mathcal{V}\to\mathbb{R}}} \left\{ \int f\, d\alpha + \int g\, d\beta : f(u) + g(v) \le c(u, v),\ \forall (u, v) \in \mathcal{U} \times \mathcal{V} \right\}. \tag{7.1}$$

This equivalence is known as Kantorovich duality and is detailed, for example, in [83]. Moreover, it is established in [17, Proposition 2.4] that the entropic optimal transport problem satisfies

$$T_c^\eta(\alpha, \beta) = \max_{f:\mathcal{U}\to\mathbb{R}} \left\{ \int f\, d\alpha + \int \tilde{f}_\beta\, d\beta \right\} = \max_{g:\mathcal{V}\to\mathbb{R}} \left\{ \int \tilde{g}_\alpha\, d\alpha + \int g\, d\beta \right\}, \tag{7.2}$$

where

$$\tilde{f}_\beta(v) = \eta \log \beta(v) - \eta \log \left(\sum_u \exp\left\{ \frac{1}{\eta}(f(u) - c(u, v)) \right\} \right)$$

and

$$\tilde{g}_\alpha(u) = \eta \log \alpha(u) - \eta \log \left(\sum_v \exp\left\{ \frac{1}{\eta}(g(v) - c(u, v)) \right\} \right).$$

The formulation (7.2) is referred to as the semidual of the entropic optimal transport problem while the quantities $\tilde{f}_\beta$ and $\tilde{g}_\alpha$ are referred to as the (c, η)-transforms of f and g with respect to β and α, respectively. In what follows, we will let

$$\tilde{g}(u) = -\eta \log \left(\sum_v \exp\left\{ \frac{1}{\eta}(g(v) - c(u, v)) \right\} \right),$$

to simplify notation. Note that $\tilde{g}_\alpha(u) = \eta \log \alpha(u) + \tilde{g}(u)$. Our proof of Lemma 13 will leverage the duality (7.1) and (7.2) as well as the following basic facts about $\tilde{f}$ and $\tilde{g}$.

Lemma 12. *Let $(\mathcal{U}, d_\mathcal{U})$ and $(\mathcal{V}, d_\mathcal{V})$ be finite pseudometric spaces, and let $f : \mathcal{U} \to \mathbb{R}$ and $g : \mathcal{V} \to \mathbb{R}$ be real-valued functions. Furthermore, let $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$ be a non-negative cost function satisfying $|c(u, v) - c(u', v')| \le M(d_\mathcal{U}(u, u') + d_\mathcal{V}(v, v'))$ for all $u, u' \in \mathcal{U}$ and $v, v' \in \mathcal{V}$ for some $M \in \mathbb{R}$. Then for any $\eta > 0$, $\tilde{f}$ and $\tilde{g}$ satisfy $|\tilde{f}(v) - \tilde{f}(v')| \le M d_\mathcal{V}(v, v')$ and $|\tilde{g}(u) - \tilde{g}(u')| \le M\, d_\mathcal{U}(u, u')$ for all $u, u' \in \mathcal{U}$ and $v, v' \in \mathcal{V}$.*

A proof of Lemma 12 is provided in Appendix D. A detailed discussion of the (c, η)-transform and its use in optimal transport can be found in [67]. Now we may proceed to the result of interest.

Lemma 13. *Let $\mathcal{U}$ and $\mathcal{V}$ be finite and let $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$ be a cost function. Then for any $\eta \ge 0$, any $\alpha, \alpha' \in \mathcal{M}(\mathcal{U})$, and any $\beta \in \mathcal{M}(\mathcal{V})$, it holds that*

$$T_c^\eta(\alpha, \beta) - T_c^\eta(\alpha', \beta) \le T_{c_\mathcal{U}}(\alpha, \alpha') + \eta(H(\alpha') - H(\alpha)). \tag{7.3}$$

The analogous bound for a pair of measures in $\mathcal{M}(\mathcal{V})$ also holds.

Proof. We begin by considering the case $\eta > 0$. Let $\alpha, \alpha' \in \mathcal{M}(\mathcal{U})$ and let $\beta \in \mathcal{M}(\mathcal{V})$. By (7.2) there exists $g : \mathcal{V} \to \mathbb{R}$ be such that

$$T_c^\eta(\alpha, \beta) = \int \tilde{g}_\alpha \, d\alpha + \int g \, d\beta. \tag{7.4}$$

Rewriting $\tilde{g}_\alpha$, we have

$$\begin{aligned} T_c^\eta(\alpha, \beta) &= \int (\tilde{g} + \eta \log \alpha) \, d\alpha + \int g \, d\beta \\ &= \int \tilde{g} \, d\alpha + \int g \, d\beta - \eta H(\alpha). \end{aligned}$$

As g is feasible for the semidual problem of $T_c^\eta(\alpha', \beta)$, (7.2) implies that

$$\begin{aligned} T_c^\eta(\alpha', \beta) &\geq \int \tilde{g}_{\alpha'} \, d\alpha' + \int g \, d\beta \\ &= \int \tilde{g} \, d\alpha' + \int g \, d\beta - \eta H(\alpha'). \end{aligned} \tag{7.5}$$

Combining (7.4) and (7.5), we find that

$$T_c^\eta(\alpha, \beta) - T_c^\eta(\alpha', \beta) \leq \int \tilde{g} \, d\alpha - \int \tilde{g} \, d\alpha' + \eta(H(\alpha') - H(\alpha)).$$

Note that $|c(u, v) - c(\tilde{u}, \tilde{v})| \leq c_{\mathcal{U}}(u, \tilde{u}) + c_{\mathcal{V}}(v, \tilde{v})$, and therefore by Lemma 12, $\tilde{g}$ satisfies $\tilde{g}(u) - \tilde{g}(\tilde{u}) \leq c_{\mathcal{U}}(u, \tilde{u})$. Thus the pair $(\tilde{g}, -\tilde{g})$ is feasible for the dual (7.1) of $T_{c_{\mathcal{U}}}(\alpha, \alpha')$ and it follows that

$$T_c^\eta(\alpha, \beta) - T_c^\eta(\alpha', \beta) \leq T_{c_{\mathcal{U}}}(\alpha, \alpha') + \eta(H(\alpha') - H(\alpha)).$$

Taking the limit as $\eta \to 0$ of (7.3) and applying [17, Proposition 2.1], we obtain the result for $\eta = 0$. □

The next proposition details the implication of Lemma 13 for the k-step entropic optimal transport cost. The proof follows from a straightforward application of Lemma 13 and the pointwise inequalities $(c_k)_{\mathcal{X}^k} \leq c_{\mathcal{X},k}$ and $(c_k)_{\mathcal{Y}^k} \leq c_{\mathcal{Y},k}$.

Proposition 14. *For any $\eta \geq 0$, $n \geq 1$, and $k \in \{1, ..., n\}$,*

$$\begin{aligned} \left| \tilde{S}_k(X_1^n, Y_1^n) - \frac{1}{k} T_{c_k}^\eta(\mu_k, \nu_k) \right| \leq \frac{1}{k} T_{c_{\mathcal{X},k}}(\hat{\mu}_{k,n}, \mu_k) + \frac{\eta}{k} \left| H(\mu_k) - H(\hat{\mu}_{k,n}) \right| \\ + \frac{1}{k} T_{c_{\mathcal{Y},k}}(\hat{\nu}_{k,n}, \nu_k) + \frac{\eta}{k} \left| H(\nu_k) - H(\hat{\nu}_{k,n}) \right|. \end{aligned}$$

7.3 Proofs from Section 3

In this section, we prove Theorem 3 regarding the consistency of the proposed estimates without entropic regularization.

Theorem 3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ be ergodic. Then there exists a sequence $k = k(n)$ with $k(n) \to \infty$ such that $\hat{S}_{k,n} \to S(\mu, \nu)$ and $\hat{\lambda}^{k,n} \Rightarrow \mathcal{J}_{min}(\mu, \nu)$ almost surely as $n \to \infty$.*

Proof. We begin by constructing a sequence $\{k(n)\}$ such that the $k(n)$-step empirical optimal transport cost converges to the optimal joining cost almost surely. As noted by [51], due to the ergodic theorem, μ and ν have admissible sequences $\{\ell(n)\}$ and $\{m(n)\}$. Using the same reasoning, one may verify that $\{k(n)\}$ where $k(n) = \min\{\ell(n), m(n)\}$ is also admissible for both processes. Since any admissible sequence is also c-admissible, $\{k(n)\}$ is c-admissible for both μ and ν. In order to simplify notation in

the rest of the proof, we suppress the dependence of $k(n)$ on n. For any $n \geq 1$, an application of the triangle inequality gives

$$\left|\hat{S}_k(X_1^n, Y_1^n) - S(\mu,\nu)\right| \leq \left|\hat{S}_k(X_1^n, Y_1^n) - \frac{1}{k}T_{c_k}(\mu_k,\nu_k)\right| + \left|\frac{1}{k}T_{c_k}(\mu_k,\nu_k) - S(\mu,\nu)\right|. \tag{7.6}$$

Applying Proposition 14, we have

$$\left|\hat{S}_k(X_1^n, Y_1^n) - \frac{1}{k}T_{c_k}(\mu_k,\nu_k)\right| \leq \frac{1}{k}T_{c_{\mathcal{X},k}}(\hat{\mu}_k[X_1^n],\mu_k) + \frac{1}{k}T_{c_{\mathcal{Y},k}}(\hat{\nu}_k[Y_1^n],\nu_k),$$

which by the c-admissibility of k for μ and ν implies that the first term on the right hand side in (7.6) goes to zero, almost surely as $n \to \infty$. An application of Proposition 2 with the fact that $k(n) \to \infty$ shows that the second term on the right hand side in (7.6) goes to zero. It follows that

$$\left|\hat{S}_k(X_1^n, Y_1^n) - S(\mu,\nu)\right| \to 0, \quad \text{almost surely.} \tag{7.7}$$

Next we show that the sequence of estimated optimal joinings indexed by k converges weakly to the set of optimal joinings $\mathcal{J}_{\min}(\mu,\nu)$, almost surely. Fix sequences $X = X_1, X_2, ...$ and $Y = Y_1, Y_2, ...$ in sets of μ- and ν-measure one on which (7.7) holds. Let $\{\hat{\lambda}^{k,n}\}_{n\geq 1}$ be the corresponding sequence of estimated optimal joinings. By Lemma 23 in Appendix E, for any subsequence $\{\hat{\lambda}^{k,n_\ell}\}_{\ell\geq 1}$, there is a further subsequence converging weakly to a joining $\lambda \in \mathcal{J}(\mu,\nu)$. For ease of notation, we refer to this further subsequence again as $\{\hat{\lambda}^{k,n_\ell}\}_{\ell\geq 1}$. Then

$$\int c\, d\lambda_1 = \lim_{\ell\to\infty} \int c\, d\hat{\lambda}_1^{k,n_\ell} = \lim_{\ell\to\infty} \hat{S}_k(X_1^{n_\ell}, Y_1^{n_\ell}) = S(\mu,\nu),$$

where the first equality follows from the continuity and boundedness of c, the second equality follows from Proposition 21 in Appendix B, and the third equality follows from (7.7). Thus, $\lambda \in \mathcal{J}_{\min}(\mu,\nu)$ and since the subsequence was arbitrary, we conclude that $\hat{\lambda}^{k,n_\ell} \Rightarrow \mathcal{J}_{\min}(\mu,\nu)$. By the choice of sequences X and Y, this convergence occurs almost surely. □

7.4 Proofs from Section 4

In this section, we prove Theorem 5 regarding the expected error of the estimated optimal joining cost $\hat{S}_{k,n}$. Our argument may be broken down into three steps. First, we prove a Lipschitz result for the optimal joining cost akin to Lemma 13 in terms of Ornstein's $\overline{d}$-distance. Second, we prove a novel upper bound on these $\overline{d}$ terms using the ϕ-mixing coefficients of the process measures μ and ν. Finally, we use a covering number bound to control the error of the estimated k-step optimal transport cost.

Bound on Optimal Joining Cost Discrepancy. To begin, we establish an inequality for the difference between the entropic optimal joining costs of two pairs of processes akin to the result stated in Lemma 13. Since we will require the bound for the proof of Theorem 9 as well, we prove Lemma 15 more generally for the regularized optimal joining cost $S^\eta(\mu,\nu)$. Briefly, we remind the reader that the $\overline{d}$-distance between two processes, introduced in [62], may be defined as the optimal joining cost with respect to the single-letter Hamming metric $(\mathbf{u},\mathbf{u}') \mapsto \mathbb{1}(u_1 \neq u_1')$. The distance $\overline{d}$ may be thought of as the process analogue to the total variation distance.

Lemma 15. *Let $\alpha, \alpha' \in \mathcal{M}_s(\mathcal{X}^\mathbb{N})$ be stationary process measures. Then for any $\eta \geq 0$ and $\beta \in \mathcal{M}_s(\mathcal{Y}^\mathbb{N})$,*

$$S^\eta(\alpha,\beta) - S^\eta(\alpha',\beta) \leq \|c\|_\infty \overline{d}(\alpha,\alpha') + \eta(h(\alpha') - h(\alpha)).$$

The analogous bound for stationary process measures in $\mathcal{M}_s(\mathcal{Y}^\mathbb{N})$ also holds.

Proof. Fix $k \geq 1$, $\eta \geq 0$, α, $\alpha' \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$, and $\beta \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$. Recall that by Lemma 13,

$$T^{\eta}_{c_k}(\alpha_k, \beta_k) - T^{\eta}_{c_k}\left(\alpha'_k, \beta_k\right) \leq T_{(c_k)_{\mathcal{X}^k}}(\alpha_k, \alpha'_k) + \eta(H(\alpha'_k) - H(\alpha_k)).$$

Let $\delta_k(x_1^k, \tilde{x}_1^k) = \sum_{\ell=1}^{k} \mathbb{1}(x_\ell \neq \tilde{x}_\ell)$ be the k-step Hamming distance and note that the pointwise upper bound $(c_k)_{\mathcal{X}^k} \leq \|c\|_\infty \delta_k$ holds. Thus,

$$T^{\eta}_{c_k}(\alpha_k, \beta_k) - T^{\eta}_{c_k}\left(\alpha'_k, \beta_k\right) \leq \|c\|_\infty T_{\delta_k}(\alpha_k, \alpha'_k) + \eta(H(\alpha'_k) - H(\alpha_k)).$$

Dividing by k, letting $k \to \infty$, and applying Proposition 7, we find

$$S^{\eta}(\alpha, \beta) - S^{\eta}(\alpha', \beta) \leq \|c\|_\infty S_\delta(\alpha, \alpha') + \eta(h(\alpha') - h(\alpha)).$$

Recognizing that $S_\delta(\alpha, \alpha') = \overline{d}(\alpha, \alpha')$, the result follows. □

Bound on $\overline{d}$. Next, we prove an upper bound on the $\overline{d}$-distance between a stationary process measure and an approximation constructed from its finite dimensional distributions. The approximation of interest is defined as follows:

Definition 8 (Block approximation with gaps)**.** *Let $\mathcal{U}$ be a finite space and $k, g \geq 1$. We define $\tilde{\Lambda}^k : \mathcal{M}_s(\mathcal{U}^{\mathbb{N}}) \times \mathcal{M}(\mathcal{U}^g) \to \mathcal{M}(\mathcal{U}^{\mathbb{N}})$ to be the map that takes a process $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ and a probability measure $\alpha \in \mathcal{M}(\mathcal{U}^g)$ to the unique probability measure on $\mathcal{U}^{\mathbb{N}}$ obtained by independently concatenating γ_k and α together infinitely many times. Formally, for any $\ell(k+g)$-dimensional cylinder set $C \subset \mathcal{U}^{\mathbb{N}}$,*

$$\tilde{\Lambda}^k[\gamma, \alpha](C) = \prod_{i=0}^{\ell-1} \gamma_k(C_{i(k+g)+1}^{i(k+g)+k}) \alpha(C_{i(k+g)+k+1}^{(i+1)(k+g)}).$$

Moreover, we define $\Lambda^k : \mathcal{M}_s(\mathcal{U}^{\mathbb{N}}) \times \mathcal{M}(\mathcal{U}^g) \to \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ to be the map defined by randomizing the start of the output of $\tilde{\Lambda}^k$ over the first $k+g$ coordinates. Formally, for any set $U \subset \mathcal{U}^{\mathbb{N}}$,

$$\Lambda^k[\gamma, \alpha](U) = \frac{1}{k+g} \sum_{\ell=0}^{k+g-1} \tilde{\Lambda}^k[\gamma, \alpha](\mathcal{U}^\ell \times U).$$

We will refer to $\tilde{\Lambda}^k[\gamma, \alpha]$ as the independent k-block process approximation of γ with gap *g and $\Lambda^k[\gamma, \alpha]$ as the* stationary k-block process approximation of γ with gap *g.*

Note the abuse of notation for the block approximation with gaps compared to the block approximations without gaps. In the rest of the paper, it will be understood that we omit gaps when $\tilde{\Lambda}^k$ or Λ^k take one argument and include gaps when either takes two arguments. Note also that we omit α when referring to either approximation because our arguments do not depend on the choice of α. For simplicity, we will use the same notation for these approximations regardless of the alphabet of the processes under consideration. As such, $\Lambda^k[\mu, \alpha]$ and $\Lambda^k[\nu, \beta]$ are well-defined. We will show later that $\Lambda^k[\mu, \alpha]$ and $\Lambda^k[\nu, \beta]$ arise naturally in the proof of Theorem 5. In particular, it will be necessary to control the error of these approximations as measured by the $\overline{d}$-distances to μ and ν, respectively.

Lemma 16. *Let $\mathcal{U}$ be finite and $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ have ϕ-mixing coefficient ϕ_γ. Then for every $k \geq 1$, $g \geq 0$ and $\alpha \in \mathcal{M}(\mathcal{U}^g)$,*

$$\overline{d}(\gamma, \Lambda^k[\gamma, \alpha]) \leq \frac{g}{k+g} + \frac{k}{k+g}\phi_\gamma(g+1).$$

Proof. Fix $k \geq 1$, $g \geq 0$, $\alpha \in \mathcal{M}(\mathcal{U}^g)$. To simplify notation, let $\tilde{\xi} = \tilde{\Lambda}^k[\gamma, \alpha]$ and $\xi = \Lambda^k[\gamma, \alpha]$. We begin by defining an intermediate process $\zeta \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$. Let $\tilde{\zeta} \in \mathcal{M}(\mathcal{U}^{\mathbb{N}})$ be the probability measure corresponding to the distribution of the process $V = V_1, V_2, \ldots$ generated by drawing a sequence

$U = U_1, U_2, ...$ according to γ and replacing the g-blocks of U with independent draws $(G_1^i, ..., G_g^i)$ from α to obtain

$$V = \underbrace{U_1, \ldots, U_k}_{k-\text{block 1}}, \underbrace{G_1^1, \ldots, G_g^1}_{g-\text{block 1}}, \underbrace{U_{k+g+1}, \ldots, U_{2k+g}}_{k-\text{block 2}}, \underbrace{G_1^2, \ldots, G_g^2}_{g-\text{block 2}}, \ldots. \tag{7.8}$$

Then, let $\zeta \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ be the stationary process measure obtained by randomizing the start of $\tilde{\zeta}$ over the first $k+g$ coordinates. By the triangle inequality for $\overline{d}$,

$$\overline{d}(\gamma, \xi) \leq \overline{d}(\gamma, \zeta) + \overline{d}(\zeta, \xi).$$

Since the $\overline{d}$-distance is defined as an infimum over joinings, we may upper bound both terms on the right hand side by the expected cost of some suitably chosen joinings of γ and ζ, and ζ and ξ, respectively.

We will first bound $\overline{d}(\gamma, \zeta)$ by constructing a coupling of γ and $\tilde{\zeta}$ with low expected cost and then randomizing the start to obtain a joining of γ and ζ. Since the k-blocks of $\tilde{\zeta}$ are equal in distribution to those of γ by construction, we may couple them so that the k-blocks are equal with probability one. Formally, define the coupling $\pi \in \Pi(\gamma, \tilde{\zeta})$ to be the probability measure corresponding to the distribution of the process (U', V') generated by drawing the sequence $U' = U_1', U_2', \ldots$ according to γ and letting $V' = U_1', \ldots, U_k', G_1^1, \ldots G_g^1, U_{k+g+1}', \ldots$ as in (7.8), replacing the g-blocks of U' with independent draws $(G_1^i, \ldots, G_g^i)$ from α. In particular, $U_\ell' = V_\ell'$ with π-probability one when $\ell = i(k+g) + j$ for some $i \geq 0$ and $j \in \{1, ..., k\}$. Letting $\lambda \in \mathcal{J}(\gamma, \zeta)$ be the joining obtained by randomizing the start of π over the first $k+g$ coordinates, we obtain

$$\begin{aligned}
\overline{d}(\gamma, \zeta) &\leq \int \mathbb{1}(u \neq v)\, d\lambda_1(u, v) \\
&= \int \left(\frac{1}{k+g} \sum_{\ell=1}^{k+g} \mathbb{1}(u_\ell \neq v_\ell) \right) d\pi_{k+g}(u_1^{k+g}, v_1^{k+g}) \\
&= \int \left(\frac{1}{k+g} \sum_{\ell=k+1}^{k+g} \mathbb{1}(u_\ell \neq v_\ell) \right) d\alpha(u_{k+1}^{k+g}, v_{k+1}^{k+g}) \\
&\leq \frac{g}{k+g}.
\end{aligned}$$

Next we bound $\overline{d}(\zeta, \xi)$. By Proposition 2,

$$\overline{d}(\zeta, \xi) = \lim_{m \to \infty} \frac{1}{m} T_{\delta_m}(\zeta_m, \xi_m)$$

and thus, fixing a subsequence $m(\ell) := \ell(k+g) + k$ for $\ell \in \mathbb{N}_0$, we have

$$\overline{d}(\zeta, \xi) = \lim_{L \to \infty} \frac{1}{m(L)} T_{\delta_{m(L)}}\left(\zeta_{m(L)}, \xi_{m(L)}\right). \tag{7.9}$$

It suffices to obtain a bound on

$$\frac{1}{m(L)} T_{\delta_{m(L)}}\left(\zeta_{m(L)}, \xi_{m(L)}\right)$$

for fixed $L \in \mathbb{N}$ and take a limit as $L \to \infty$. Similar to the first bound, we will achieve this by constructing a coupling of $\tilde{\zeta}_{m(L)}$ and $\tilde{\xi}_{m(L)}$ with low expected cost and randomizing the start to obtain a coupling of $\zeta_{m(L)}$ and $\xi_{m(L)}$. Fix $L \in \mathbb{N}$ and recall that both $\tilde{\zeta}$ and $\tilde{\xi}$ are comprised of alternating blocks of size k and g, with the difference between the two measures being that the k-blocks of $\tilde{\zeta}$ depend upon one another while those of $\tilde{\xi}$ are independent of one another. In order to obtain the desired bound, we will bridge the gap between $\tilde{\zeta}_{m(L)}$ and $\tilde{\xi}_{m(L)}$ with a series of intermediate process measures

$\rho^0, ..., \rho^L \in \mathcal{M}(\mathcal{U}^{\mathbb{N}})$ where the first $\ell+1$ k-blocks of ρ^ℓ are dependent on one another (as in $\tilde{\zeta}_{m(L)}$) and the rest are independent (as in $\tilde{\xi}_{m(L)}$). In other words, ρ^ℓ describes the distribution of the process

$$\underbrace{U_1,\dots,U_k}_{k-\text{block } 1},\underbrace{G_1^1,\dots,G_g^1}_{g-\text{block}},\dots,\underbrace{U_{m(\ell)-k+1},\dots,U_{m(\ell)}}_{k-\text{block } \ell+1},\underbrace{\dots}_{g-\text{block}},\underbrace{\tilde{U}_1^1,\dots,\tilde{U}_k^1}_{k-\text{block } \ell+2},\underbrace{\dots}_{g-\text{block}},\underbrace{\tilde{U}_1^2,\dots,\tilde{U}_k^2}_{k-\text{block } \ell+3},\dots$$

where the sequence $U_1, U_2, \dots$ is drawn according to γ and for each $i \geq 1$, $(G_1^i, \dots, G_g^i)$ and $(\tilde{U}_1^i, ..., \tilde{U}_k^i)$ are independent draws from α and γ_k, respectively. In other words, ρ^ℓ is equal to $\tilde{\zeta}$ on the first $\ell+1$ $(k+g)$-blocks and equal to $\tilde{\xi}$ on the remaining blocks. Note that $\rho^0 = \tilde{\xi}$ and $\rho^L_{m(L)} = \tilde{\zeta}_{m(L)}$. Applying the triangle inequality for the optimal transport cost,

$$T_{\delta_{m(L)}}\left(\tilde{\zeta}_{m(L)}, \tilde{\xi}_{m(L)}\right) \leq \sum_{\ell=0}^{L-1} T_{\delta_{m(L)}}\left(\rho^\ell_{m(L)}, \rho^{\ell+1}_{m(L)}\right). \tag{7.10}$$

In order to bound the terms on the right hand side of (7.10), we will couple each pair $\rho^\ell_{m(L)}$ and $\rho^{\ell+1}_{m(L)}$ so that they are equal on the first $\ell+1$ $(k+g)$-blocks, close on the next k-block, and equal again on the remaining k- and g-blocks. Fix $\ell \in \{0, ..., L-1\}$ and consider the coupling of $\rho^\ell_{m(L)}$ and $\rho^{\ell+1}_{m(L)}$ corresponding to the distribution of the paired process $(U', V') = (U'_1, V'_1), \dots, (U'_{m(L)}, V'_{m(L)})$, defined as

$$\begin{aligned} U' &= U_1,\dots,U_k,\ \dots\ ,\dots,U_{m(\ell)-k+1},\dots,U_{m(\ell)},\ \dots,\ \tilde{U}'_1,\ \dots,\tilde{U}'_k,\ \dots,\ \tilde{U}^1_1,\dots,\tilde{U}^1_k,\dots \\ V' &= \underbrace{U_1,\dots,U_k}_{k-\text{block } 1},\underbrace{\dots}_{g-\text{block}},\dots,\underbrace{U_{m(\ell)-k+1},\dots,U_{m(\ell)}}_{k-\text{block } \ell+1},\underbrace{\dots}_{g-\text{block}},\underbrace{\tilde{V}'_1,\dots,\tilde{V}'_k}_{k-\text{block } \ell+2},\underbrace{\dots}_{g-\text{block}},\underbrace{\tilde{U}^1_1,\dots,\tilde{U}^1_k}_{k-\text{block } \ell+3},\dots \end{aligned} \tag{7.11}$$

with elements defined as follows: The sequence $U_1, U_2, \cdots$ is drawn according to γ. The g-blocks of U' and V', which are omitted from (7.11), are equal and drawn independently according to α. The k-blocks $(\tilde{U}^i_1, \dots, \tilde{U}^i_k)$ are drawn independently of one another according to the k-dimensional distribution γ_k of γ. And the paired sequence $(\tilde{U}'_1, \tilde{V}'_1)$, $\dots$, $(\tilde{U}'_k, \tilde{V}'_k)$ is drawn according to an optimal coupling of γ_k and $\gamma_k(\cdot|[U_1^{m(\ell)}]_k)$ with respect to δ_k, where $\gamma_k(\cdot|[U_1^{m(\ell)}]_k) \in \mathcal{M}(\mathcal{U}^k)$ is the distribution of $U_{m(\ell+1)-k+1}, ..., U_{m(\ell+1)}$ conditioned on the previous k-blocks $(U_1, \dots, U_k), \dots, (U_{m(\ell)-k+1}, \dots, U_{m(\ell)})$.

It is clear from (7.11) that the coupling (U', V') only incurs a cost at k-block $\ell+2$. By construction, this cost is equal to the optimal transport cost of γ_k and $\gamma_k(\cdot|[U_1^{m(\ell)}]_k)$ with respect to δ_k. Thus

$$T_{\delta_{m(L)}}\left(\rho^\ell_{m(L)}, \rho^{\ell+1}_{m(L)}\right) \leq \mathbb{E}\left[\delta_{m(L)}(U', V')\right] = \mathbb{E}\left[T_{\delta_k}\left(\gamma_k, \gamma_k(\cdot|[U_1^{m(\ell)}]_k)\right)\right].$$

Finally, using the fact that for any (u_1^k, v_1^k), $\delta_k(u_1^k, v_1^k) \leq k\delta(u_1^k, v_1^k)$, we have

$$T_{\delta_{m(L)}}\left(\rho^\ell_{m(L)}, \rho^{\ell+1}_{m(L)}\right) \leq k\mathbb{E}\left[T_\delta\left(\gamma_k, \gamma_k(\cdot|[U_1^{m(\ell)}]_k)\right)\right] \leq k \max_{u_1^{m(\ell)}} T_\delta\left(\gamma_k, \gamma_k(\cdot|[u_1^{m(\ell)}]_k)\right).$$

Finally, using the fact that the optimal transport cost with respect to δ is equal to the total variation distance, we have

$$T_{\delta_{m(L)}}\left(\rho^\ell_{m(L)}, \rho^{\ell+1}_{m(L)}\right) \leq k \max_{u_1^{m(\ell)}} \max_{A \in \mathcal{U}^k} \left|\gamma_k(A|[u_1^{m(\ell)}]_k) - \gamma_k(A)\right|.$$

Letting $\phi_\gamma : \mathbb{N} \to \mathbb{R}_+$ be the mixing coefficient of γ, it follows that

$$T_{\delta_{m(L)}}\left(\rho^\ell_{m(L)}, \rho^{\ell+1}_{m(L)}\right) \leq k\phi_\gamma(g+1).$$

Plugging this result into (7.10),

$$T_{\delta_{m(L)}}\left(\tilde{\zeta}_{m(L)}, \tilde{\xi}_{m(L)}\right) \leq \sum_{\ell=0}^{L-1} k\phi_\gamma(g+1) = Lk\phi_\gamma(g+1).$$

By randomizing the start of the couplings considered above, one may further establish that

$$T_{\delta_{m(L)}}\left(\zeta_{m(L)}, \xi_{m(L)}\right) \leq Lk\phi_\gamma(g+1).$$

Plugging this into (7.9) and recalling that $m(L) = L(k+g)+k$, we find that

$$\overline{d}(\zeta,\xi) = \lim_{L\to\infty} \frac{Lk}{L(k+g)+k}\phi_\gamma(g+1) = \frac{k}{k+g}\phi_\gamma(g+1).$$

Combining this and the earlier bound yields the result. □

Bound on the Mean k-step Optimal Transport Cost. In the final step before proving Theorem 5, we prove an upper bound on the mean optimal transport cost between $\hat{\mu}_{k,n}$ and μ_k in terms of ϕ_μ (and the analogous result for ν). In order to do this, we leverage [8, Proposition 1.7], stated below as Theorem B, regarding the expectation of the p-Wasserstein distance from an empirical measure to its target measure for stationary, ρ-mixing sequences. We will say that a process measure $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ has ρ-mixing coefficient $\rho_\gamma : \mathbb{N}_0 \to \mathbb{R}_+$ if $\rho_\gamma(0) = 1$ and for $g > 1$ and any random variable $U = (U_1, U_2, ...) : \Omega \to \mathcal{U}^{\mathbb{N}}$ distributed according to γ,

$$\rho_\gamma(g) := \sup\left\{|\text{Corr}(F,G)| : \ell \geq 1, F \in \mathcal{L}^2(\sigma(U_1,...,U_\ell)), G \in \mathcal{L}^2(\sigma(U_{\ell+g},...))\right\},$$

where for $i \leq j \leq \infty$, $\sigma(U_i,...,U_j)$ is the smallest sigma field in $(\Omega, \mathcal{B}, \mathbb{P})$ with respect to which U_i^j is measurable and for a sigma field $\mathcal{F} \subset \mathcal{B}$, $\mathcal{L}^2(\mathcal{F})$ is the set of square-integrable, $\mathcal{F}$-measurable random variables. The result is stated below in a form that is adapted to our notation and the case of $p = 1$.

Theorem B ([8])**.** *Let $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ be a stationary process measure on a Polish space $\mathcal{U}$ with metric d and let γ have ρ-mixing coefficient ρ_γ. Define $\chi_n = n^{-2}\sum_{m=0}^n \sum_{g=0}^m \rho_\gamma(g)$ and let $\Delta = \text{Diam}(\mathcal{U})$. If $\gamma_1^n := \gamma_1^n[U_1^n] \in \mathcal{M}(\mathcal{U})$ is the empirical measure constructed from samples U_1^n drawn according to γ, then there exists a constant $C < \infty$ such that for any $t \in (0, \Delta/4]$,*

$$\mathbb{E}T_d(\gamma_1^n, \gamma_1) \leq C\left(t + \chi_n^{1/2}\int_t^{\frac{1}{4}\Delta} \mathcal{N}(\mathcal{U}, d, \varepsilon)^{1/2}\, d\varepsilon\right).$$

As we show in the next proposition, we may translate this result into an upper bound on the expectation of the adapted optimal transport costs between $\hat{\mu}_{k,n}$ and μ_k, and $\hat{\nu}_{k,n}$ and ν_k under a ϕ-mixing assumption.

Proposition 17. *Let μ and ν have ϕ-mixing coefficients ϕ_μ and ϕ_ν, respectively. Then, there exists a constant $C < \infty$ such that for any $n \geq 1$, $k \in \{1,...,n\}$, and $t \in (0, \frac{1}{4}\|c\|_\infty]$,*

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{k}T_{c_{\mathcal{X},k}}(\hat{\mu}_{k,n}, \mu_k) + \frac{1}{k}T_{c_{\mathcal{Y},k}}(\hat{\nu}_{k,n}, \nu_k)\right] \\
&\quad \leq C\left(t + \left(\frac{1}{n^2}\sum_{g=0}^n (n-g+1)\phi_\mu^{1/2}(g)\right)^{1/2} \int_t^{\frac{1}{4}\|c\|_\infty} \mathcal{N}\left(\mathcal{X}^k, \frac{1}{k}c_{\mathcal{X},k}, \varepsilon\right)^{1/2} d\varepsilon\right. \\
&\quad\quad \left. + \left(\frac{1}{n^2}\sum_{\ell=0}^n (n-g+1)\phi_\nu^{1/2}(g)\right)^{1/2} \int_t^{\frac{1}{4}\|c\|_\infty} \mathcal{N}\left(\mathcal{Y}^k, \frac{1}{k}c_{\mathcal{Y},k}, \varepsilon\right)^{1/2} d\varepsilon\right).
\end{aligned}$$

Proof. The result follows from two applications of Theorem B for μ and ν. Considering first the case of μ, let $\mathcal{U} = \mathcal{X}^k$ with pseudo-metric $d = \frac{1}{k}c_{\mathcal{X},k}$. Moreover, let $\tilde{\mu} \in \mathcal{M}_s((\mathcal{X}^k)^{\mathbb{N}})$ be the distribution of the stationary process $(X_1, ..., X_k), (X_2, ..., X_{k+1}), ...$ Note that the ρ- and ϕ-mixing coefficients of $\tilde{\mu}$ satisfy $\rho_{\tilde{\mu}}(g) \leq 2\phi_{\tilde{\mu}}^{1/2}(g)$ for every $g \geq 0$ [9]. One may also easily establish that $\phi_{\tilde{\mu}}(g) \leq \phi_\mu(g)$ for every $g \geq 0$. Then a direct application of Theorem B to $\tilde{\mu}$ yields

$$\mathbb{E}\left[\frac{1}{k}T_{c_{\mathcal{X},k}}(\hat{\mu}_{k,n}, \mu_k)\right] \leq C\left(t + \left(\frac{2}{n^2}\sum_{m=0}^{n}\sum_{g=0}^{m}\phi_\mu^{1/2}(g)\right)^{1/2}\int_t^{\frac{1}{4}\Delta}\mathcal{N}\left(\mathcal{X}^k, \frac{1}{k}c_{\mathcal{X},k}, \varepsilon\right)^{1/2}d\varepsilon\right)$$
$$= C\left(t + \left(\frac{2}{n^2}\sum_{g=0}^{n}(n-g+1)\phi_\mu^{1/2}(g)\right)^{1/2}\int_t^{\frac{1}{4}\Delta}\mathcal{N}\left(\mathcal{X}^k, \frac{1}{k}c_{\mathcal{X},k}, \varepsilon\right)^{1/2}d\varepsilon\right),$$

for some constant $C < \infty$ and any $t \in (0, \frac{1}{4}\Delta]$, where the $n - k + 1$ term comes from the fact that there are $n - k + 1$ k-blocks in the sequence X_1^n. In this case $\Delta = \|c\|_\infty$ and an identical argument for ν yields the result. □

Proof of Main Results. Gathering the results proven above, we may proceed with the proofs of Theorem 5 and Corollary 6.

Theorem 5. *Let μ and ν have ϕ-mixing coefficients ϕ_μ and ϕ_ν, respectively. Then there exists a constant $C < \infty$ such that for every $n \geq 1$, $k \in \{1, ..., n\}$, $g \geq 0$ and $t \in (0, \frac{1}{4}\|c\|_\infty]$,*

$$\mathbb{E}\left|\hat{S}_{k,n} - S(\mu,\nu)\right| \leq \|c\|_\infty\left(\frac{k(\phi_\mu(g+1) + \phi_\nu(g+1))}{k+g} + \frac{3g}{k}\right) + C\left(t + u_t(k,n) + v_t(k,n)\right),$$

where

$$u_t(k,n) = \left(\frac{1}{n^2}\sum_{\ell=0}^{n}(n-\ell+1)\phi_\mu^{1/2}(\ell)\right)^{1/2}\int_t^{\frac{1}{4}\|c\|_\infty}\mathcal{N}\left(\mathcal{X}^k, \frac{1}{k}c_{\mathcal{X},k}, \varepsilon\right)^{1/2}d\varepsilon,$$

and $v_t(k,n)$ is defined similarly in terms of ϕ_ν and $\mathcal{Y}$.

Proof. Let n, k and g be as in the statement of the theorem. By the triangle inequality,

$$\left|\hat{S}_k(X_1^n, Y_1^n) - S(\mu,\nu)\right| \leq \left|\hat{S}_k(X_1^n, Y_1^n) - \frac{1}{k}T_{c_k}(\mu_k,\nu_k)\right| + \left|\frac{1}{k}T_{c_k}(\mu_k,\nu_k) - S(\mu,\nu)\right|.$$

We begin by establishing an upper bound on $\left|\frac{1}{k}T_{c_k}(\mu_k,\nu_k) - S(\mu,\nu)\right|$. Note by Proposition 2 that $\frac{1}{k}T_{c_k}(\mu_k,\nu_k) \leq S(\mu,\nu)$, so it suffices to upper bound $S(\mu,\nu) - \frac{1}{k}T_{c_k}(\mu_k,\nu_k)$. Let $\pi \in \Pi(\mu_k,\nu_k)$ achieve the minimum in the problem $T_{c_k}(\mu_k,\nu_k)$ and let $\gamma \in \mathcal{M}(\mathcal{X}^g \times \mathcal{Y}^g)$ be a probability measure with marginals $\alpha \in \mathcal{M}(\mathcal{X}^g)$ and $\beta \in \mathcal{M}(\mathcal{Y}^g)$. Finally, let $\lambda^{k,g} \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$ be the stationary process measure satisfying $\lambda^{k,g} = \Lambda^{k+g}[\pi \otimes \gamma]$. In other words, $\lambda^{k,g}$ is obtained by independently concatenating π and γ infinitely many times and randomizing the start over the first $k+g$ coordinates. Note that $\lambda^{k,g} \in \mathcal{J}(\Lambda^k[\mu,\alpha], \Lambda^k[\nu,\beta])$. Then by the construction of $\lambda^{k,g}$,

$$S\left(\Lambda^k[\mu,\alpha], \Lambda^k[\nu,\beta]\right) \leq \int c\,d\lambda^{k,g} = \frac{1}{k+g}\left(\int c_k\,d\pi + \int c_g\,d\gamma\right) \leq \frac{1}{k+g}\left(T_{c_k}(\mu_k,\nu_k) + g\|c\|_\infty\right).$$

Rearranging terms, multiplying by $(k+g)/k$, and adding $S(\mu,\nu)$ to both sides, we obtain that

$$S(\mu,\nu) - \frac{1}{k}T_{c_k}(\mu_k,\nu_k) \leq S(\mu,\nu) - \frac{k+g}{k}S\left(\Lambda^k[\mu,\alpha], \Lambda^k[\nu,\beta]\right) + \frac{g}{k}\|c\|_\infty$$

$$\leq S(\mu,\nu) - S\left(\Lambda^k[\mu,\alpha],\Lambda^k[\nu,\beta]\right) + \frac{g}{k}\|c\|_\infty, \tag{7.12}$$

where in (7.12) we use the fact that the cost c, and thus the optimal joining cost, is non-negative. By Lemmas 15 and 16, we see that

$$\begin{aligned} S(\mu,\nu) - S\left(\Lambda^k[\mu,\alpha],\Lambda^k[\nu,\beta]\right) &\leq \|c\|_\infty \left(\overline{d}\left(\Lambda^k[\mu,\alpha],\mu\right) + \overline{d}\left(\Lambda^k[\nu,\beta],\nu\right)\right) \\ &\leq \|c\|_\infty \left(\frac{2g}{k+g} + \frac{k}{k+g}(\phi_\mu(g+1)+\phi_\nu(g+1))\right) \\ &\leq \|c\|_\infty \left(\frac{2g}{k} + \frac{k}{k+g}(\phi_\mu(g+1)+\phi_\nu(g+1))\right). \end{aligned}$$

It follows that

$$\left|\frac{1}{k}T_{c_k}(\mu_k,\nu_k) - S(\mu,\nu)\right| \leq \|c\|_\infty \left(\frac{k(\phi_\mu(g+1)+\phi_\nu(g+1))}{k+g} + \frac{3g}{k}\right).$$

In order to bound the other term, we apply Proposition 14 to obtain the bound

$$\left|\hat{S}_k(X_1^n,Y_1^n) - \frac{1}{k}T_{c_k}(\mu_k,\nu_k)\right| \leq \frac{1}{k}T_{c_{\mathcal{X},k}}(\hat{\mu}_{k,n},\mu_k) + \frac{1}{k}T_{c_{\mathcal{Y},k}}(\hat{\nu}_{k,n},\nu_k).$$

Combining the bounds proven above, taking an expectation, and applying Proposition 17, we obtain the result. □

Corollary 6. *Let μ and ν have ϕ-mixing coefficients ϕ_μ and ϕ_ν, respectively, satisfying*

$$\sum_{\ell=0}^{n}(n-\ell)\phi_\mu^{1/2}(\ell) = \mathcal{O}(n^p) \qquad \textit{and} \qquad \sum_{\ell=0}^{n}(n-\ell)\phi_\nu^{1/2}(\ell) = \mathcal{O}(n^p)$$

for some $p \in [1,2)$. Then there exists a constant $C < \infty$ depending only on ϕ_μ and ϕ_ν such that for every $k \geq 1$, $g \geq 0$, and n large enough,

$$\mathbb{E}\left|\hat{S}_{k,n} - S(\mu,\nu)\right| \leq \|c\|_\infty \left(\frac{k(\phi_\mu(g+1)+\phi_\nu(g+1))}{k+g} + \frac{3g}{k} + \frac{C(|\mathcal{X}|^{k/2}+|\mathcal{Y}|^{k/2})}{n^{1-p/2}}\right). \tag{4.1}$$

In particular, if $k(n) < \frac{(2-p)\log n}{\log(|\mathcal{X}|\vee|\mathcal{Y}|)\vee 1}$ and $g(n) = o(k(n))$ with $k(n), g(n) \to \infty$, then the upper bound converges to zero as $n \to \infty$.

Proof. The summability condition for ϕ_μ implies that $n^{-2}\sum_{\ell=0}^{n}(n-\ell+1)\phi_\mu^{1/2}(\ell) = \mathcal{O}(n^{p-2})$ and thus $(n^{-2}\sum_{\ell=0}^{n}(n-\ell+1)\phi_\mu^{1/2}(\ell))^{1/2} = \mathcal{O}(n^{p/2-1})$. Moreover, for every $\varepsilon \in (0, \frac{1}{4}\|c\|_\infty]$, we have $\mathcal{N}(\mathcal{X}^k, \frac{1}{k}c_{\mathcal{X},k}, \varepsilon) \leq |\mathcal{X}|^k$. Thus for large enough n, there is a constant $C < \infty$ such that

$$u_t(k,n) \leq Cn^{p/2-1}\left(\frac{1}{4}\|c\|_\infty - t\right)|\mathcal{X}|^{k/2} \leq \frac{C|\mathcal{X}|^{k/2}}{n^{1-p/2}}.$$

Using the same line of reasoning, one may prove the analogous bound for $v_t(k,n)$. Plugging these bounds into Theorem 5, we obtain the result by letting $t \to 0$. □

7.5 Proofs from Section 5

In this section, we prove the results stated in Section 5. We begin with Lemma 18, which states that the limit in Proposition 7 exists and is equal to a supremum.

Lemma 18. *For any $\eta \geq 0$,*

$$\lim_{k\to\infty} \frac{1}{k} T^{\eta}_{c_k}(\mu_k, \nu_k) = \sup_{k\geq 1} \frac{1}{k} T^{\eta}_{c_k}(\mu_k, \nu_k).$$

Proof. By Fekete's lemma, it suffices to show that the sequence $\{T^{\eta}_{c_k}(\mu_k, \nu_k)\}_{k\geq 1}$ is superadditive. Fix $k, \ell \geq 1$ and let $\pi \in \Pi(\mu_{k+\ell}, \nu_{k+\ell})$ be a solution to $T^{\eta}_{c_{k+\ell}}(\mu_{k+\ell}, \nu_{k+\ell})$. Let $\pi_k \in \mathcal{M}(\mathcal{X}^k \times \mathcal{Y}^k)$ and $\pi_\ell \in \mathcal{M}(\mathcal{X}^\ell \times \mathcal{Y}^\ell)$ be the measures corresponding to the first k coordinates and last ℓ coordinates of π, respectively. Using the stationarity of μ and ν, it is straightforward to show that $\pi_k \in \Pi(\mu_k, \nu_k)$ and $\pi_\ell \in \Pi(\mu_\ell, \nu_\ell)$. Moreover, using the subadditivity of $H_{k+\ell}(\cdot)$,

$$\begin{aligned}
T^{\eta}_{c_{k+\ell}}(\mu_{k+\ell}, \nu_{k+\ell}) &= \int c_{k+\ell}\, d\pi - \eta H_{k+\ell}(\pi) \\
&\geq \int c_k\, d\pi_k - \eta H_k(\pi_k) + \int c_\ell\, d\pi_\ell - \eta H_\ell(\pi_\ell) \\
&\geq T^{\eta}_{c_k}(\mu_k, \nu_k) + T^{\eta}_{c_\ell}(\mu_\ell, \nu_\ell).
\end{aligned}$$

So the sequence $\{T^{\eta}_{c_k}(\mu_k, \nu_k)\}_{k\geq 1}$ is superadditive and the conclusion follows. □

Proposition 7. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$. Then for any $\eta \geq 0$,*

$$\lim_{k\to\infty} \frac{1}{k} T^{\eta}_{c_k}(\mu_k, \nu_k) \;=\; S^{\eta}(\mu, \nu).$$

Proof. Fix $\varepsilon > 0$ and $\eta \geq 0$ and let $\lambda \in \mathcal{J}(\mu, \nu)$ be a joining of μ and ν such that

$$\int c\, d\lambda_1 - \eta h(\lambda) \leq S^{\eta}(\mu, \nu) + \varepsilon.$$

Since the k-dimensional distribution of λ, written as λ_k, satisfies $\lambda_k \in \Pi(\mu_k, \nu_k)$, we have

$$T^{\eta}_{c_k}(\mu_k, \nu_k) \leq \int c_k\, d\lambda_k - \eta H(\lambda_k).$$

As λ is stationary and $h(\lambda) \leq \frac{1}{k} H(\lambda_k)$,

$$\begin{aligned}
S^{\eta}(\mu, \nu) + \varepsilon &\geq \int c\, d\lambda_1 - \eta h(\lambda) \\
&= \frac{1}{k} \int c_k\, d\lambda_k - \eta h(\lambda) \\
&\geq \frac{1}{k} \int c_k\, d\lambda_k - \frac{\eta}{k} H(\lambda_k) \\
&\geq \frac{1}{k} T^{\eta}_{c_k}(\mu_k, \nu_k).
\end{aligned}$$

By Lemma 18 we may take a limit in k and let $\varepsilon \to 0$ to establish that

$$S^{\eta}(\mu, \nu) \geq \lim_{k\to\infty} \frac{1}{k} T^{\eta}_{c_k}(\mu_k, \nu_k).$$

Now let $\{\pi^k\}$ be a sequence with $\pi^k \in \Pi(\mu_k, \nu_k)$ such that

$$\frac{1}{k} \int c_k\, d\pi^k - \frac{\eta}{k} H(\pi_k) \leq \frac{1}{k} T^{\eta}_{c_k}(\mu_k, \nu_k) + \varepsilon_k,$$

where $\varepsilon_k \to 0$. From this sequence, we wish to construct a sequence of joinings converging to a joining of μ and ν. For every $k \geq 1$, let $\lambda^k \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$ be the stationary process measure satisfying

$\lambda^k = \Lambda^k[\pi^k]$. We will now show that the $\mathcal{X}^{\mathbb{N}}$- and $\mathcal{Y}^{\mathbb{N}}$-marginals of λ^k converge weakly to μ and ν. Let $\sigma : \mathcal{X}^{\mathbb{N}} \to \mathcal{X}^{\mathbb{N}}$ and $\tau : \mathcal{Y}^{\mathbb{N}} \to \mathcal{Y}^{\mathbb{N}}$ be the left-shift maps on $\mathcal{X}^{\mathbb{N}}$ and $\mathcal{Y}^{\mathbb{N}}$, respectively. Fix a measurable cylinder set $C = C_1 \times \cdots \times C_m \subset \mathcal{X}^m$ and let $\tilde{C} \subset \mathcal{X}^{\mathbb{N}}$ be its extension to $\mathcal{X}^{\mathbb{N}}$ such that $\tilde{C} = C \times \mathcal{X} \times \mathcal{X} \cdots$. Then for $k \geq m$,

$$\begin{aligned}
\lambda^k(\tilde{C} \times \mathcal{Y}^{\mathbb{N}}) &= \frac{1}{k}\sum_{\ell=0}^{k-1} \tilde{\Lambda}^k[\pi^k](\sigma^{-\ell}\tilde{C} \times \tau^{-\ell}\mathcal{Y}^{\mathbb{N}}) \\
&= \frac{1}{k}\sum_{\ell=0}^{k-1} \tilde{\Lambda}^k[\pi^k](\sigma^{-\ell}\tilde{C} \times \mathcal{Y}^{\mathbb{N}}) \\
&= \frac{1}{k}\sum_{\ell=0}^{k-1} \tilde{\Lambda}^k[\mu_k](\sigma^{-\ell}\tilde{C}) \\
&= \frac{k-m+1}{k}\mu_m(C) + \frac{1}{k}\sum_{\ell=1}^{m-1} \mu_{m-\ell}\left(C_1 \times \cdots \times C_{m-\ell}\right)\mu_\ell\left(C_{m-\ell+1} \times \cdots \times C_m\right).
\end{aligned}$$

Fixing m and taking a limit in k, we see that

$$\lim_{k\to\infty} \lambda^k(\tilde{C} \times \mathcal{Y}^{\mathbb{N}}) = \mu_m(C) = \mu(\tilde{C}).$$

Thus the $\mathcal{X}^{\mathbb{N}}$-marginal of λ^k converges weakly to μ and one may use a similar argument to show that the $\mathcal{Y}^{\mathbb{N}}$-marginal of λ^k converges weakly to ν. So by Lemma 23, $\lambda^{k_\ell} \Rightarrow \lambda \in \mathcal{J}(\mu,\nu)$ for some subsequence $\{\lambda^{k_\ell}\}$. Now for each $\ell \geq 1$, one may show using the definition of λ^{k_ℓ} that $\int c\, d\lambda_1^{k_\ell} = \frac{1}{k_\ell}\int c_{k_\ell}\, d\pi^{k_\ell}$ and $h(\lambda^{k_\ell}) = \frac{1}{k_\ell}H(\pi^{k_\ell})$. Thus, by the upper semicontinuity of $h(\cdot)$ and the continuity and boundedness of c,

$$\begin{aligned}
S^\eta(\mu,\nu) &\leq \int c\, d\lambda_1 - \eta h(\lambda) \\
&\leq \liminf_{\ell\to\infty}\left\{\int c\, d\lambda_1^{k_\ell} - \eta h(\lambda^{k_\ell})\right\} \\
&= \liminf_{\ell\to\infty}\left\{\frac{1}{k_\ell}\int c_{k_\ell}\, d\pi^{k_\ell} - \frac{\eta}{k_\ell}H(\pi^{k_\ell})\right\} \\
&\leq \liminf_{\ell\to\infty}\left\{\frac{1}{k_\ell}T^\eta_{c_{k_\ell}}(\mu_{k_\ell},\nu_{k_\ell}) + \varepsilon_{k_\ell}\right\} \\
&= \lim_{k\to\infty}\frac{1}{k}T^\eta_{c_k}(\mu_k,\nu_k),
\end{aligned}$$

giving the result. □

Proposition 8. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$. Then the entropic optimal joining cost satisfies*

$$\lim_{\eta\to 0} S^\eta(\mu,\nu) = S(\mu,\nu).$$

Proof. Let $\{\eta_n\}$ be a sequence of non-negative integers such that $\eta_n \to 0$ and for every $n \geq 1$ let $\lambda^n \in \mathcal{J}^{\eta_n}_{\min}(\mu,\nu)$. As $\mathcal{J}(\mu,\nu)$ is compact in the weak topology, there exists a subsequence of $\{\lambda^n\}$, which we also refer to as $\{\lambda^n\}$, such that $\lambda^n \Rightarrow \lambda$ for some $\lambda \in \mathcal{J}(\mu,\nu)$. Now let $\lambda^* \in \mathcal{J}_{\min}(\mu,\nu)$. Using the feasibility of λ^n for $S(\mu,\nu)$ and λ^* for $S^{\eta_n}(\mu,\nu)$, it follows that for every $n \geq 1$,

$$\int c\, d\lambda_1^* \leq \int c\, d\lambda_1^n$$

and

$$\int c\, d\lambda_1^n - \eta_n h(\lambda^n) \leq \int c\, d\lambda_1^* - \eta_n h(\lambda^*).$$

Rearranging, we obtain

$$0 \leq \int c\, d\lambda_1^n - \int c\, d\lambda_1^* \leq \eta_n(h(\lambda^n) - h(\lambda^*)). \tag{7.13}$$

As $h(\cdot)$ is bounded, we have $\lim_{n\to\infty} \eta_n(h(\lambda^n) - h(\lambda^*)) = 0$. Taking limits in (7.13) and using the continuity and boundedness of c,

$$\int c\, d\lambda_1 = \lim_{n\to\infty} \int c\, d\lambda_1^n = \int c\, d\lambda_1^* = S(\mu, \nu).$$

It follows that $\lambda \in \mathcal{J}_{\min}(\mu, \nu)$ and $\lim_{n\to\infty} \int c\, d\lambda_1^n = S(\mu, \nu)$. Again using the boundedness of $h(\cdot)$,

$$\lim_{n\to\infty} S^{\eta_n}(\mu, \nu) = \lim_{n\to\infty} \left\{ \int c\, d\lambda_1^n - \eta_n h(\lambda^n) \right\} = \lim_{n\to\infty} \int c\, d\lambda_1^n = S(\mu, \nu).$$

Since $\{\eta_n\}$ was arbitrary, we obtain the result. □

Theorem 9. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ be ergodic. Then for any $\eta > 0$, there exists a sequence $k = k(n)$ with $k \to \infty$ such that $\tilde{S}_{k,n} \to S^\eta(\mu, \nu)$ and $\tilde{\lambda}^{k,n} \Rightarrow \mathcal{J}^\eta_{\min}(\mu, \nu)$ almost surely as $n \to \infty$.*

Proof. Fix some $\eta > 0$. To begin, we would like to construct a sequence $\{k(n)\}$ satisfying

$$\lim_{n\to\infty} \left| \tilde{S}_k(X_1^n, Y_1^n) - S^\eta(\mu, \nu) \right| = 0, \quad \text{almost surely.} \tag{7.14}$$

Our approach will be similar to the unregularized case with the exception that we also have to control the error in the entropies of the estimates. In particular, by Proposition 14,

$$\begin{aligned} \left| \tilde{S}_k(X_1^n, Y_1^n) - \frac{1}{k} T^\eta_{c_k}(\mu_k, \nu_k) \right| \leq \frac{1}{k} T_{c_{\mathcal{X},k}}(\hat{\mu}_{k,n}, \mu_k) + \eta \left| \frac{1}{k} H(\hat{\mu}_{k,n}) - \frac{1}{k} H(\mu_k) \right| \\ + \frac{1}{k} T_{c_{\mathcal{Y},k}}(\hat{\nu}_{k,n}, \nu_k) + \eta \left| \frac{1}{k} H(\hat{\nu}_{k,n}) - \frac{1}{k} H(\nu_k) \right|. \end{aligned}$$

So it is necessary to ensure that the entropy error terms also decay to zero along the sequence $\{k(n)\}$ that we construct. To see that such a sequence exists, fix $\varepsilon > 0$ and $i \in \mathbb{N}$ and recall that since $\hat{\mu}_{i,n} \Rightarrow \mu_i$ and $H(\cdot)$ is weakly continuous, there exists an $n(i) \in \mathbb{N}$ such that

$$\mu\left(\left| \frac{1}{i} H(\hat{\mu}_{i,n(i)}) - \frac{1}{i} H(\mu_i) \right| > \varepsilon \right) \leq 2^{-i}.$$

Then by Borel-Cantelli,

$$\mu\left(\limsup_{i\to\infty} \left| \frac{1}{i} H(\hat{\mu}_{i,n(i)}) - \frac{1}{i} H(\mu_i) \right| > \varepsilon \right) = 0,$$

and it follows that $\left| \frac{1}{i} H(\hat{\mu}_{i,n(i)}) - \frac{1}{i} H(\mu_i) \right| \to 0$, μ-almost surely. Abusing notation somewhat, we obtain a sequence $\{i(n)\}$ by letting $i(n') = \inf_i \{n(i) = n'\}$. Thus

$$\lim_{n\to\infty} \left| \frac{1}{i(n)} H(\hat{\mu}_{i(n),n}) - \frac{1}{i(n)} H(\mu_{i(n)}) \right| \to 0, \qquad \mu\text{-almost surely.}$$

Let $\{j(n)\}$ be a sequence constructed in the analogous manner for ν and let $\{\ell(n)\}$ and $\{m(n)\}$ be admissible sequences for μ and ν. Then letting $\{k(n)\}$ be the sequence defined by $k(n) = \min\{\, i(n), j(n), \ell(n), m(n)\}$, we obtain the desired convergence.

Next we show that the sequence of estimated entropic optimal joinings indexed by $k(n)$ converges weakly to the set of entropic optimal joinings $\mathcal{J}^\eta_{\min}(\mu, \nu)$, almost surely. In order to simplify notation,

we will suppress the dependence of $k(n)$ on n in the rest of the proof. Fix sequences $X_1, X_2, ...$ and $Y_1, Y_2, ...$ in the sets of μ- and ν-measure one on which (7.14) holds. Let $\{\tilde{\lambda}^{k,n}\}_{n\geq 1}$ be the corresponding sequence of estimated entropic optimal joinings. By Lemma 23 in Appendix E, for any subsequence $\{\tilde{\lambda}^{k,n_\ell}\}_{\ell\geq 1}$, there is a further subsequence converging weakly to a joining $\lambda \in \mathcal{J}(\mu,\nu)$. For ease of notation, we refer to this further subsequence again as $\{\tilde{\lambda}^{k,n_\ell}\}_{\ell\geq 1}$. Using the upper semicontinuity of $h(\cdot)$ and the continuity and boundedness of c, one may establish

$$\begin{aligned}\int c\, d\lambda_1 - \eta h(\lambda) &\leq \liminf_{\ell\to\infty}\left\{\int c\, d\tilde{\lambda}_1^{k,n_\ell} - \eta h(\tilde{\lambda}^{k,n_\ell})\right\}\\ &= \liminf_{\ell\to\infty} \frac{1}{k} T^\eta_{c_k}(\mu_k^{n_\ell}, \nu_k^{n_\ell})\\ &= T^\eta(\mu,\nu).\end{aligned}$$

Thus, $\lambda \in \mathcal{J}^\eta_{\min}(\mu,\nu)$ and since the subsequence was arbitrary, we conclude that $\tilde{\lambda}^{k,n} \Rightarrow \mathcal{J}^\eta_{\min}(\mu,\nu)$. By the choice of sequences $X_1, X_2, ...$ and $Y_1, Y_2, ...$, we conclude that $\tilde{\lambda}^{k,n} \Rightarrow \mathcal{J}^\eta_{\min}(\mu,\nu)$ almost surely. □

Before proving Theorem 10, we state a lemma regarding the entropy of the k-block empirical measure $\hat{\mu}_{k,n}$.

Lemma 19. *Suppose that for some $u : \mathbb{N}\times\mathbb{N} \to \mathbb{R}_+$ with $\lim_{n\to\infty} u(k,n) = 0$ for every $k \geq 1$, it holds that*

$$\mathbb{E}\|\mu_k - \hat{\mu}_{k,n}\|_1 \leq u(k,n)$$

for large enough n. Then for any $k \geq 1$ and n large enough,

$$\mathbb{E}|H(\mu_k) - H(\hat{\mu}_{k,n})| \leq u(k,n)\log\left(\frac{|\mathcal{X}|^{3k}}{u(k,n)}\right).$$

A proof of Lemma 19 may be found in Appendix A.

Theorem 10. *Let μ and ν have ϕ-mixing coefficients ϕ_μ and ϕ_ν, respectively, satisfying*

$$\sum_{\ell=0}^{n}(n-\ell)\phi_\mu^{1/2}(\ell) = \mathcal{O}(n^p) \qquad \text{and} \qquad \sum_{\ell=0}^{n}(n-\ell)\phi_\nu^{1/2}(\ell) = \mathcal{O}(n^p)$$

for some $p \in [1,2)$. Then there exists a constant $C < \infty$ depending only on ϕ_μ and ϕ_ν such that for every $\eta > 0$, $k \geq 1$, $g \geq 0$, and large enough n,

$$\begin{aligned}\mathbb{E}\left|\tilde{S}_{k,n} - S^\eta(\mu,\nu)\right| \;\leq\; &\|c\|_\infty(\phi_\mu(g+1)+\phi_\nu(g+1))\frac{k}{k+g} + (3\|c\|_\infty + 2\eta(\log|\mathcal{X}| + \log|\mathcal{Y}|))\frac{g}{k}\\ &+ u(k,n)\left(\frac{\|c\|_\infty}{2} + \frac{\eta}{k}\log\left(\frac{|\mathcal{X}|^{3k}}{u(k,n)}\right)\right) + v(k,n)\left(\frac{\|c\|_\infty}{2} + \frac{\eta}{k}\log\left(\frac{|\mathcal{Y}|^{3k}}{v(k,n)}\right)\right),\end{aligned}$$

where $u(k,n) = C|\mathcal{X}|^{k/2}n^{p/2-1}$ and $v(k,n) = C|\mathcal{Y}|^{k/2}n^{p/2-1}$. In particular, if $k(n) < \frac{(2-p)\log n}{\log(|\mathcal{X}|\vee|\mathcal{Y}|)\vee 1}$ and $g(n) = o(k(n))$ with $k(n), g(n) \to \infty$, then the upper bound converges to zero as $n \to \infty$.

Proof. Let $\eta \geq 0$ and fix $k \geq 1$, $g \geq 0$, and $n \geq k$. By the triangle inequality,

$$\left|\tilde{S}_k(X_1^n, Y_1^n) - S^\eta(\mu,\nu)\right| \leq \underbrace{\left|\tilde{S}_k(X_1^n, Y_1^n) - \frac{1}{k}T^\eta_{c_k}(\mu_k,\nu_k)\right|}_{T_1} + \underbrace{\left|\frac{1}{k}T^\eta_{c_k}(\mu_k,\nu_k) - S^\eta(\mu,\nu)\right|}_{T_2},$$

where T_1 and T_2 denote the two differences on the right hand side, respectively. By Proposition 14, T_1 is bounded as

$$T_1 \leq \frac{1}{k}T_{c_{\mathcal{X},k}}(\hat{\mu}_{k,n},\mu_k) + \frac{\eta}{k}|H(\mu_k) - H(\hat{\mu}_{k,n})| + \frac{1}{k}T_{c_{\mathcal{Y},k}}(\hat{\nu}_{k,n},\nu_k) + \frac{\eta}{k}|H(\nu_k) - H(\hat{\nu}_{k,n})|$$

$$\leq \frac{\|c\|_\infty}{2}(\|\hat{\mu}_{k,n} - \mu_k\|_1 + \|\hat{\nu}_{k,n} - \nu_k\|_1) + \frac{\eta}{k}(|H(\mu_k) - H(\hat{\mu}_{k,n})| + |H(\nu_k) - H(\hat{\nu}_{k,n})|).$$

Taking an expectation of both sides, we obtain

$$\mathbb{E}T_1 \leq \frac{\|c\|_\infty}{2}(\mathbb{E}\|\hat{\mu}_{k,n} - \mu_k\|_1 + \mathbb{E}\|\hat{\nu}_{k,n} - \nu_k\|_1) + \frac{\eta}{k}(\mathbb{E}|H(\mu_k) - H(\hat{\mu}_{k,n})| + \mathbb{E}|H(\nu_k) - H(\hat{\nu}_{k,n})|).$$

As the ℓ_1 norm is equivalent (up to a constant factor) to the optimal transport cost with respect to the Hamming cost δ, one has $\mathbb{E}\|\hat{\mu}_{k,n} - \mu_k\|_1 = 2\mathbb{E}T_\delta(\hat{\mu}_{k,n}, \mu_k)$. Then one may repeat the arguments in the proof of Proposition 17 with $\frac{1}{k}c_{\mathcal{X},k}$ replaced by δ to show that there exists a constant $C < \infty$ such that

$$\mathbb{E}\|\hat{\mu}_{k,n} - \mu_k\|_1 \leq C\left(t + \left(\frac{1}{n^2}\sum_{\ell=0}^{n}(n-\ell+1)\phi_\mu^{1/2}(\ell)\right)^{1/2}\int_t^{1/4}\mathcal{N}(\mathcal{X}^k, \delta, \varepsilon)^{1/2}\,d\varepsilon\right),$$

for any $t \in (0, 1/4]$. Using the summability condition for ϕ_μ and the fact that $\mathcal{N}(\mathcal{X}^k, \delta, \varepsilon) \leq |\mathcal{X}|^k$ for every $\varepsilon > 0$, we obtain for large enough n,

$$\mathbb{E}\|\hat{\mu}_{k,n} - \mu_k\|_1 \leq C\left(t + \left(\frac{1}{4} - t\right)\frac{|\mathcal{X}|^{k/2}}{n^{1-p/2}}\right),$$

where we have absorbed any additional constants arising from the summability condition on ϕ_μ into C. Finally, taking the limit $t \to 0$, we have that for large enough n, $\mathbb{E}\|\hat{\mu}_{k,n} - \mu_k\|_1 \leq u(k,n)$ where $u(k,n) = C|\mathcal{X}|^{k/2}n^{p/2-1}$, thereby satisfying the conditions of Lemma 19. Using the same arguments, one may also establish the analogous bound $\mathbb{E}\|\hat{\nu}_{k,n} - \nu_k\|_1 \leq v(k,n)$ where $v(k,n) = C|\mathcal{Y}|^{k/2}n^{p/2-1}$. Thus, Lemma 19 implies

$$\mathbb{E}T_1 \leq u(k,n)\left(\frac{\|c\|_\infty}{2} + \frac{\eta}{k}\log\left(\frac{|\mathcal{X}|^{3k}}{u(k,n)}\right)\right) + v(k,n)\left(\frac{\|c\|_\infty}{2} + \frac{\eta}{k}\log\left(\frac{|\mathcal{Y}|^{3k}}{v(k,n)}\right)\right),$$

for n large enough.

Next we upper bound T_2. Note that according to Lemma 18 and Proposition 7, $k^{-1}T_{c_k}^\eta(\mu,\nu) \leq S^\eta(\mu,\nu)$. Thus it suffices to bound $S^\eta(\mu,\nu) - k^{-1}T_{c_k}^\eta(\mu,\nu)$. In order to do this, we will approximate an optimal joining of μ and ν by an iid block process with gaps. Let $\pi \in \Pi(\mu_k, \nu_k)$ achieve the minimum in the problem $T_{c_k}^\eta(\mu_k, \nu_k)$ and let $\gamma \in \mathcal{M}(\mathcal{X}^g \times \mathcal{Y}^g)$ be a probability measure with marginals $\alpha \in \mathcal{M}(\mathcal{X}^g)$ and $\beta \in \mathcal{M}(\mathcal{Y}^g)$. Finally, let $\lambda^{k,g} \in \mathcal{M}_s(\mathcal{X}^\mathbb{N} \times \mathcal{Y}^\mathbb{N})$ be the stationary process measure satisfying $\lambda^{k,g} = \Lambda^{k+g}[\pi \otimes \gamma]$. In other words, $\lambda^{k,g}$ is the stationary process obtained by independently concatenating $\pi \otimes \gamma$ and randomizing the start over the first $k+g$ symbols. Note that $\lambda^{k,g} \in \mathcal{J}(\Lambda^{k,g}[\mu,\alpha], \Lambda^{k,g}[\nu,\beta])$ and by Lemma 20,

$$h(\lambda^{k,g}) = h(\Lambda^k[\pi \otimes \gamma]) = h(\tilde{\Lambda}^k[\pi \otimes \gamma]) = \frac{1}{k+g}(H(\pi) + H(\gamma)).$$

Then by the construction of $\lambda^{k,g}$,

$$\begin{aligned} S^\eta\left(\Lambda^k[\mu,\alpha], \Lambda^k[\nu,\beta]\right) &\leq \int c\,d\lambda^{k,g} - \eta h(\lambda^{k,g}) \\ &= \frac{1}{k+g}\left(\int c_k\,d\pi + \int c_g\,d\gamma - \eta H(\pi) - \eta H(\gamma)\right) \\ &\leq \frac{1}{k+g}\left(T_{c_k}^\eta(\mu_k,\nu_k) + g\|c\|_\infty\right). \end{aligned}$$

Rearranging terms, multiplying by $(k+g)/k$, and adding $S^\eta(\mu,\nu)$ to both sides, we obtain

$$T_2 \leq S^\eta(\mu,\nu) - \frac{k+g}{k}S^\eta\left(\Lambda^k[\mu,\alpha], \Lambda^k[\nu,\beta]\right) + \frac{g}{k}\|c\|_\infty$$

$$= S^\eta(\mu,\nu) - S^\eta\left(\Lambda^k[\mu,\alpha],\Lambda^k[\nu,\beta]\right) - \frac{g}{k}S^\eta\left(\Lambda^k[\mu,\alpha],\Lambda^k[\nu,\beta]\right) + \frac{g}{k}\|c\|_\infty.$$

By Lemmas 15 and 16, we see that

$$\begin{aligned}
&S^\eta(\mu,\nu) - S^\eta\left(\Lambda^k[\mu,\alpha],\Lambda^k[\nu,\beta]\right)\\
&\quad\le \|c\|_\infty\left(\overline{d}\left(\Lambda^k[\mu,\alpha],\mu\right) + \overline{d}\left(\Lambda^k[\nu,\beta],\nu\right)\right) + \eta(h(\mu) - h(\Lambda^k[\mu,\alpha]) + h(\nu) - h(\Lambda^k[\nu,\beta]))\\
&\quad\le \|c\|_\infty\left(\frac{2g}{k+g} + \frac{k}{k+g}(\phi_\mu(g+1) + \phi_\nu(g+1))\right) + \eta(h(\mu) - h(\Lambda^k[\mu,\alpha]) + h(\nu) - h(\Lambda^k[\nu,\beta])).
\end{aligned}$$

In order to simplify this bound, note that

$$\begin{aligned}
h(\mu) - h(\Lambda^k[\mu,\alpha]) &= h(\mu) - \frac{1}{k+g}(H(\mu_k) + H(\alpha))\\
&\le \frac{1}{k}H(\mu_k) - \frac{1}{k+g}H(\mu_k)\\
&= \frac{g}{k(k+g)}H(\mu_k)\\
&\le \frac{g\log|\mathcal{X}|^k}{k(k+g)}\\
&= \frac{g\log|\mathcal{X}|}{k+g},
\end{aligned}$$

and similarly $h(\nu) - h(\Lambda^k[\nu,\beta]) \le \frac{g\log|\mathcal{Y}|}{k+g}$. Thus,

$$\begin{aligned}
&S^\eta(\mu,\nu) - S^\eta\left(\Lambda^k[\mu,\alpha],\Lambda^k[\nu,\beta]\right)\\
&\quad\le \|c\|_\infty(\phi_\mu(g+1) + \phi_\nu(g+1))\frac{k}{k+g} + (2\|c\|_\infty + \eta(\log|\mathcal{X}| + \log|\mathcal{Y}|))\frac{g}{k+g}.
\end{aligned}$$

Moreover, using the fact that the independent joining has maximal entropy rate among all joinings with the same marginals, one may establish

$$-S^\eta\left(\Lambda^k[\mu,\alpha],\Lambda^k[\nu,\beta]\right) \le \eta h(\Lambda^k[\mu,\alpha]\otimes\Lambda^k[\nu,\beta]) = \frac{\eta}{k+g}\left(H(\mu_k) + H(\alpha) + H(\nu_k) + H(\beta)\right).$$

As the measure $\gamma \in \mathcal{M}(\mathcal{X}^g \times \mathcal{Y}^g)$ and marginals $\alpha \in \mathcal{M}(\mathcal{X}^g)$ and $\beta \in \mathcal{M}(\mathcal{Y}^g)$ were arbitrary, there is no loss of generality in assuming that $H(\alpha) = H(\beta) = 0$. Thus

$$\begin{aligned}
-S^\eta\left(\Lambda^k[\mu,\alpha],\Lambda^k[\nu,\beta]\right) &\le \frac{\eta}{k+g}\left(H(\mu_k) + H(\nu_k)\right)\\
&\le \frac{\eta}{k+g}\left(\log|\mathcal{X}|^k + \log|\mathcal{Y}|^k\right)\\
&= \frac{\eta k}{k+g}\left(\log|\mathcal{X}| + \log|\mathcal{Y}|\right).
\end{aligned}$$

Combining this and our earlier bounds, we have

$$\begin{aligned}
T_2 &\le \|c\|_\infty(\phi_\mu(g+1) + \phi_\nu(g+1))\frac{k}{k+g} + (2\|c\|_\infty + \eta(\log|\mathcal{X}| + \log|\mathcal{Y}|))\frac{g}{k+g}\\
&\quad + \frac{\eta g}{k}\left(\log|\mathcal{X}| + \log|\mathcal{Y}|\right) + \|c\|_\infty\frac{g}{k}\\
&\le \|c\|_\infty(\phi_\mu(g+1) + \phi_\nu(g+1))\frac{k}{k+g} + (3\|c\|_\infty + 2\eta(\log|\mathcal{X}| + \log|\mathcal{Y}|))\frac{g}{k}.
\end{aligned}$$

Combining the bounds on $\mathbb{E}T_1$ and T_2 proven above, we obtain the result. □

Acknowledgments

KO and ABN were supported in part by NSF grants DMS-1613072 and DMS-1613261. KM gratefully acknowledges the support of NSF CAREER grant DMS-1847144.

References

[1] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[3] F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. *Statistics & Probability Letters*, 76(12):1298–1302, 2006.

[4] M. Beiglböck. Cyclical monotonicity and the ergodic theorem. *Ergodic Theory and Dynamical Systems*, 35(3):710–713, 2015.

[5] E. Bernton, P. Ghosal, and M. Nutz. Entropic optimal transport: geometry and large deviations. *arXiv preprint arXiv:2102.04397*, 2021.

[6] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

[7] P. Berthet, J. Dedecker, and F. Merlevède. Central limit theorem and almost sure results for bivariate empirical W_1 distances. 2020.

[8] E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

[9] R. C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[10] X. Bressaud, R. Fernández, and A. Galves. Speed of $\overline{d}$-convergence for Markov approximations of chains with complete connections. a coupling approach. *Stochastic Processes and Their Applications*, 83(1):127–138, 1999.

[11] Y. Chen, Z. Lin, and H.-G. Müller. Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

[12] N. Courty, R. Flamary, D. Tuia, and T. Corpetti. Optimal transport for data fusion in remote sensing. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 3571–3574. IEEE, 2016.

[13] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016.

[14] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[15] I. Csiszár and Z. Talata. On rate of convergence of statistical estimation of stationary ergodic processes. *IEEE Transactions on Information theory*, 56(8):3637–3641, 2010.

[16] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[17] M. Cuturi and G. Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4):941–965, 2018.

[18] Y. A. Davydov. Mixing conditions for Markov chains. *Theory of Probability & Its Applications*, 18(2):312–328, 1974.

[19] T. de la Rue. An introduction to joinings in ergodic theory. *Discrete & Continuous Dynamical Systems-A*, 15(1):121, 2005.

[20] N. Deb, P. Ghosal, and B. Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *arXiv preprint arXiv:2107.01718*, 2021.

[21] S. Dede. An empirical central limit theorem in L_1 for stationary sequences. *Stochastic Processes and Their Applications*, 119(10):3494–3515, 2009.

[22] J. Dedecker and F. Merlevède. Behavior of the Wasserstein distance between the empirical and the marginal distributions of stationary α-dependent sequences. *Bernoulli*, 23(3):2083–2127, 2017.

[23] E. del Barrio and J.-M. Loubes. The statistical effect of entropic regularization in optimal transportation. *arXiv preprint arXiv:2006.05199*, 2020.

[24] R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[25] P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International Conference on Machine Learning*, pages 1367–1376. PMLR, 2018.

[26] M. H. Ellis. The $\overline{d}$-distance between two Markov processes cannot always be attained by a Markov joining. *Israel Journal of Mathematics*, 24(3-4):269–273, 1976.

[27] S. N. Evans and F. A. Matsen. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

[28] R. Flamary, C. Févotte, N. Courty, and V. Emiya. Optimal spectral transportation with application to music transcription. In *Advances in Neural Information Processing Systems*, 2016.

[29] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[30] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[31] H. Furstenberg. Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Theory of Computing Systems*, 1(1):1–49, 1967.

[32] S. Gallo, M. Lerasle, and D. Takahashi. Markov approximations of chains of infinite order in the $\overline{d}$-metric. *Markov Processes and Related Fields*, 19(1):51–82, 2013.

[33] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

[34] E. Glasner. *Ergodic Theory via Joinings*. Number 101. American Mathematical Soc., 2003.

[35] R. M. Gray, D. L. Neuhoff, and P. C. Shields. A generalization of Ornstein's $\overline{d}$ distance with applications to information theory. *The Annals of Probability*, pages 315–328, 1975.

[36] W. Guo, N. Ho, and M. Jordan. Fast algorithms for computational optimal transport and Wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pages 2088–2097. PMLR, 2020.

[37] N. Ho, X. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via Wasserstein means. In *International Conference on Machine Learning*, pages 1501–1509. PMLR, 2017.

[38] S. Hundrieser, M. Klatt, and A. Munk. Entropic optimal transport on countable spaces: Statistical theory and asymptotics. 2021.

[39] J.-C. Hütter and P. Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

[40] H. Janati, M. Cuturi, and A. Gramfort. Wasserstein regularization for sparse multi-task regression. In *International Conference on Artificial Intelligence and Statistics*, volume 89, 2019.

[41] H. Janati, B. Muzellec, G. Peyré, and M. Cuturi. Entropic optimal transport between unbalanced Gaussian measures has a closed form. *Advances in Neural Information Processing Systems*, 33, 2020.

[42] O. Jenkinson. Ergodic optimization. *Discrete and Continuous Dynamical Systems*, 15(1):197, 2006.

[43] O. Jenkinson. Ergodic optimization in dynamical systems. *Ergodic Theory and Dynamical Systems*, 39(10):2593–2618, 2019.

[44] M. Klatt, C. Tameling, and A. Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443, 2020.

[45] A. V. Kolesnikov and D. A. Zaev. Optimal transportation of processes with infinite Kantorovich distance: Independence and symmetry. *Kyoto Journal of Mathematics*, 57(2):293–324, 2017.

[46] C. Laclau, I. Redko, B. Matei, Y. Bennani, and V. Brault. Co-clustering through optimal transport. In *International Conference on Machine Learning*, pages 1955–1964, 2017.

[47] T. Lin, N. Ho, and M. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pages 3982–3991. PMLR, 2019.

[48] T. Lin, N. Ho, and M. I. Jordan. On the acceleration of the Sinkhorn and Greenkhorn algorithms for optimal transport. *arXiv preprint arXiv:1906.01437*, 2019.

[49] A. Lopes, J. Mengue, J. Mohr, and R. Souza. Entropy, pressure and duality for Gibbs plans in ergodic transport. *Bulletin of the Brazilian Mathematical Society, New Series*, 46(3):353–389, 2015.

[50] A. O. Lopes, J. K. Mengue, J. Mohr, and R. R. Souza. Entropy and variational principle for one-dimensional lattice systems with a general a priori probability: positive and zero temperature. *Ergodic Theory and Dynamical Systems*, 35(6):1925–1961, 2015.

[51] K. Marton and P. C. Shields. Entropy and the consistent estimation of joint distributions. *The Annals of Probability*, pages 960–977, 1994.

[52] K. McGoff, S. Mukherjee, and A. B. Nobel. Gibbs posterior convergence and the thermodynamic formalism. *Annals of Applied Probability*, forthcoming.

[53] K. McGoff and A. B. Nobel. Empirical risk minimization and complexity of dynamical models. *Annals of Statistics*, 48(4):2031–2054, 2020.

[54] K. McGoff and A. B. Nobel. Empirical risk minimization for dynamical systems and stationary processes. *Information and Inference: A Journal of the IMA*, 2021.

[55] G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

[56] L. Mi, W. Zhang, X. Gu, and Y. Wang. Variational Wasserstein clustering. In *European Conference on Computer Vision*, pages 322–337, 2018.

[57] L. Montrucchio and G. Pistone. Kantorovich distance on finite metric spaces: Arens–Eells norm and CUT norms. *Information Geometry*, pages 1–37, 2021.

[58] M. Muskulus and S. Verduyn-Lunel. Wasserstein distances in the analysis of time series and dynamical systems. *Physica D: Nonlinear Phenomena*, 240(1):45–58, 2011.

[59] M. Nutz and J. Wiesel. Entropic optimal transport: Convergence of potentials. *arXiv preprint arXiv:2104.11720*, 2021.

[60] K. O'Connor, K. McGoff, and A. Nobel. Optimal transport for stationary Markov chains via policy iteration. *Journal of Machine Learning Research*, forthcoming.

[61] K. O'Connor, B. Yi, K. McGoff, and A. B. Nobel. Graph optimal transport with transition couplings of random walks. *arXiv preprint arXiv:2106.07106*, 2021.

[62] D. S. Ornstein. An application of ergodic theory to probability theory. *The Annals of Probability*, 1(1):43–58, 1973.

[63] D. S. Ornstein and P. C. Shields. The $\overline{d}$-recognition of processes. *Advances in Mathematics*, 104(2):182–224, 1994.

[64] D. S. Ornstein and B. Weiss. How sampling reveals a process. *The Annals of Probability*, 18(3):905–930, 1990.

[65] V. M. Panaretos and Y. Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application*, 6:405–431, 2019.

[66] M. Perrot, N. Courty, R. Flamary, and A. Habrard. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29:4197–4205, 2016.

[67] G. Peyré and M. Cuturi. *Computational Optimal Transport*, volume 11. Now Publishers, Inc., 2019.

[68] T. Rippl, A. Munk, and A. Sturm. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109, 2016.

[69] L. Rüschendorf and T. Sei. On optimal stationary couplings between stationary processes. *Electronic Journal of Probability*, 17, 2012.

[70] F. Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87. Birkhäuser, 2015.

[71] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[72] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large scale optimal transport and mapping estimation. In *International Conference on Learning Representations*, 2018.

[73] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[74] P. C. Shields. *The Ergodic Theory of Discrete Sample Paths*, volume 13. American Mathematical Soc., 1996.

[75] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

[76] M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society Series B*, 80(1):219–238, 2018.

[77] Z. Talata. Divergence of information-criterion based Markov order estimators for infinite memory processes. In *IEEE International Symposium on Information Theory*, pages 1378–1382, 2010.

[78] Z. Talata. Divergence rates of Markov order estimators and their application to statistical estimation of stationary ergodic processes. *Bernoulli*, 19(3):846–885, 2013.

[79] C. Tameling, M. Sommerfeld, and A. Munk. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5):2744–2781, 2019.

[80] V. Titouan, N. Courty, R. Tavenard, and R. Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284, 2019.

[81] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

[82] Q. Tong and K. Kobayashi. Entropy-regularized optimal transport on multivariate normal and q-normal distributions. *Entropy*, 23(3):302, 2021.

[83] C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

[84] P. Walters. *An Introduction to Ergodic Theory*, volume 79. Springer Science & Business Media, 2000.

[85] S. Wang, T. T. Cai, and H. Li. Optimal estimation of Wasserstein distance on a tree with an application to microbiome studies. *Journal of the American Statistical Association*, pages 1–17, 2020.

[86] W. Wang, J. A. Ozolek, D. Slepčev, A. B. Lee, C. Chen, and G. K. Rohde. An optimal transportation approach for nuclear structure-based pathology. *IEEE Transactions on Medical Imaging*, 30(3):621–631, 2010.

[87] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[88] H. Xu, D. Luo, and L. Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. In *Advances in Neural Information Processing Systems*, pages 3052–3062, 2019.

[89] H. Xu, D. Luo, H. Zha, and L. C. Duke. Gromov-Wasserstein learning for graph matching and node embedding. In *International Conference on Machine Learning*, pages 6932–6941. PMLR, 2019.

[90] D. Zaev. On ergodic decompositions related to the Kantorovich problem. *Journal of Mathematical Sciences*, 216(1):65–83, 2016.

[91] D. A. Zaev. On the Monge-Kantorovich problem with additional linear constraints. *Mathematical Notes*, 98(5-6):725–741, 2015.

A Properties of Entropy and Entropy Rate

Lemma 20. *Let $\mathcal{U}$ be finite, let σ be the left-shift on $\mathcal{U}^{\mathbb{N}}$, and let $\tilde{\gamma} \in \mathcal{M}(\mathcal{U}^{\mathbb{N}})$ be n-stationary, i.e. satisfying $\tilde{\gamma} \circ \sigma^{-n} = \tilde{\gamma}$. Then if $\gamma = \frac{1}{n}\sum_{\ell=0}^{n-1} \tilde{\gamma} \circ \sigma^{-\ell}$, it holds that $h(\gamma) = h(\tilde{\gamma})$.*

Proof. The proof follows the argument outlined in [74]. We find it most convenient to adopt the notation of random variables: let $\tilde{X} = \tilde{X}_1, \tilde{X}_2, ...$ be an n-stationary process with distribution $\tilde{\gamma}$, let S be uniformly distributed on the set $\{0, ..., n-1\}$, and let $X = X_1, X_2, ...$ be the stationary process such that for any $s \in \{0, ..., n-1\}$, $X|S = s$ is equal in distribution to $\tilde{X}_s, \tilde{X}_{s+1}, ...$ Note that, by construction, X has distribution equal γ. We will assume that $\tilde{X}$, X and S are all defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$.

Fix some $k \geq 1$ and note that by the chain rule for entropy, we have

$$H(X_1^k) - H(X_1^k|S) = H(S) - H(S|X_1^k).$$

This implies that

$$\left|\frac{1}{k}H(X_1^k) - \frac{1}{k}H(X_1^k|S)\right| = \frac{1}{k}|H(S) - H(S|X_1^k)| \leq \frac{\log n}{k}.$$

Letting $k \to \infty$, we find

$$h(X) = \lim_{k\to\infty} \frac{1}{k}H(X_1^k) = \lim_{k\to\infty} \frac{1}{k}H(X_1^k|S). \tag{A.1}$$

Now writing out $H(X_1^k|S)$,

$$H(X_1^k|S) = \sum_{s=0}^{n-1} \mathbb{P}(S = s)H(X_1^k|S = s) = \frac{1}{n}\sum_{s=0}^{n-1} H(X_1^k|S = s) = \frac{1}{n}\sum_{s=0}^{n-1} H(\tilde{X}_{s+1}^{s+k}). \tag{A.2}$$

The rest of the argument will follow by showing that $\lim_{k\to\infty} \frac{1}{k}H(\tilde{X}_{s+1}^{s+k})$ is independent of s and equal to the entropy rate $h(\tilde{X})$ of the n-stationary process $\tilde{X}$. To see this, note that the chain rule for the entropy implies

$$H(\tilde{X}_1^{s+k}) = H(\tilde{X}_{s+1}^{s+k}) + H(\tilde{X}_1^s|\tilde{X}_{s+1}^{s+k})$$

for every $s = 0, ..., n-1$. This implies the bound

$$\left|\frac{1}{k}H(\tilde{X}_{s+1}^{s+k}) - \frac{1}{k}H(\tilde{X}_1^{s+k})\right| \leq \frac{\log |\mathcal{X}|^s}{k}.$$

Letting $k \to \infty$ once again, we have

$$\lim_{k\to\infty} \frac{1}{k}H(\tilde{X}_{s+1}^{s+k}) = \lim_{k\to\infty} \frac{1}{k}H(\tilde{X}_1^{s+k}) = \lim_{k\to\infty} \frac{1}{s+k}H(\tilde{X}_1^{s+k}) = h(\tilde{X}).$$

Now, using this fact with (A.1) and (A.2), we obtain

$$h(X) = \frac{1}{n}\sum_{s=0}^{n-1} \lim_{k\to\infty} \frac{1}{k}H(\tilde{X}_{s+1}^{s+k}) = \frac{1}{n}\sum_{s=0}^{n-1} h(\tilde{X}) = h(\tilde{X}).$$

Translating this result back to the measures γ and $\tilde{\gamma}$, we obtain the desired result. □

Next we prove Lemma 19 regarding the entropy of the k-block empirical measure $\hat{\mu}_{k,n}$. Our proof relies on the following well-known bound in information theory. A proof may be found in [14, Theorem 17.3.3].

Lemma C. *Let α and β be two probability measures on a finite set $\mathcal{U}$ such that $\|\alpha - \beta\|_1 \leq \frac{1}{2}$. Then*

$$|H(\alpha) - H(\beta)| \leq \|\alpha - \beta\|_1 \log\left(\frac{|\mathcal{U}|}{\|\alpha - \beta\|_1}\right).$$

Lemma 19. *Suppose that for some $u : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ with $\lim_{n\to\infty} u(k,n) = 0$ for every $k \geq 1$, it holds that*

$$\mathbb{E}\|\mu_k - \hat{\mu}_{k,n}\|_1 \leq u(k,n)$$

for large enough n. Then for any $k \geq 1$ and n large enough,

$$\mathbb{E}|H(\mu_k) - H(\hat{\mu}_{k,n})| \leq u(k,n) \log\left(\frac{|\mathcal{X}|^{3k}}{u(k,n)}\right).$$

Proof. Fix $k \geq 1$ and $n \geq k$ to be chosen later on, and let $\Delta_{k,n} = \|\mu_k - \hat{\mu}_{k,n}\|_1$. For a fixed sequence X_1^n, Lemma C gives us

$$\begin{aligned}
|H(\mu_k) - H(\hat{\mu}_{k,n})| &\leq \Delta_{k,n} \log\left(\frac{|\mathcal{X}|^k}{\Delta_{k,n}}\right) \mathbb{1}(\Delta_{k,n} \leq 1/2) + (\log|\mathcal{X}|^k)\mathbb{1}(\Delta_{k,n} > 1/2) \\
&\leq \Delta_{k,n} \log\left(\frac{|\mathcal{X}|^k}{\Delta_{k,n}}\right) + (\log|\mathcal{X}|^k)\mathbb{1}(\Delta_{k,n} > 1/2).
\end{aligned}$$

Taking an expectation, we have

$$\begin{aligned}
\mathbb{E}|H(\mu_k) - H(\hat{\mu}_{k,n})| &\leq \mathbb{E}\left[\Delta_{k,n} \log\left(\frac{|\mathcal{X}|^k}{\Delta_{k,n}}\right)\right] + \mathbb{P}(\Delta_{k,n} > 1/2) \log|\mathcal{X}|^k \\
&\leq \mathbb{E}\left[\Delta_{k,n} \log\left(\frac{|\mathcal{X}|^k}{\Delta_{k,n}}\right)\right] + 2(\mathbb{E}\Delta_{k,n}) \log|\mathcal{X}|^k \\
&\leq \mathbb{E}\left[\Delta_{k,n} \log\left(\frac{|\mathcal{X}|^k}{\Delta_{k,n}}\right)\right] + u(k,n) \log|\mathcal{X}|^{2k},
\end{aligned}$$

for n large enough. An application of Jensen's inequality to the first term yields

$$\mathbb{E}\left[\Delta_{k,n} \log\left(\frac{|\mathcal{X}|^k}{\Delta_{k,n}}\right)\right] \leq (\mathbb{E}\Delta_{k,n}) \log\left(\frac{|\mathcal{X}|^k}{\mathbb{E}\Delta_{k,n}}\right).$$

Since $\mathbb{E}\Delta_{k,n} \leq u(k,n)$ and $\lim_{n\to\infty} u(k,n) = 0$, for large enough n, $\mathbb{E}\Delta_{k,n}$ and $u(k,n)$ are both less than or equal to $1/2$. Since $x \mapsto x\log(|\mathcal{X}|^k/x)$ is increasing on $x \in [0, 1/2]$, this implies that

$$\mathbb{E}\left[\Delta_{k,n} \log\left(\frac{|\mathcal{X}|^k}{\Delta_{k,n}}\right)\right] \leq u(k,n) \log\left(\frac{|\mathcal{X}|^k}{u(k,n)}\right).$$

Combining this with our earlier bound and gathering terms, we obtain the result. □

B Properties of the Proposed Estimates

Proposition 21. *For any $\eta \geq 0$, the proposed estimates satisfy $\tilde{\lambda}^{k,n} \in \mathcal{J}(\Lambda^k[\hat{\mu}_{k,n}], \Lambda^k[\hat{\nu}_{k,n}])$ and*

$$\int c\, d\tilde{\lambda}_1^{k,n} - \eta h(\tilde{\lambda}^{k,n}) = \tilde{S}_{k,n}.$$

Proof. Let $\pi \in \mathcal{M}(\mathcal{X}^k \times \mathcal{Y}^k)$ be as defined in Section 5. We start by proving that $\tilde{\lambda}^{k,n}$ is invariant under the left-shift $\sigma \times \tau$ on $\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$. Note first that by construction, $\tilde{\Lambda}^k[\pi] \circ (\sigma \times \tau)^{-k} = \tilde{\Lambda}^k[\pi]$. Then,

$$\begin{aligned}
k\tilde{\lambda}^{k,n} \circ (\sigma \times \tau)^{-1} &= \sum_{\ell=0}^{k-1} \tilde{\Lambda}^k[\pi] \circ (\sigma \times \tau)^{-\ell-1} \\
&= \sum_{\ell=1}^{k-1} \tilde{\Lambda}^k[\pi] \circ (\sigma \times \tau)^{-\ell} + \tilde{\Lambda}^k[\pi] \circ (\sigma \times \tau)^{-k} \\
&= \sum_{\ell=1}^{k-1} \tilde{\Lambda}^k[\pi] \circ (\sigma \times \tau)^{-\ell} + \tilde{\Lambda}^k[\pi] \\
&= \sum_{\ell=0}^{k-1} \tilde{\Lambda}^k[\pi] \circ (\sigma \times \tau)^{-\ell} \\
&= k\tilde{\lambda}^{k,n}.
\end{aligned}$$

Thus $\tilde{\lambda}^{k,n} \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$. Next we prove that $\tilde{\lambda}^{k,n} \in \Pi(\Lambda^k[\hat{\mu}_{k,n}], \Lambda^k[\hat{\nu}_{k,n}])$. Fix a measurable set $C \subset \mathcal{X}^{\mathbb{N}}$. Then

$$\begin{aligned}
\tilde{\lambda}^{k,n}(C \times \mathcal{Y}^{\mathbb{N}}) &= \frac{1}{k} \sum_{\ell=0}^{k-1} \tilde{\Lambda}^k[\pi](\sigma^{-\ell} C \times \tau^{-\ell} \mathcal{Y}^{\mathbb{N}}) \\
&= \frac{1}{k} \sum_{\ell=0}^{k-1} \tilde{\Lambda}^k[\pi](\sigma^{-\ell} C \times \mathcal{Y}^{\mathbb{N}}) \\
&= \frac{1}{k} \sum_{\ell=0}^{k-1} \tilde{\Lambda}^k[\hat{\mu}_{k,n}](\sigma^{-\ell} C) \\
&= \Lambda^k[\hat{\mu}_{k,n}](C).
\end{aligned}$$

Since C was arbitary, it follows that the $\mathcal{X}^{\mathbb{N}}$-marginal of $\tilde{\lambda}^{k,n}$ is $\Lambda^k[\hat{\mu}_{k,n}]$. A similar argument will show that the $\mathcal{Y}^{\mathbb{N}}$-marginal of $\tilde{\lambda}^{k,n}$ is $\Lambda^k[\hat{\nu}_{k,n}]$. Thus $\tilde{\lambda}^{k,n} \in \mathcal{J}(\Lambda^k[\hat{\mu}_{k,n}], \Lambda^k[\hat{\nu}_{k,n}])$. Finally, by the construction of $\tilde{\lambda}^{k,n}$,

$$\int c \, d\tilde{\lambda}_1^{k,n} - \eta h(\tilde{\lambda}^{k,n}) = \frac{1}{k} \int c_k \, d\pi - \frac{\eta}{k} H(\pi) = \frac{1}{k} T_{c_k}^{\eta}(\hat{\mu}_{k,n}, \hat{\nu}_{k,n}) = \tilde{S}_{k,n},$$

where the first equality follows Lemma 20. □

C Existence of an Entropic Optimal Joining

Proposition 22. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite, $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$, $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$, and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a non-negative cost function. Then for every $\eta \geq 0$, the set of entropic optimal joinings $\mathcal{J}_{min}^{\eta}(\mu, \nu)$ is non-empty.*

Proof. Fix $\eta \geq 0$ and a sequence $\{\lambda^n\} \subset \mathcal{J}(\mu, \nu)$ such that $\int c \, d\lambda_1^n - \eta h(\lambda^n) \to S^{\eta}(\mu, \nu)$. As $\mathcal{J}(\mu, \nu)$ is compact in the weak topology, we may extract a subsequence $\{\lambda^{n_\ell}\}$ such that $\lambda^{n_\ell} \Rightarrow \lambda$ as $\ell \to \infty$ for some $\lambda \in \mathcal{J}(\mu, \nu)$. As the entropy rate $h(\cdot)$ is weakly upper semicontinuous on $\mathcal{M}_s(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$ and c is continuous and bounded,

$$\int c \, d\lambda_1 - \eta h(\lambda) \leq \liminf_{\ell \to \infty} \left\{ \int c \, d\lambda_1^{n_\ell} - \eta h(\lambda^{n_\ell}) \right\} = S^{\eta}(\mu, \nu).$$

Thus we conclude that $\lambda \in \mathcal{J}_{\text{min}}^{\eta}(\mu, \nu)$ and $\mathcal{J}_{\text{min}}^{\eta}(\mu, \nu)$ is non-empty. □

D Properties of the (c, η)-Transform

Lemma 12. *Let $(\mathcal{U}, d_{\mathcal{U}})$ and $(\mathcal{V}, d_{\mathcal{V}})$ be finite pseudometric spaces, and let $f : \mathcal{U} \to \mathbb{R}$ and $g : \mathcal{V} \to \mathbb{R}$ be real-valued functions. Furthermore, let $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$ be a non-negative cost function satisfying $|c(u, v) - c(u', v')| \leq M(d_{\mathcal{U}}(u, u') + d_{\mathcal{V}}(v, v'))$ for all $u, u' \in \mathcal{U}$ and $v, v' \in \mathcal{V}$ for some $M \in \mathbb{R}$. Then for any $\eta > 0$, $\tilde{f}$ and $\tilde{g}$ satisfy $|\tilde{f}(v) - \tilde{f}(v')| \leq M d_{\mathcal{V}}(v, v')$ and $|\tilde{g}(u) - \tilde{g}(u')| \leq M\, d_{\mathcal{U}}(u, u')$ for all $u, u' \in \mathcal{U}$ and $v, v' \in \mathcal{V}$.*

Proof. We will prove the bound for $\tilde{g}$ and the bound for $\tilde{f}$ will follow from a similar argument. Let the conditions of the proposition hold. Then for any $u, u' \in \mathcal{U}$,

$$\begin{aligned}
\tilde{g}(u) &= -\eta \log \left(\sum_v \exp \left\{ \frac{1}{\eta}(g(v) - c(u, v)) \right\} \right) \\
&= -\eta \log \left(\sum_v \exp \left\{ \frac{1}{\eta}(g(v) - c(u', v) + c(u', v) - c(u, v)) \right\} \right) \\
&= -\eta \log \left(\sum_v \exp \left\{ \frac{1}{\eta}(g(v) - c(u', v)) \right\} \exp \left\{ \frac{1}{\eta}(c(u', v) - c(u, v)) \right\} \right) \\
&\leq -\eta \log \left(\exp \left\{ -\frac{M}{\eta} d_{\mathcal{U}}(u, u') \right\} \sum_v \exp \left\{ \frac{1}{\eta}(g(v) - c(u', v)) \right\} \right) \\
&= M d_{\mathcal{U}}(u, u') + \tilde{g}(u').
\end{aligned}$$

Applying the same argument after exchanging u and u' and using the symmetry of $d_{\mathcal{U}}(\cdot, \cdot)$, the result for $\tilde{g}$ follows. □

E Weak Convergence of Couplings and Joinings

Lemma 23. *Let $\mathcal{U}$ and $\mathcal{V}$ be Polish spaces and $\{\mu^n\} \subset \mathcal{M}(\mathcal{U})$ and $\{\nu^n\} \subset \mathcal{M}(\mathcal{V})$ be sequences satisfying $\mu^n \Rightarrow \mu$ and $\nu^n \Rightarrow \nu$ for some $\mu \in \mathcal{M}(\mathcal{U})$ and $\nu \in \mathcal{M}(\mathcal{V})$. Then for any sequence $\{\pi^n\}$ satisfying $\pi^n \in \Pi(\mu^n, \nu^n)$ for every $n \geq 1$, there exists a subsequence $\{\pi^{n_\ell}\}$ such that $\pi^{n_\ell} \Rightarrow \pi$ for some $\pi \in \Pi(\mu, \nu)$. Moreover, if $\mathcal{U} = \mathcal{X}^{\mathbb{N}}$ and $\mathcal{V} = \mathcal{Y}^{\mathbb{N}}$ for Polish alphabets $\mathcal{X}$ and $\mathcal{Y}$ and for every $n \geq 1$, $\mu^n \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$, $\nu^n \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ and $\pi^n \in \mathcal{J}(\mu^n, \nu^n)$, then $\pi \in \mathcal{J}(\mu, \nu)$.*

Proof. The first part of the lemma follows from basic weak convergence arguments (see for example [83]). Suppose that the second set of conditions hold. Then from the first part of the lemma, it suffices to show that π is invariant under the joint left-shift $\sigma \times \tau : \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$. Since $\sigma \times \tau$ is continuous, for any bounded and continuous $f : \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to \mathbb{R}$, $f \circ (\sigma \times \tau)$ is also bounded and continuous and it follows that

$$\int f \, d[\pi \circ (\sigma \times \tau)^{-1}] = \int f \circ (\sigma \times \tau) \, d\pi = \lim_{\ell \to \infty} \int f \circ (\sigma \times \tau) \, d\pi^{n_\ell} = \lim_{\ell \to \infty} \int f \, d\pi^{n_\ell} = \int f \, d\pi.$$

Since f was arbitrary, we conclude that π is invariant under the left-shift $\sigma \times \tau$ and the second part of the lemma follows. □