COMPUTATION AND CONSISTENT ESTIMATION OF STATIONARY OPTIMAL TRANSPORT PLANS

Kevin O'Connor

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2021

Approved by:

Sayan Banerjee

Kevin McGoff

Andrew B Nobel

Vladas Pipiras

Quoc Tran-Dinh

ABSTRACT

Kevin O'Connor: Computation and Consistent Estimation of Stationary Optimal Transport Plans
(Under the direction of Kevin McGoff and Andrew B Nobel)

Informally, the optimal transport (OT) problem is to align, or couple, two distributions of interest as best as possible with respect to some prespecified cost. A coupling that achieves the minimum cost among all couplings is referred to as an OT plan; the cost of the OT plan is referred to as the OT cost. Researchers in statistics and machine learning have expended a great deal of effort to understand the properties of OT plans and costs. The motivation for this work stems partly from the fact that, unlike many other divergence measures and metrics between distributions, OT plans and costs describe relationships between distributions in a manner that respects the geometry of the underlying space (by way of the specified cost). However, this advantage does not necessarily carry over when standard OT techniques are applied to distributions with specific structure. In the case that the two distributions describe stationary stochastic processes, the OT problem may ignore the differences in the sequential dependence of either process. One must find a way to make the OT problem account for the stationary dependence of the marginal processes.

In this thesis, we study OT for stationary processes, a field that we refer to as *stationary optimal transport*. Through example and theory, we argue that when applying OT to stationary processes, one should incorporate the stationarity into the problem directly – constraining the set of allowed transport plans to those that are stationary themselves. In this way, we only consider transport plans that respect the dependence structure of the marginal processes. We study this constrained OT problem from statistical and computational perspectives, with an eye toward applications in machine learning and data science. In particular, we

1. develop algorithms for computing stationary OT plans of Markov chains.
2. extend these tools for Markov OT to the alignment and comparison of weighted graphs.
3. propose estimates of stationary OT plans based on finite sequences of observations.

We build upon existing techniques in OT as well as draw from a variety of fields including Markov decision processes, graph theory, and ergodic theory. In doing this, we uncover new perspectives on OT and pave the way for additional applications and approaches in future work.

ACKNOWLEDGEMENTS

I am enormously grateful to my advisors, Andrew Nobel and Kevin McGoff, for their mentorship and encouragement. I feel extremely fortunate for all that Andrew has taught me about statistics, doing research, and thinking clearly. I am also thankful for Kevin's insightful questions and ideas, and willingness to advise me remotely. My time working with both has deeply affected the way I do research and I will carry the lessons I learned from them with me for the rest of my career.

I have had the great pleasure of collaborating with some inspiring researchers during my time at UNC. I thank Junier Oliva for welcoming me into his group very early on in my PhD. I learned a great deal from working with him, Manzil Zaheer, Chris Bender, Yang Li, and Juan Jose Garcia on exchangeable density estimation. In addition to my advisors, I thank Bongsoo Yi for somehow finding time during his first and second years to contribute to the content of Chapter 4. Moreover, I thank Carson Mosso, Sayan Mukherjee, Charles Perou, Vladas Pipiras, Aatish Thennavan, and Fred Wright for their time and energy spent as collaborators on their respective projects. Finally, I thank the remainder of my committee, Sayan Banerjee and Quoc Tran-Dinh, for their helpful questions and advice.

I also had an enriching experience as a quantitative research intern at Optiver in Summer 2021. I thank the education and research teams there for giving me an opportunity and the resources to learn something new. I am particularly grateful to Levi Dudte, Randall Laurence, Matt Scholl, and Peter Stobbe. I look forward to returning as a full-time quantitative researcher in the Spring.

The students in the Statistics and Operations Research department have been great company for these four and a half years. I thank Jongwhan Yoo for getting me through the first year courses. I am also grateful to Jonathan Williams and Iain Carmichael for their career guidance. I would also like to thank Samopriya Basu, Elyse Borgert, Kentaro Hoffman, Thomas Keefe, Benjy Leinwand, Carson Mosso, Alexander Murph, Taylor Petty, Nhan Pham, and Alex Touzov for letting me distract them when I needed a break.

I also owe much of this dissertation to my family. I thank my parents, Darlise DiMatteo and Tim O'Connor, and step mom, Julie O'Connor, for giving me the space and resources to pursue my interests. I thank my sister, Sarah O'Connor, and step-brother, Jack Sisofo, for keeping me grounded. I thank my parents-in-law, Ken and Glynis Whitcomb, for their support and hospitality in the last year and a half of my PhD. And I thank the rest of my family for reminding me to have fun once in a while.

Finally, I would like to thank my loving wife, Cambria Whitcomb, for her unending support and encouragement. The journey over the past four and a half years has offered plenty of unexpected twists and turns. She took each of these obstacles in stride and I cannot adequately express my appreciation for this in words. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

OT	Optimal transport
OTC	Optimal transition coupling
HMM	Hidden Markov model
MDP	Markov decision process
TC-MDP	Transition coupling - Markov decision process
GraphOTC	Graph optimal transition coupling
FusedOTC	Fused graph optimal transition coupling
GOT	Graph optimal transport
GW	Gromov-Wasserstein
FGW	Fused Gromov-Wasserstein
COPT	Coordinated optimal transport

LIST OF SYMBOLS

$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_+$	The set of non-negative real numbers
$\mathbb{R}_{>0}$	The set of positive real numbers
$\mathcal{M}(\mathcal{U})$	The set of Borel probability measures on the Polish space $\mathcal{U}$
$\mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$	The set of stationary process measures on the product space $\mathcal{U}^{\mathbb{N}}$
γ_k	The k-dimensional distribution of $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$
$\Pi(\mu, \nu)$	The set of couplings of μ and ν
$\mu \otimes \nu$	The independent coupling of μ and ν
$\mathcal{J}(\mu, \nu)$	The set of joinings of μ and ν
$\mathcal{J}_{\min}(\mu, \nu)$	The set of optimal joinings of μ and ν
$\mathcal{J}_{\min}^{\eta}(\mu, \nu)$	The set of entropic optimal joinings of μ and ν
$\Pi_{\mathrm{M}}(\mu, \nu)$	The set of Markov joinings of μ and ν
$\Pi_{\mathrm{TC}}(\mu, \nu)$	The set of transition couplings of μ and ν
$\Pi_{\mathrm{TC}}(P, Q)$	The set of transition couplings of transition matrices P and Q
$\mathcal{T}(c; \mu, \nu)$	The optimal transport cost of μ and ν with respect to c
$\mathcal{T}_{\eta}(c; \mu, \nu)$	The entropic optimal transport cost of μ and ν with respect to c
$\mathcal{S}(c; \mu, \nu)$	The optimal joining cost of μ and ν with respect to c
$\mathcal{S}_{\eta}(c; \mu, \nu)$	The entropic optimal joining cost of μ and ν with respect to c
c_k	The k-step cost $(u_1^k, v_1^k) \mapsto \sum_{\ell=1}^{k} c(u_\ell, v_\ell)$
$\overline{c}$	The long-run average cost $\limsup_{k\to\infty} \frac{1}{k} c_k(u_1^k, v_1^k)$
$f \oplus g$	The function on $\mathcal{U} \times \mathcal{V}$ defined by $f \oplus g(u, v) = f(u) + g(v)$
$H(\gamma)$	The Shannon entropy of the measure γ
$h(\mu)$	The entropy rate of the process μ
$\tilde{\Lambda}[\gamma]$	The iid block process obtained by concatenating γ independently
$\Lambda[\gamma]$	The stationary process obtained by randomizing the start of $\tilde{\Lambda}[\gamma]$
$c^{\mathcal{X}}$	The $\mathcal{X}$-adapted cost satisfying $c^{\mathcal{X}}(x, x') = \sup_y \lvert c(x, y) - c(x', y) \rvert$
$c^{\mathcal{Y}}$	The $\mathcal{Y}$-adapted cost satisfying $c^{\mathcal{Y}}(y, y') = \sup_x \lvert c(x, y) - c(x, y') \rvert$

CHAPTER 1

Introduction

Optimal transport has captured the interest of researchers in a wide range of fields for more than two centuries now. What began as a straightforward optimization problem motivated by the transportation of physical goods has evolved into a robust framework for developing statistical tools and analyzing data. Indeed, optimal transport has led to new approaches in statistical estimation (Chen et al., 2021; Janati et al., 2019; Frogner et al., 2015; Bassetti et al., 2006; Bernton et al., 2019), deep generative modeling (Arjovsky et al., 2017; Bousquet et al., 2017; Tolstikhin et al., 2018; Salimans et al., 2018), clustering (Ho et al., 2017; Laclau et al., 2017; Mi et al., 2018), and other tasks. Ideas from optimal transport have also found their way into applications such as image processing (Rabin et al., 2014; Papadakis, 2015; Rabin and Papadakis, 2015) and cell modeling (Schiebinger et al., 2019; Demetci et al., 2020; Tong et al., 2020; Yang et al., 2020; Moriel et al., 2021). Why has the optimal transport problem so captivated the statistical community? What advantages do transport-based techniques offer in statistical applications? And finally, what can go wrong when one naively applies optimal transport to stationary processes? After introducing the optimal transport problem, we will spend the rest of this chapter addressing these questions and setting the stage for our exploration into stationary optimal transport.

1.1 What is Optimal Transport?

An early form of the optimal transport problem was first posed by Gaspard Monge in 1781 (Monge, 1781). Monge was interested in the problem of transporting dirt to form embankments with as little work as possible. In particular, if one wishes to construct an embankment in a certain shape and location, what is the rule for moving each piece of dirt from an excavation site to the embankment that requires the minimum total transportation distance? Despite its roots in the literal transportation of mass, we will see that Monge's question and the questions that follow have useful interpretations and applications in mathematics more generally.

Monge's problem may be formulated mathematically by thinking of the collection of dirt to be excavated and embankment to be constructed as probability measures μ and ν on spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. In a sense, $d\mu(x)$ describes the amount of dirt to be excavated at a point $x \in \mathcal{X}$ and $d\nu(y)$ describes the amount of dirt to be piled at a point $y \in \mathcal{Y}$. The cost of moving a piece of dirt between each pair of points is encoded by a real-valued cost function c on $\mathcal{X} \times \mathcal{Y}$, so that $c(x, y)$ represents the cost of moving dirt from x to y. Finally, Monge's problem is to find a map T from $\mathcal{X}$ to $\mathcal{Y}$ minimizing the total cost

$$\int c(x, T(x))\, d\mu(x), \tag{1.1}$$

subject to the constraint that $\mu \circ T^{-1} = \nu$. In this context, any map T satisfying the constraint above may be thought of as a map for *transporting* the pile described by μ to form that described by ν. To minimize the quantity (1.1) is to minimize the total cost of transportation from μ to ν.

In his memoir, Monge considered the properties of such optimal transport rules but was not able to provide an explicit solution at the time. In fact, even though the problem above has a simple formulation and clear interpretation, it need not have a solution. To see this, one may take the simple case of $\mu = \delta_0$ and $\nu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$. In other words, one wishes to transport a pile at a single point to form a pile concentrated on two points. Despite μ and ν describing two very benign distributions, it is clear that no map T satisfying the transport constraint $\mu \circ T^{-1} = \nu$ can exist. More generally, one may observe this same phenomenon when transporting between any discrete and continuous distributions. This phenomenon aligns with the intuition that the act of transportation cannot create or destroy matter and thus the number of objects before transport must be the same as the number of objects after transport.

Over a century and a half later, Leonid Kantorovich proposed an alternative optimal transport problem in the 1940's. As we have described optimal transport above, it is implicitly assumed that the transport map takes mass at each point x to a *single* point y. While this makes sense when physically transporting discrete objects, this restriction is not necessary when working with probability measures. Instead, we may add to our consideration transport plans that take mass at a point x and distribute it in some way across points in $\mathcal{Y}$. This is precisely the problem Leonid Kantorovich studied in the 1940's (Kantorovich, 2006a,b).

Mathematically, the transport plans considered by Kantorovich are formalized as couplings. A coupling of μ and ν is a joint probability measure π on $\mathcal{X} \times \mathcal{Y}$ such that $\pi_{\mathcal{X}} = \mu$ and $\pi_{\mathcal{Y}} = \nu$. Informally, a coupling may be thought of as a joint distribution admitting μ and ν as $\mathcal{X}$ and $\mathcal{Y}$ marginals, respectively. We will use $\Pi(\mu, \nu)$ to refer to the set of couplings of μ and ν. The problem posed by Kantorovich is

$$\begin{aligned} \text{minimize} \quad & \int c(x, y)\, d\pi(x, y) \\ \text{subject to} \quad & \pi \in \Pi(\mu, \nu) \end{aligned} \tag{1.2}$$

We will denote the optimal value attained in (1.2) by $\mathcal{T}(c; \mu, \nu)$. Problem (1.2) is similar in spirit to Monge's problem with the exception that the former minimizes the transport cost over couplings while the other minimizes the transport cost over transport maps. Monge's problem may even be seen as a constrained form of (1.2) by noting that any map T from $\mathcal{X}$ to $\mathcal{Y}$ satisfying the transport constraint $\mu \circ T^{-1} = \nu$ may be used to construct a coupling $d\pi(x, y) = d\mu(x)\delta(T(x) = y)$. Throughout this dissertation, we will be most interested in Kantorovich formulation of the optimal transport problem (1.2). However, we expect that interesting insights may be gained by considering the Monge problem in the context of stationary optimal transport in future work.

By construction, optimal transport plans describe a certain joint distribution of the distributions of interest. As a minimizer of the expected cost, the particular joint distribution specified by an optimal transport plan is one in which the distributions are coupled as closely as possible with respect to the specified cost. In this sense, optimal transport plans define *alignments* of the two marginal distributions. Viewed differently, optimal transport plans are constructed so that if one draws paired samples iid from the joint distribution they define, then these pairs (X, Y) will have low cost $c(X, Y)$ on average. In contrast, drawing the realizations X and Y independently of one another might yield arbitrarily high cost on average.

On the other hand, the optimal transport cost describes the minimum cost of transportation between the marginal distributions. For similar distributions, the optimal transport cost will be small, while for distributions that differ in shape or assign most of their mass to regions far apart from one another, the optimal transport cost will be large. Note that the notions of *similar*, *different*, and *far apart* are with respect to the cost function c that is specified between the spaces $\mathcal{X}$ and $\mathcal{Y}$. Specifically, the cost function in the optimal transport problem encodes the similarities and differences of interest for comparing the two marginal distributions. For example, when one is working in Euclidean space, one commonly uses the standard Euclidean metric as a cost function. In this way, the optimal transport cost may be thought of as lifting a cost between points to a cost between distributions. This allows one to use the optimal transport cost as a means of comparing distributions in a way that incorporates the ambient geometry of the spaces of interest or any additional geometric information specific to the problem at hand.

1.2 Why Optimal Transport?

As we have tried to make clear above, the optimal transport problem provides a natural means of comparing and aligning probability distributions. But what advantages does this approach offer over existing methods for comparing distributions? One key advantage touched upon in the previous section was the fact that the optimal transport problem incorporates the geometry of the spaces of interest via the cost function. When the spaces being considered are identical, it is common practice to use their metric as the cost function. At a high level, this ensures that nearness in samples gives nearness in distribution. As an example, the optimal transport cost between the Gaussian distributions $\mathcal{N}(0, 1)$ and $\mathcal{N}(m, 1)$ with respect to the squared Euclidean distance is simply m. When additional information is available, the cost function also offers a means for the practitioner to tune the optimal transport problem to emphasize features of interest when comparing the distributions. We leverage this flexibility in the experiments detailed in both Chapters 3 and 4.

A second favorable property of optimal transport is that it allows one to compare discrete and continuous distributions. This is not the case for, say, the Kullback-Leibler (KL) divergence, which requires that one of the measures be absolutely continuous with respect to the other. This can cause the KL divergence to fail in very simple examples, such as comparing a point mass at -1 to the uniform distribution on the interval $[0, 1]$. This feature of optimal transport is particularly useful for comparing an empirical probability measure derived from data with elements in a family of models over a continuous space. This is a common consideration in statistical optimal transport described in Chapter 2 and we exploit this feature heavily in our work on estimation described in Chapter 5.

Finally, unlike most other divergences and metrics for probability distributions, the optimal transport problem offers both a transport plan and a cost. In addition to being of interest itself, the optimal transport plan implicitly describes how the optimal transport problem arrived at the cost it did. It also illustrates the parts of the space in which the distributions are most similar or different. This is an incredibly valuable feature that is absent in many other distribution metrics and divergences. We use this feature of optimal transport throughout our work, with a particular emphasis in Chapter 4.

These advantages of optimal transport over other probability divergences and metrics makes transport-based techniques especially attractive for researchers in statistics and machine learning. A wide array of statistical problems may be recast in terms of comparing or aligning distributions and thus lend themselves to transport-based approaches. For example, the task of model fitting may be viewed as selecting an element from a family of distributions that is "closest" to the empirical distribution associated with the data observed up to that point. The optimal transport cost offers a natural measure for quantifying the closeness between the empirical distribution and elements in the family of model distributions that incorporates the geometry of the space of interest. Importantly, with the optimal transport cost, there is no issue considering continuous model distributions despite the empirical distribution being discrete. Similarly, the task of generative modeling may be viewed as learning a map from a simple distribution with low-dimensional support to a more complicated distribution described by some observed data. One may use the optimal transport plan between these two distributions to find a map or alignment for these two distributions. By drawing samples conditionally on the points in the low-dimensional space, one obtains a generative model that fits the observed distribution of the data by construction.

In general, by simply viewing the statistical objects of interest as distributions, one avails herself of the powerful toolkit of optimal transport. This paradigm has given rise to a myriad of new tools in statistics and machine learning. These new tools have made it especially important to understand the properties and behavior of the optimal transport problem in different settings. At a high level, this has been the focus of the optimal transport community in recent years. As we will elaborate upon in Section 1.4, this dissertation is focused on understanding and adapting the optimal transport problem to the case when the distributions of interest are stationary processes. In the next section, we begin to explore this setting through example.

1.3 Two Motivating Examples

In most existing work on optimal transport, 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. When multiple samples are available, these are typically iid replicates of the fixed time behavior. 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 (see Chapter 2 for an overview).

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. While standard optimal transport techniques are applicable to problems such as these, they do not account for the structure of the underlying measures, which reflect dynamic processes rather than static quantities.

In order to illustrate this, we provide an example below in which a coupling of stationary processes does not share the same dependence structure as the marginal processes.

Example 1.1. Let X and Y be iid Bernoulli($1/2, 1/2$) processes, independent of each other, defined on the same probability space. For $i \geq 0$ let $\tilde{X}_i = X_i$, and let $\tilde{Y}_i = X_i$ if i is a power of 2 and $\tilde{Y}_i = Y_i$ otherwise. Then the joint process $(\tilde{X}, \tilde{Y}) = (\tilde{X}_0, \tilde{Y}_0), (\tilde{X}_1, \tilde{Y}_1), \ldots$ is a coupling of X and Y, but it is not stationary.

In Example 1.1, we find that despite being a coupling of iid processes, the coupling $(\tilde{X}, \tilde{Y})$ exhibits long-range, non-stationary behavior. In this sense, couplings of relatively simple processes may have complicated dependence structure themselves. Given this example, it is natural to wonder whether *optimal* couplings of stationary processes are always stationary? In the next example, we show that this is not the case.

Example 1.2. Let $X = X_0, X_1, ...$ and $Y = Y_0, Y_1, ...$ be stationary Markov processes on $\{0, 1\}$ with transition matrices

$$P = \begin{array}{c|cc} & 0 & 1 \\ \hline 0 & 1/2 & 1/2 \\ 1 & 1/2 & 1/2 \end{array} \quad \text{and} \quad Q = \begin{array}{c|cc} & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 1 & 0 \end{array},$$

respectively. Note that the process X corresponding to P is iid, while the process Y corresponding to Q is deterministic (after conditioning on the initial symbol Y_0). Moreover, note that both processes have identical one-dimensional stationary distributions on $\{0, 1\}$ coinciding with the $(1/2, 1/2)$ measure. Stated differently, for any $i \geq 0$, $X_i = 0$ or 1 each with probability $1/2$ and similarly for Y_i.

Let c be the cost function on $\{0, 1\}^{\mathbb{N}} \times \{0, 1\}^{\mathbb{N}}$ defined by $c(x, y) = \delta(x_0 \neq y_0)$. It is easy to see that since the stationary distributions of X and Y are identical, the optimal transport distance between X and Y is zero. In particular, let $(\tilde{X}_0, \tilde{Y}_0)$ be the coupling of X_0 and Y_0 such that $(\tilde{X}_0, \tilde{Y}_0) = (0, 0)$ or $(1, 1)$ each with probability $1/2$. Moreover let $(\tilde{X}, \tilde{Y}) = (\tilde{X}_0, \tilde{Y}_0), (\tilde{X}_1, \tilde{Y}_1), ...$ be the coupling of X and Y obtained by initializing $(\tilde{X}_0, \tilde{Y}_0)$ as above and then letting $\tilde{X}_i$ and $\tilde{Y}_i$ evolve independently of one another for $i \geq 1$. One will find that $(\tilde{X}, \tilde{Y})$ has expected cost equal to zero and is thus optimal despite being non-stationary.

In Example 1.2, we find that the optimal transport cost may fail to capture differences in the sequential dependence of the two processes, focusing entirely on their one-dimensional distributions. Moreover, an optimal coupling of stationary processes need not be stationary itself. These two problems present serious obstacles to the application of optimal transport to stationary processes. As we will see in the next section, these issues are intertwined and may be resolved by incorporating stationarity directly into the optimal transport problem.

1.4 Adapting Optimal Transport for Stationary Processes

Based on the examples above, it is clear that couplings and the optimal transport problem do not necessarily capture the stationary dynamics of stationary processes. This is because the standard optimal transport problem is formulated in terms of static distributions. On the other hand, stationary processes evolve through time. A transport plan which is optimal initially, need not remain optimal as the processes evolve. In fact, an optimal transport plan between stationary processes can have arbitrarily high expected cost in the long run. As such, naively applying standard optimal transport to measures evolving dynamically through time can give one a false sense of closeness between the two systems.

A simple reason why optimal transport falls short in this setting is that it does not take the dynamics of the marginals into account. It is greedy by construction, looking for an optimal plan in a single time step. In a dependent setting, one needs to consider instead optimality over the lifetime of the system. In order to do this, one necessarily needs to take the stationary dependence into account, finding a plan that performs best over time.

A natural approach to incorporate the stationary dependence into the optimal transport problem is to consider only couplings that are also stationary, referred to as *joinings*. We will refer to the set of joinings of two stationary processes μ and ν as $\mathcal{J}(\mu, \nu)$. Note that the independent coupling $\mu \otimes \nu$ of two stationary processes is stationary so the set $\mathcal{J}(\mu, \nu)$ is always non-empty. Joinings were first introduced by Furstenberg (Furstenberg, 1967) and have been studied at length in the ergodic theory literature (Glasner, 2003; de la Rue, 2006, 2020). Additional background on joinings may be found in Chapter 2.

The optimal transport problem may be adapted for stationary processes in a straightforward manner by optimizing the expected cost over the set of joinings rather than couplings:

$$\begin{aligned} \text{minimize} \quad & \int c(x, y)\, d\lambda_1(x, y) \\ \text{subject to} \quad & \lambda \in \mathcal{J}(\mu, \nu). \end{aligned} \tag{1.3}$$

We refer to Problem (1.3) as the *optimal joining problem* or equivalently the *stationary optimal transport problem* and denote the optimal value attained in (1.3) by $\mathcal{S}(c; \mu, \nu)$. Note that since λ is a probability measure over the product space $\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$, we integrate the cost with respect to the one-dimensional distribution λ_1 of λ. Note also that the optimal joining problem is simply a constrained form of the optimal transport problem. It follows that $\mathcal{T}(c; \mu, \nu) \leq \mathcal{S}(c; \mu, \nu)$.

The optimal joining problem may be motivated further by previewing the following result detailed in Chapter 5.

Theorem 1.3. *Let $\overline{c}$ be a real-valued cost function defined on $\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ by $\overline{c}(x, y) = \limsup_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} c(x_\ell, y_\ell)$. Then under appropriate assumptions the optimal joining cost satisfies $\mathcal{S}(c;\mu,\nu) = \mathcal{T}(\overline{c};\mu,\nu)$.*

The function $\overline{c}$ captures the long-run average cost of a paired sequence $(x, y) \in \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$. Thus, the theorem above states that the optimal joining cost is equivalent to performing optimal transport with respect to the long-run average cost $\overline{c}$. In this sense, the optimal joining problem minimizes long-run average cost rather than the expected cost at a fixed time point. Considering the optimal joining problem associated with $\mathcal{S}(c;\mu,\nu)$ rather than the long-run average optimal transport problem associated with $\mathcal{T}(\overline{c};\mu,\nu)$ is much simpler since the set of joinings has nice properties such as compactness and convexity while the function $\overline{c}$ may be highly irregular even if c is well-behaved.

1.5 Overview of Contributions

The optimal joining problem has been studied to some extent (see Chapter 2 for an overview). However, some important questions remain. How can one solve for optimal joinings? How can one estimate optimal joinings from data? And how might one apply the optimal joining problem to tasks in machine learning and data science?

In this dissertation, we take a step toward answering these questions, exploring stationary optimal transport from both computational and statistical perspectives. Our primary aim is to adapt tools from optimal transport and other fields to draw insights about stationary optimal transport plans in a variety of settings. A secondary objective of this dissertation is to draw attention to new applications and future research directions in the field. Our particular contributions are as follows:

1. We develop tractable algorithms for computing stationary optimal couplings of Markov chains. By example, we demonstrate that stationary couplings of Markov chains need not be Markov or stationary themselves. To remedy this, we propose a constrained problem called the optimal transition coupling problem that accounts for both Markovity and stationarity in a

natural manner. In studying this problem, we draw a novel connection to Markov decision processes and exploit this connection to develop tractable algorithms for solving the optimal transition coupling problem. As a proof-of-concept, we apply the optimal transition coupling problem to the task of synchronizing and comparing computer generated music.

2. We leverage the tools developed for Markov OT to the alignment and comparison of weighted graphs. In particular, by noting that a weighted graph may be associated with a Markov chain by means of a simple random walk on its nodes, one may immediately apply Markov OT to weighted graphs. We find that optimal transition couplings of these simple random walks provide soft alignments of the nodes and edges of the two graphs. Moreover, the expected cost of the optimal transition coupling provides a measure of dissimilarity between the two graphs that incorporates differences in global and local structure. This represents a novel approach to graph optimal transport that improves upon existing approaches theoretically and empirically.

3. We propose estimates of stationary optimal transport plans and the stationary optimal transport cost based on finite sequences of observations from the marginal processes. We prove that these estimates are consistent in the large sample limit, extending previous work in ergodic theory to more general cost functions. Under strong mixing assumptions, we also establish a finite-sample upper bound on the expected error of the estimated optimal joining cost. As a special case, we obtain new insights into the estimation of existing joining-based metrics such as Ornstein's $\overline{d}$-distance and the $\overline{\rho}$-distance introduced by Gray, Neuhoff, and Shields. Finally, we introduce an entropically regularized optimal joining problem and extend our estimation scheme and results to this new problem.

1.6 Organization

In Chapter 2, we cover preliminaries and existing work related to optimal transport and stationary processes. In Chapter 3, we develop and study the optimal transition coupling problem for joining Markov chains. In Chapter 4, we extend the optimal transition coupling problem to weighted graphs. Finally, in Chapter 5, we study the problem of estimating optimal joinings from data. Appendices for each chapter may be found at the end of bibliography.

1.7 Bibliographic Notes

The content of this dissertation draws upon three papers, one of which has been accepted for publication and two of which are currently in preparation for submission. In particular, the work on optimal transport for Markov chains described in Chapter 3 appears in (O'Connor et al., 2021b). The extension of Markov optimal transport to weighted graphs described in Chapter 4 appears in (O'Connor et al., 2021c). Finally, the content of Chapter 5 describing an estimation procedure for optimal joinings from finite sequences of observations appears in (O'Connor et al., 2021a).

CHAPTER 2

Background and Related Work

The work described in this dissertation draws upon ideas from several areas of interest within optimal transport. In this chapter, we lay the groundwork for these results by covering some background on these areas. We begin by discussing some general results and references in optimal transport. We then shift our focus to computational and algorithmic considerations for the optimal transport problem. Next, we address some relevant work regarding the statistical aspects of optimal transport. Finally, we conclude with a discussion of existing results in stationary optimal transport.

Notation. Let $\mathbb{R}_+$ denote the non-negative real numbers and $\mathbb{R}_{>0}$ denote the positive real numbers. For a metric space $\mathcal{U}$, let $\mathcal{M}(\mathcal{U})$ denote the set of probability measures on $\mathcal{U}$. We will say that a function $f : \mathcal{U} \to \mathbb{R}$ is $\mathcal{O}(g)$ for some function $g : \mathcal{U} \to \mathbb{R}$ if $\lim_{u\to\infty} f(u)/g(u) = C$ for some $C \in \mathbb{R}$. We will say that $f = o(g)$ if $\lim_{u\to\infty} f(u)/g(u) = 0$. We will use the shorthand U_1^n for a sequence $U_1, ..., U_n \in \mathcal{U}$.

2.1 Optimal Transport

As mentioned in Chapter 1, the study of optimal transport dates back to the work of Monge (Monge, 1781) who proposed the optimal transport problem motivated by the efficient transportation of earth to form embankments. Little progress was made on Monge's problem until Kantorovich (Kantorovich, 2006b) introduced his formulation of the optimal transport problem in 1942, copied below for the convenience of the reader.

$$\mathcal{T}(c; \mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int c(x, y)\, d\pi(x, y).$$

In the case that $\mathcal{X}$ and $\mathcal{Y}$ are equal and compact metric spaces, Kantorovich proved that optimal solutions $\pi \in \Pi(\mu, \nu)$ are characterized by the existence of potential functions $f : \mathcal{X} \to \mathbb{R}$ such that

$|f(x) - f(y)| \leq c(x,y)$ for every $x, y \in \mathcal{X}$ and $f(x) - f(y) = c(x,y)$ with π-probability one. This result is generally referred to as *Kantorovich duality.* It was only in 1948 that Kantorovich noticed the connection to Monge's problem (Kantorovich, 2006a).

(Beiglböck and Schachermayer, 2011) present the proof of Kantorovich duality in the most general setting necessary for our purposes. In particular, it is proven that if $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ is non-negative, measurable, $\mu \otimes \nu$-almost surely finite, then

$$\mathcal{T}(c; \mu, \nu) = \sup_{f \oplus g \leq c} \left\{ \int f(x) \, d\mu(x) + \int g(y) \, d\nu(y) \right\}, \tag{2.1}$$

where $f \oplus g$ is the function on $\mathcal{X} \times \mathcal{Y}$ satisfying $f \oplus g(x, y) = f(x) + g(y)$. Kantorovich duality provides a useful alternate perspective from which to study the optimal transport problem. For one, it provides a means of easily lower bounding the optimal transport cost, i.e. by carefully choosing potential functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ satisfying $f \oplus g \leq c$ and adding their expectations. Secondly, it provides a new avenue for solving optimal transport problems numerically. As we discuss in Section 2.2, one may derive efficient algorithms for obtaining optimal transport plans via the dual problem for a regularized variant of the optimal transport problem.

Under the additional assumption that there is a finite transport plan, (Beiglböck and Schachermayer, 2011) prove that optimality in this setting is characterized by a property of the transport plan called *c-cyclical monotonicity.* A set measurable set $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ is said to be c-cyclically monotone if for every $n \geq 1$ and every collection $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$,

$$\sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_i, y_{i+1}),$$

where we use the convention that $y_{n+1} = y_1$. A transport plan $\pi \in \Pi(\mu, \nu)$ is said to be c-cyclically monotone if there exists a c-cyclically monotone set $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ such that $\pi(\Gamma) = 1$. The significance of the relationship between c-cyclical monotonicity and the optimal transport plan is that it provides a simple condition that one may check to verify or disprove optimality of a transport plan. In Chapter 5, we show that ergodic optimal joinings also satisfy a type of cyclical monotonicity and use this to establish an analogue to Kantorovich duality for the optimal joining problem. We remark that an interesting connection between cyclical monotonicity and ergodic theory was established when Beiglböck gave a succinct proof that cyclical monotonicity for a finite transport plan implies optimality of the plan using the pointwise ergodic theorem (Beiglböck, 2015).

For a comprehensive treatment of the theoretical aspects of optimal transport, we refer the reader to (Villani, 2008), particularly the first six chapters. A number of other texts on optimal transport (Villani, 2003; Santambrogio, 2015; Ambrosio et al., 2008; Rachev and Rüschendorf, 1998, 2006; Peyré and Cuturi, 2019) have been written from a variety of perspectives that may also appeal to the curious reader.

2.2 Computational Optimal Transport

Over the past decade or so, researchers have become increasingly interested in understanding computational aspects of the optimal transport problem. This focus area within the field is referred to as *computational optimal transport*. Computational optimal transport is focused on developing algorithmic approaches to solving the optimal transport problem, particularly in the case that the spaces of interest are finite. In this setting, probability measures may be encoded as vectors and couplings as matrices. The optimal transport problem then becomes a linear program and may be solved using standard solvers such as the network simplex algorithm (Peyré and Cuturi, 2019). Unfortunately, these algorithms do not typically scale well with the dimension d of the marginal probability vectors, with the best solvers exhibiting runtimes scaling like $\mathcal{O}(d^3 \log d)$.

In his seminal paper (Cuturi, 2013), Cuturi demonstrated that one may improve upon this runtime by adding a negative entropy term to the discrete optimal transport problem. Letting $H(\pi) = -\sum_{x,y} \pi(x,y) \log \pi(x,y)$ be the Shannon entropy of a coupling $\pi \in \Pi(\mu, \nu)$, the *entropic optimal transport problem* with coefficient $\eta > 0$ is defined as

$$\mathcal{T}_\eta(c; \mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int c \, d\pi - \eta H(\pi) \right\}. \tag{2.2}$$

Using Lagrangian duality, one may show that optimal solutions π to (2.2) are unique and have the form

$$\pi(x, y) = u(x) \exp\left\{ -\frac{1}{\eta} c(x, y) \right\} v(y), \tag{2.3}$$

for some $u : \mathcal{X} \to \mathbb{R}_{>0}$ and $v : \mathcal{Y} \to \mathbb{R}_{>0}$. Cuturi recognized that this problem can be solved by a matrix scaling algorithm known as the Sinkhorn-Knopp algorithm (Sinkhorn, 1967). This approach is much more efficient than using standard linear program solvers and has opened the door to larger scale applications of optimal transport. Later on, we develop adaptations of the Sinkhorn-Knopp algorithm for performing optimal transport between Markov chains (Chapter 3) and weighted graphs (Chapter 4).

In general, the addition of the negative entropy penalty does cause a discrepancy between $\mathcal{T}_\eta(c; \mu, \nu)$ and $\mathcal{T}(c; \mu, \nu)$. However, in the limit as $\eta \to 0$, this discrepancy goes to zero (Cuturi and Peyré, 2018). On the other hand, in the limit $\eta \to \infty$, the unique solution to (2.2) is the coupling of μ and ν with maximal entropy: the independent coupling $\mu \otimes \nu$. As is clear from the form (2.3) of solutions to (2.2), when $\eta > 0$, each element of an entropic optimal transport plan is positive. When viewing couplings as probability measures rather than matrices, this is equivalent to saying that entropic optimal transport plans have full support. This is in contrast with unregularized optimal transport plans which, as solutions of linear programs, are comprised mostly of zeroes. Finally, we note that problem (2.2) does admit a dual problem akin to (2.1) (see e.g. (Cuturi and Peyré, 2018)). We defer a more detailed discussion of the entropic dual problem to Chapter 5, where it plays a key role in our results.

(Altschuler et al., 2017) provided a more refined analysis of the Sinkhorn-Knopp algorithm for entropic optimal transport. It was shown that this algorithm (with an additional rounding step) produces a coupling of μ and ν with expected cost within ε of $\mathcal{T}(c; \mu, \nu)$ in time $\mathcal{O}(d^2 (\log d) \|c\|_\infty^3 \varepsilon^{-3})$. This result provided theoretical confirmation that one could use entropic optimal transport to approximate the optimal transport cost in time which is effectively linear in the dimension d^2 of the couplings of interest. Compared to the $\mathcal{O}(d^3 \log d)$ complexity observed in standard optimal transport solvers, this is a significant improvement. Additionally, the authors introduced the Greenkhorn algorithm, a greedy version of the Sinkhorn algorithm which has the same complexity but runs faster in practice. Later work (Dvurechensky et al., 2018; Lin et al., 2019) has explored refinements of the complexity bounds for the Sinkhorn-Knopp and Greenkhorn algorithms. In particular, this new analysis has shown that variants of either algorithm are able to achieve $\mathcal{O}(d^2 (\log d) \|c\|_\infty^2 \varepsilon^{-2})$ runtime complexity.

More recent work (Dvurechensky et al., 2018; Lin et al., 2019; Guo et al., 2020) has considered alternative algorithms for solving entropy-regularized optimal transport problems based on gradient descent. At a high level, this body of work proposes to use variants of gradient descent to solve the Lagrangian dual of the entropic optimal transport problem. While we do not explore the extension of gradient-based algorithms to stationary optimal transport in this dissertation, we believe this may be a promising direction for future research.

A selection of other work has proposed adaptations of the optimal transport problem and associated algorithms for application to sequential data. (Muskulus and Verduyn-Lunel, 2011) considered one of the earliest extensions of computational optimal transport to dynamical systems. Given two observed sequences of length n, they proposed that one capture the sequential dependence of the sequences via empirical k-block measures. The k-block measure corresponding to a sequence $X_1, ..., X_n$ is the probability measure on $\mathcal{X}^k$ defined by placing equal mass on each observed k-block, $X_1^k, X_2^{k+1}, ..., X_{n-k+1}^n$. The authors proposed that one compare the observed sequences via the optimal transport cost between their respective empirical k-block measures. As we discuss in Chapter 5, this approach is closely related to the problem of estimating optimal joinings. However, this connection was not addressed by the authors.

(Cazelles et al., 2020) studied an extension of computational optimal transport techniques to stationary time series. To every such time series, one can associate a normalized power spectral density (NPSD). Previous work had suggested comparing time series via the Kullback-Leibler divergence between their NPSD's, but this is only valid if one of them is absolutely continuous with respect to the other. Instead, they argue that one should use the Wasserstein distance between the NPSD's of the marginal time series in order to compare the two.

(Su and Hua, 2017, 2018) also studied an extension of the computational optimal transport problem to sequences. Given two observed sequences $X_1, ..., X_n$ and $Y_1, ..., Y_m$, they proposed to constrain the set of couplings in the optimal transport problem to those that do not disturb the relative order of the sequences too much. The degree to which a coupling π preserves the ordering of the two sequences is quantified by the inverse difference moment,

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\pi(X_i, Y_j)}{({}^{i}\!/_{n} - {}^{j}\!/_{m})^2 + 1}.$$

This quantity is large when the coupling π puts most of its mass on points (X_i, Y_j) such that ${}^{i}/_{n}$ is close to ${}^{j}/_{m}$, indicating that the order of the sequences is preserved.

Another line of work (Cohen et al., 2021; Cuturi and Blondel, 2017; Janati et al., 2020) has explored distances between time series based on dynamic time warping. Similar in spirit to the optimal transport problem, the dynamic time warping problem seeks an alignment of two observed sequences that respects the ordering of the respective sequences and minimizes an expected cost.

2.3 Statistical Optimal Transport

It is well-known (see e.g. (Villani, 2008)) that for certain choices of cost function, the optimal transport cost satisfies the conditions of a metric on a certain space of probability measures. In particular, for any Polish space $(\mathcal{X}, d)$, the optimal transport cost $(\mu, \nu) \mapsto \mathcal{T}^{1/p}(d^p; \mu, \nu)$ for any $p \in [1, \infty)$ defines a metric on the space $\mathcal{P}_p(\mathcal{X})$ of probability measures $\mu \in \mathcal{M}(\mathcal{X})$ satisfying

$$\int d^p(x, x_0)\, d\mu(x) < \infty,$$

where $x_0 \in \mathcal{X}$ is arbitrary. For example, letting c be the Euclidean metric on $\mathbb{R}^2$, $(\mu, \nu) \mapsto \mathcal{T}(c; \mu, \nu)$ defines a metric on the subset $\mathcal{P}_1(\mathbb{R}^2)$ of probability measures on $\mathbb{R}^2$.

Under the same conditions, one may also establish that the optimal transport cost actually metrizes the weak topology. In other words, one may show that the weak convergence $\mu_n \Rightarrow \mu$ holds if and only if $\mathcal{T}(c; \mu_n, \mu) \to 0$. This result has a particular significance when μ_n is defined as an empirical measure based on observations $X_1, ..., X_n$ drawn iid according to the distribution μ, i.e. $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$. Then the convergence $\mathcal{T}(\mu_n, \mu) \to 0$ follows from Varadarajan's theorem (Varadarajan, 1958) which establishes the convergence $\mu_n \Rightarrow \mu$ with probability one.

These features of the optimal transport problem naturally lead to several questions: How fast does the convergence $\mathcal{T}(c; \mu_n, \mu) \to 0$ occur? May one establish $\mathcal{T}(c; \mu_n, \nu_n) \to \mathcal{T}(c; \mu, \nu)$? If so, how fast does this convergence occur? And finally, under what conditions can one obtain finite sample bounds for $\mathbb{E}|\,\mathcal{T}(c; \mu_n, \nu_n) - \mathcal{T}(c; \mu, \nu)|$? Recent work in statistics has aimed to answer these questions, building new intuition about the application of optimal transport techniques to data.

One of the earliest results regarding the speed of convergence for $\mathcal{T}(c; \mu_n, \mu)$ was established by Dudley (Dudley, 1969), who showed that the optimal transport cost suffers from a "curse of dimensionality." For example, if $\mu \in \mathcal{M}(\mathbb{R}^d)$ is absolutely continuous with respect to the Lebesgue measure, the mean optimal transport cost between μ_n and μ goes to zero no faster than $n^{-1/d}$ (up to a constant factor). Subsequent work has sought conditions under which this exponential scaling in dimension is alleviated. (Boissard and Le Gouic, 2014) prove an upper bound the expectation of $\mathcal{T}(c; \mu_n, \mu)$ under a covering number condition on the support of μ. Similarly, (Fournier and Guillin, 2015) provide an upper bound on the same quantity under moment conditions. Finally, (Weed and Bach, 2019) showed that one may obtain faster rates of convergence when the probability measure of interest is supported on some low-dimensional subset of the ambient space. In Chapter 5, we make explicit use of a result of (Boissard and Le Gouic, 2014) establishing a bound on $\mathbb{E}\,\mathcal{T}(c; \mu_n, \mu)$ in the case that μ_n is derived from dependent observations.

A collection of other work has focused on estimating $\mathcal{T}(c; \mu, \nu)$ by means of the plug-in estimator $\mathcal{T}(c; \mu_n, \nu_n)$ where μ_n and ν_n are derived from n observations from μ and ν, respectively. Much of this work (Rippl et al., 2016; Sommerfeld and Munk, 2018; Bigot et al., 2019; Del Barrio et al., 2019; Tameling et al., 2019; Berthet et al., 2020) has focused on proving central limit theorems for the $\mathcal{T}(c; \mu_n, \nu_n)$ under increasingly general conditions on the spaces of interest. We do not explore central limit theorems for the optimal joining problem in this dissertation but acknowledge that it may be of interest in future work. Rather, we focus on showing that the analogous plug-in estimator for the optimal joining cost is consistent and proving finite sample error bounds on this estimated cost (see Chapter 5).

Finally, a more recent collection of work has studied the statistical aspects of the entropic optimal transport problem. Remarkably, it was shown that under suitable conditions like sub-Gaussianity, the entropic optimal transport cost circumvents the curse of dimensionality that the standard optimal transport cost falls victim to (Genevay et al., 2019; Mena and Niles-Weed, 2019). Other work (Mena and Niles-Weed, 2019; Klatt et al., 2020; Hundrieser et al., 2021) has also explored central limit theorems for $\mathcal{T}_\eta(c; \mu_n, \nu_n)$. We describe a consistent estimation scheme and finite sample error bound for a regularized optimal joining problem in Chapter 5.

2.4 Stationary Optimal Transport

As discussed in Chapter 1, this dissertation considers the optimal transport problem when the marginal distributions are stationary processes. Before formally defining a stationary process, it is helpful to define the more general concept of a dynamical system. A dynamical system is a triple $(\mathcal{U}, T, \mu)$ where $\mathcal{U}$ is a Polish space, μ is a probability measure on $\mathcal{U}$, and $T : \mathcal{U} \to \mathcal{U}$ is a map satisfying $\mu \circ T^{-1} = \mu$. When this condition on T is satisfied, it is said that T is a measure-preserving map for μ. Dynamical systems are the primary subject of interest in ergodic theory (see e.g. (Shields, 1996) and (Walters, 2000)) and are used as models for population growth (Zhao, 2003), cell development (Furusawa and Kaneko, 2012), economic systems (Medio and Gallo, 1995; Zhang, 2006), and other phenomena that evolve over time.

Stationary processes, on the other hand, offer a means of modeling evolving systems in a manner more suited to statistical analysis. A *stochastic process* with alphabet $\mathcal{X}$ is a random variable $X = X_1, X_2, ...$ taking values in set $\mathcal{X}^{\mathbb{N}}$. The process X is said to be *stationary* if for every $k \geq 0$, the joint distribution of $X_\ell^{\ell+k}$ is independent of ℓ. In this dissertation, we will primarily be interested in the distributions of stationary processes, which are described by *stationary process measures.* Formally, a probability measure $\mu \in \mathcal{M}(\mathcal{X}^{\mathbb{N}})$ is a stationary process measure if $\mu(\mathcal{X} \times [a_1^k]) = \mu([a_1^k])$ for any $k \geq 1$ and any cylinder set $[a_1^k] = \{x \in \mathcal{X}^{\mathbb{N}} : x_i = a_i, \forall 1 \leq i \leq k\}$. We will use $\mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ to denote the set of stationary process measures with alphabet $\mathcal{X}$. The correspondence between a stationary process X and the stationary process measure μ describing its distribution is made explicit by recognizing the $\mathbb{P}(X \in A) = \mu(A)$ for every measurable $A \subset \mathcal{X}^{\mathbb{N}}$. Whenever there is no risk of confusion, we will use "stationary process" and "stationary process measure" interchangeably throughout the rest of the dissertation.

One may show that a stationary process $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ is a special case of a dynamical system $(\mathcal{U}, T, \mu)$ by letting $\mathcal{U} = \mathcal{X}^{\mathbb{N}}$ and the map T be the map that maps sequences like $(x_1, x_2, ...) \mapsto (x_2, x_3, ...)$, known as the left-shift map on $\mathcal{X}^{\mathbb{N}}$. While this dissertation focuses on stationary processes rather than dynamical systems, we highlight this connection to give context to related work in the ergodic theory and dynamical systems literature. In particular, the optimal joining problem may be defined for dynamical systems and has been considered for example in (Rüschendorf and Sei, 2012) and (McGoff and Nobel, 2020).

2.4.1 Origins of Stationary Optimal Transport

Stationary couplings, or *joinings*, of dynamical systems were first introduced by Furstenberg (Furstenberg, 1967). Among other things, Furstenberg used joinings to study questions related to the filtering of stochastic processes. For example, under what conditions can one recover a stationary process $X_1, X_2, ...$ from the corrupted process $X_1 + Y_1, X_2 + Y_2, ...$ where $Y_1, Y_2, ...$ is also stationary? Furstenberg showed that this was possible if the two processes were integrable and the only joining of the two was the independent joining, a property referred to as *disjointness.* A detailed treatment of joinings and their role in ergodic theory can be found in the text (Glasner, 2003). For a more high-level overview, the reader may consult the surveys (de la Rue, 2006, 2020).

The first example of a stationary optimal transport problem or optimal joining problem was introduced by Ornstein (Ornstein, 1973). Ornstein proposed a distance between discrete-alphabet stationary processes measures μ and ν on a common space $\mathcal{X}$ called the $\overline{d}$-distance, defined as

$$\overline{d}(\mu, \nu) = \inf_{\lambda \in \mathcal{J}(\mu,\nu)} \int \delta(x \neq y) \, d\lambda_1(x, y).$$

In particular, we have $\overline{d}(\mu, \nu) = \mathcal{S}(\delta; \mu, \nu)$. In his paper, Ornstein studied the properties of the class of B-processes in relation to the $\overline{d}$-metric. For example, he establishes that the class of B-processes is closed under convergence in $\overline{d}$. Later work (Ornstein and Weiss, 1990) studied how processes can be recovered from a finite number of observations in a $\overline{d}$-consistent manner. Similarly, (Ornstein and Shields, 1994) considers the question of when a process can be recovered in $\overline{d}$-distance from a finite number of observations.

Shortly after Ornstein introduced the $\overline{d}$-distance, (Gray et al., 1975) proposed a generalization to continuous alphabets known as the $\overline{\rho}$-distance. For stationary process measures μ and ν with a common Polish alphabet $(\mathcal{X}, d)$, the $\overline{\rho}$-distance is defined by

$$\overline{\rho}(\mu, \nu) = \sup_{k \geq 1} \frac{1}{k} \, \mathcal{T}(d_k; \mu_k, \nu_k),$$

where we remind the reader that d_k is the map $(x_1^k, y_1^k) \mapsto \sum_{\ell=1}^k d(x_\ell, y_\ell)$. Through a subadditivity argument, the authors show that in fact $\overline{\rho}(\mu, \nu) = \mathcal{S}(d; \mu, \nu)$. This fact plays a key role in our results in Chapter 5. In particular, it provides a means of approximating the optimal joining cost through a series of k-step optimal transport costs. In Chapter 5, we also generalize this result to incorporate entropic penalization as described in Section 2.2.

2.4.2 Existing Results in Stationary Optimal Transport

Several decades after the introduction of the $\overline{\rho}$-distance, (Rüschendorf and Sei, 2012) explored conditions under which an optimal joining of dynamical systems $(\mathcal{U}, T, \mu)$ and $(\mathcal{V}, S, \nu)$ could be found explicitly. The authors defined a class of functions $\phi : \mathcal{U} \to \mathcal{V}$ in which the optimal joining of measures μ and $\nu = \mu \circ \phi^{-1}$ was achieved by the joining $\lambda \in \mathcal{J}(\mu, \nu)$ satisfying $d\lambda(x, y) = d\mu(x)\delta(\phi(x) = y)$. Proof of this result relies on an extension of the gluing lemma (see e.g. (Villani, 2008)) to stationary couplings. The authors also extended the $\overline{\rho}$-distance to random fields.

(Lopes and Mengue, 2012) studied the optimal joining problem for dynamical systems in the topological setting with compact spaces, continuous transformations, and continuous cost. They prove a duality result in this setting via Fenchel-Rockafellar duality and study conditions which guarantee uniqueness of the optimal joining. Later work (Lopes et al., 2015) studied a variant of the optimal joining problem with an entropy penalty, subsequently connecting it to the thermodynamic formalism. The authors also proved that a duality result holds for this problem as well.

(Moameni, 2016) studied the optimal transport problem in the case that the cost function and marginal measures are invariant under a given transformation. The author proved that in this setting (with some additional conditions on the cost function), the optimal transport plan may be chosen to be invariant with respect to that same transformation. We establish a similar result in Chapter 5 where the transformation of interest is the left-shift map.

Further work (Zaev, 2015) studied the more general problem of optimal transport with additional linear constraints. Given a subspace $\mathcal{W}$ of real-valued functions defined on $\mathcal{U} \times \mathcal{V}$, the author considered the constrained set of couplings

$$\Pi_{\mathcal{W}}(\mu, \nu) = \left\{ \pi \in \Pi(\mu, \nu) : \int w(u, v)\, d\pi(u, v) = 0, \, \forall w \in \mathcal{W} \right\},$$

and the associated constrained optimal transport problem,

$$\inf_{\pi \in \Pi_{\mathcal{W}}(\mu, \nu)} \int c(u, v)\, d\pi(u, v). \tag{2.4}$$

Note that by letting $\mathcal{U} = \mathcal{X}^{\mathbb{N}}$, $\mathcal{V} = \mathcal{Y}^{\mathbb{N}}$, $\sigma : \mathcal{X}^{\mathbb{N}} \to \mathcal{X}^{\mathbb{N}}$ and $\tau : \mathcal{Y}^{\mathbb{N}} \to \mathcal{Y}^{\mathbb{N}}$ be the left-shift maps, and $\mathcal{W} = \{f - f \circ (\sigma \times \tau) : f : \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to \mathbb{R} \text{ continuous and bounded}\}$, we have that $\Pi_{\mathcal{W}}(\mu, \nu) = \mathcal{J}(\mu, \nu)$. In other words, the optimal joining problem can be written as an optimal transport problem with additional linear constraints. Under certain conditions on the cost c, the author proved that the problem (2.4) satisfies the duality

$$\inf_{\pi \in \Pi_{\mathcal{W}}(\mu,\nu)} \int c(u,v)\, d\pi(u,v) = \sup_{f \oplus g + w \leq c} \left\{ \int f(u)\, d\mu(u) + \int g(v)\, d\nu(v) \right\}, \tag{2.5}$$

where the supremum is taken over bounded and continuous $f : \mathcal{U} \to \mathbb{R}$ and $g : \mathcal{V} \to \mathbb{R}$ and $w \in \mathcal{W}$. While similar, the duality (2.5) differs slightly from that which we show in Chapter 5 for the optimal joining problem. In particular, one instead takes a supremum over f and g satisfying $f \oplus g \leq \overline{c}$ where $\overline{c}$ is the long-run average cost $(x_1^k, y_1^k) \mapsto \limsup_{k \to \infty} \frac{1}{k} c_k(x_1^k, y_1^k)$.

(Zaev, 2015) also introduced a generalization of cyclical monotonicity called $(c, \mathcal{W})$-cyclical monotonicity. While we do not include the definition here, we note that for $\mathcal{W} = \{0\}$, it reduces to the original definition of c-cyclical monotonicity. In general it is distinct from the notion of $\overline{c}$-cyclical monotonicity which plays a role in our results in Chapter 5. Using (2.5), the author showed that solutions to (2.4) are necessarily $(c, \mathcal{W})$-cyclically monotone, though the reverse implication does not hold in general.

(Zaev, 2016) studied a decomposition of the optimal joining problem for dynamical systems in terms of the ergodic decomposition. Specifically, one can write the optimal joining problem as an optimal transport problem with respect to a cost derived from the optimal joining cost of the ergodic components of the marginal processes.

In other work, (Kolesnikov and Zaev, 2017) studied the existence of optimal transport maps (in the sense of (1.1)) between process measures on $\mathbb{R}^\infty$. The authors provided conditions under which such a measure admits a map that pushes forward to a Gaussian process and is a limit of finite dimensional optimal transport maps. Interestingly these conditions involve the relative entropy of the finite dimensional distributions, which plays an active role in optimal transport as described in Section 2.2 and is an active area of interest for the optimal joining problem (see Chapter 5).

More recently, it has been shown that optimal joinings arise naturally as limiting objects of inferential procedures for dynamical systems. (McGoff and Nobel, 2020, 2021) consider the problem of fitting dynamical models to observed data via empirical risk minimization. In the large sample limit, they show that minimization of the empirical risk is equivalent to solving an optimal joining problem for the observed process and the family of dynamical models. Similarly, (McGoff et al., 2021) consider the problem from a Bayesian perspective, performing Gibbs posterior inference for a family of Gibbs processes. It is shown that the partition function has a rate function given by an optimal joining cost.

In parallel with optimal stationary transport, researchers have considered related questions about dynamical systems in a field known as *ergodic optimization*. Given a map $T : \mathcal{U} \to \mathcal{U}$ and a real-valued function $\beta : \mathcal{U} \to \mathbb{R}$, the general problem of interest in ergodic optimization is to find a measure $\mu \in \mathcal{M}(\mathcal{U})$ that is invariant under T and maximizes $\int \beta \, d\mu$. One may view the optimal joining problem as a constrained version of this problem on the space $\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ by letting T be the left-shift, $\beta = -c$, and constraining the optimization problem further to invariant measures satisfying the coupling condition. For more details, we refer the reader to the surveys (Jenkinson, 2006, 2019).

2.4.3 Stationary Optimal Transport for Markov Chains

A collection of other work has considered stationary optimal transport problems specifically for Markov chains. In a series of papers (Ellis, 1976, 1978, 1980a,b), Ellis considered the $\overline{d}$-distance in the case of Markov chains as well as the optimal Markovian joining problem. In particular, he showed that the $\overline{d}$-distance between two Markov chains was not always achieved by a Markov joining. This follows from the fact that the optimal Markovian joining cost does not satisfy the triangle inequality. This result plays a role in our study of optimal transition couplings in Chapter 3, proving that the optimal transition coupling cost is not equal to the optimal Markovian joining cost in general.

Other work has considered couplings specifically tailored to Markov processes called *Markovian couplings*. A Markovian coupling of finite-state, stationary Markov processes μ and ν with transition matrices P and Q is a joining $\lambda \in \mathcal{J}(\mu, \nu)$ that is Markov with a transition matrix R satisfying $R((x, y), \cdot) \in \Pi(P(x, \cdot), Q(y, \cdot))$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We denote the set transition matrices satisfying the condition above by $\Pi(P, Q)$ and the set of Markovian couplings of μ and ν by $\Pi_{\text{TC}}(\mu, \nu)$. In the interest of precision, we refer to such couplings as *transition couplings*. Transition couplings feature prominently in Chapters 3 and 4. In particular, we define the *optimal transition coupling problem*,

$$\min_{\lambda \in \Pi_{\mathrm{TC}}(\mu,\nu)} \int c(x,y)\, d\lambda_1(x,y). \tag{2.6}$$

As we will discuss later on, this problem incorporates the stationarity and Markovity directly into the optimal transport problem.

Existing work has considered an alternative to the optimal transition coupling problem that we refer to as the *one-step optimal transition coupling problem.* Rather than problem (2.6), this other work considers the problem of finding $R \in \Pi(P, Q)$ minimizing $\sum_{x',y'} R((x,y),(x',y'))c(x',y')$ for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$. In other words, the transition matrix R is chosen to minimize the expected cost after one step rather than the long-term average cost as in the optimal transition coupling problem.

(Mufa, 1994) considered one-step optimal transition couplings in both discrete and continuous time. In particular, they illustrated how one may obtain lower bounds on the spectral gap of a Markov process via one-step optimal transition couplings. Along the same lines, (Zhang, 1999) proved the existence of a one-step optimal transition coupling for Markov processes with Polish alphabets. The author also provided conditions under which a Markov process possesses and converges to a unique stationary distribution with respect to the one-step optimal transition coupling cost. (Zhang, 2000) later proved the existence of a one-step optimal transition coupling to non-negative, lower semicontinuous costs. The author also proved a result characterizing the stochastic dominance of one transition kernel by another in terms of the existence of a certain transition coupling.

CHAPTER 3
Optimal Transport for Stationary Markov Chains via Policy Iteration

In this chapter, we study the optimal transport problem for pairs of stationary finite-state Markov chains, with an emphasis on the computation of optimal transition couplings. Transition couplings are a constrained family of transport plans that capture the dynamics of Markov chains. Solutions of the optimal transition coupling (OTC) problem correspond to alignments of the two chains that minimize long-term average cost. We establish a connection between the OTC problem and Markov decision processes, and show that solutions of the OTC problem can be obtained via an adaptation of policy iteration. For settings with large state spaces, we develop a fast approximate algorithm based on an entropy-regularized version of the OTC problem, and provide bounds on its per-iteration complexity. We establish a stability result for both the regularized and unregularized algorithms, from which a statistical consistency result follows as a corollary. We validate our theoretical results empirically through a simulation study, demonstrating that the approximate algorithm exhibits faster overall runtime with low error. Finally, we extend the setting and application of our methods to hidden Markov models, and illustrate the potential use of the proposed algorithms in practice with an application to computer-generated music.

3.1 Introduction

In this chapter, we study the optimal transport (OT) problem in the case where the objects of interest are stationary Markov chains or processes possessing hidden Markov structure. The problem of interest to us is distinct from traditional applications of coupling to Markov chains, e.g., to establish convergence to a stationary distribution. Our interest is in the computation of optimal transport plans for Markov chains that explicitly account for both stationarity and Markovian structure. In particular, we develop algorithms for computing solutions to a Markov-constrained form of the OT problem. The algorithms leverage recent advances in computational OT as well as techniques from Markov decision processes.

The principled extension of computational OT techniques to classes of distributions that possess additional structure, such as martingales or dependent processes, is an important direction of research. Indeed, some variations of constrained OT have been considered in recent work (Beiglböck et al., 2013; Zaev, 2015; Forrow et al., 2019; Moulos, 2021; Backhoff et al., 2020), and several recent applications of OT have focused on dependent observations (Schiebinger et al., 2019; Xu et al., 2018). Extensions of OT to dependent processes open the door to new applications in climate science, finance, epidemiology and other fields, where it is common for observations to possess temporal or spatial structure. The OT problem that we consider is tailored to the alignment and comparison of Markov chains and hidden Markov models (HMMs). As an illustration, we describe in Section 4.5 an application of the proposed techniques to the analysis of computer-generated music.

The primary contributions of this chapter are as follows:

- We formulate a constrained version of the OT problem for stationary Markov chains, referred to as the optimal transition coupling (OTC) problem. The OTC problem aims to align the two chains of interest so as to minimize long-term average cost while preserving Markovity and stationarity.

- We detail an extension of the OTC problem to HMMs. In particular, we describe how one may couple a pair of HMMs via a coupling of their hidden chains using a cost that is derived from the OT cost between their emission distributions.

- We establish a useful connection between the OTC problem and Markov decision processes (MDPs) that provides a means of computing optimal solutions in an efficient manner. Leveraging this connection, we arrive at an algorithm combining policy iteration (Howard, 1960) with OT solvers that we refer to as `ExactOTC` (Algorithm 1). We state in Theorem 3.7 that if the two Markov chains of interest are irreducible, then `ExactOTC` converges to a solution of the OTC problem in a finite number of iterations.

- We introduce an entropically-constrained OTC problem and an associated regularized algorithm, referred to as `EntropicOTC` (Algorithm 2), that exhibits improved computational efficiency in theory and in practice. In Theorems 3.9 and 3.12, we establish upper bounds on the computational complexity of this algorithm, demonstrating that the runtime of each iteration is nearly-linear in the dimension of the couplings under study. This dependence is comparable to the state-of-the-art for computational OT.

- We prove a stability result for the OTC problem, stated formally in Theorem 3.13. Consistency of the plug-in estimate of the optimal transition coupling and its expected cost follows as a corollary (see Corollary 3.14).

The rest of this chapter is organized as follows: We begin by providing some background on optimal transport and define the OTC problem in Section 3.2. In Section 3.3, we detail our extension of the OTC problem to HMMs. In Section 3.4, we establish the connection between the OTC problem and MDPs and state our result regarding `ExactOTC` for obtaining optimal transition couplings. A faster, regularized algorithm `EntropicOTC` for computing optimal transition couplings is described in Section 3.5. In Section 3.6 we present our result regarding the stability of the OTC problem and the statistical consistency of optimal transition couplings computed from data. In Section 3.7 we describe a simulation study and an application of our algorithms to computer-generated music. We close with a discussion of our results in Section 3.8. Proofs for all stated results may be found in Section 3.9. Finally, an appendix containing some supplementary results and information may be found at the end of this dissertation.

Notation. Let $\mathbb{R}_+$ be the non-negative reals and $\Delta_n = \{u \in \mathbb{R}^n_+ \,|\, \sum_{i=1}^n u_i = 1\}$ denote the probability simplex in $\mathbb{R}^n$. Given a metric space $\mathcal{U}$, let $\mathcal{M}(\mathcal{U})$ denote the set of Borel probability measures on $\mathcal{U}$. For a vector $u \in \mathbb{R}^n$, let $\|u\|_\infty = \max_i |u_i|$ and $\|u\|_1 = \sum_i |u_i|$. Occasionally we will treat matrices in $\mathbb{R}^{n\times n}$ as vectors in $\mathbb{R}^{n^2}$.

3.2 The Optimal Transition Coupling Problem

Let $\mathcal{U}$ and $\mathcal{V}$ be metric spaces and $\mu \in \mathcal{M}(\mathcal{U})$ and $\nu \in \mathcal{M}(\mathcal{V})$ be probability measures. Moreover, let $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$ be a non-negative function on $\mathcal{U} \times \mathcal{V}$. Recall from Chapter 1 that the optimal transport problem associated with μ, ν, and c is the program

$$\begin{aligned} \text{minimize} \quad & \int c \, d\pi \\ \text{subject to} \quad & \pi \in \Pi(\mu, \nu), \end{aligned} \tag{3.1}$$

where $\Pi(\mu, \nu)$ is the set of couplings of μ and ν. As a natural first step toward computational OT for dependent processes, we consider the case where the marginal probability measures of interest μ and ν represent stationary Markov chains $X = (X_0, X_1, ...)$ and $Y = (Y_0, Y_1, ...)$ with values in finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Markov chains are a natural choice: their simple dependence structure is conducive to computation, and they can be studied in terms of transition matrices. Without loss of generality, assume that $\mathcal{X}$ and $\mathcal{Y}$ both contain d points. Let $P, Q \in [0,1]^{d \times d}$ be the transition matrices, and let $p, q \in \Delta_d$ be the corresponding stationary distributions, of the chains X and Y, respectively. For a brief overview of the necessary background on Markov chains, we refer the reader to Section 3.9.1. For a more in-depth review of Markov chain theory, we refer the reader to (Levin and Peres, 2017). The extension of the OTC problem to hidden Markov models, detailed in Section 3.3, enables us to apply our approach to non-Markovian processes with long-range dependence and Polish alphabets.

Remark 3.1. The optimal transport problem traces its roots back to the physical transportation of goods. In particular, the optimal coupling offers a means of stochastically matching a supply of some goods to their demand so as to minimize the expected cost of transporting the goods. In his book on the topic, Villani (Villani, 2008) offers an example of transporting loaves of bread between bakeries and cafés to build intuition for the optimal transport problem:

> *Consider a large number of bakeries, producing loaves, that should be transported each morning to cafés where consumers will eat them. The amount of bread that can be produced at each bakery, and the amount that will be consumed at each café are known in advance, and can be modeled as probability measures ... on a certain space ... (equipped with the natural metric such that the distance between two points is the shortest path joining them). The problem is to find in practice where each unit of bread should go, in such a way as to minimize the total transport cost.*

In our setting, the collections of bakeries and cafés correspond to the finite sets $\mathcal{X}$ and $\mathcal{Y}$. However, unlike the static problem described by Villani, we consider a dynamic problem in which the number of loaves produced and consumed at the bakeries and cafés evolves over time. Indeed, we suppose that the amounts produced and consumed are determined by the distributions of stationary Markov chains X and Y. As we now have dependence over time to consider, the new problem is to synchronize the supply with the demand so as to minimize the total cost of transportation over the long term while still ensuring that the bakery and cafe owners are satisfied. To make things easier for the delivery driver, one might agree to consider only transport plans that do not change over time (stationary) and under which the deliveries tomorrow only depend on the deliveries today (Markov).

In principle, one may apply the standard optimal transport problem in the Markov setting by taking $\mathcal{U} = \mathcal{X}$, $\mathcal{V} = \mathcal{Y}$ and identifying an optimal coupling of the stationary distributions p and q. However, this marginal approach does not capture the dependence structure of the chains X and Y. This is made evident when revisiting Example 1.2 from Chapter 1. Recall the setting: $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ with single-letter cost $c(x, y) = \delta(x \neq y)$, and

$$P = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \end{array} \quad \text{and} \quad Q = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}.$$

In particular, the process X corresponding to P is iid, while the process Y corresponding to Q evolves deterministically after the initial symbol Y_0 is drawn randomly. Nevertheless, under a marginal analysis, the optimal transport distance between X and Y is zero since their stationary distributions p and q each coincide with the $(1/2, 1/2)$ measure. *In general, optimal coupling of stationary distributions yields a joint distribution on the product $\mathcal{X} \times \mathcal{Y}$, but it does not provide a means of generating a joint process having X and Y as marginals.* We seek a variation of (3.1) that captures and preserves the stochastic structure, namely stationarity and Markovity, of the processes X and Y.

As an alternative to a marginal analysis, one may consider instead the full measures $\mathbb{P} \in \mathcal{M}(\mathcal{X}^{\mathbb{N}})$ and $\mathbb{Q} \in \mathcal{M}(\mathcal{Y}^{\mathbb{N}})$ of the processes X and Y. Formally, $\mathbb{P}$ is the unique probability measure on $\mathcal{X}^{\mathbb{N}}$

such that for any cylinder set $[a_i^j] := \{(x_0, x_1, ...) \in \mathcal{X}^{\mathbb{N}} : x_k = a_k, i \leq k \leq j\}$,

$$\mathbb{P}([a_i^j]) := p(a_i) \prod_{k=i}^{j-1} P(a_k, a_{k+1}).$$

The measure $\mathbb{Q}$ is defined similarly in terms of q and Q. By definition, the measures $\mathbb{P}$ and $\mathbb{Q}$ are stationary, and Markovian. However, a coupling of $\mathbb{P}$ and $\mathbb{Q}$ on the joint sequence space $\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ need not be stationary or Markovian. To illustrate, we may revisit Example 1.1 from Chapter 1: let X' and Y' be iid Bernoulli($1/2, 1/2$) processes, independent of each other, defined on the same probability space. For $i \geq 0$ let $\tilde{X}_i = X'_i$, and let $\tilde{Y}_i = X'_i$ if i is a power of 2 and $\tilde{Y}_i = Y'_i$ otherwise. One may establish that the joint process $(\tilde{X}, \tilde{Y}) = (\tilde{X}_0, \tilde{Y}_0), (\tilde{X}_1, \tilde{Y}_1), \ldots$ is a coupling of X' and Y', but it is neither stationary nor Markovian. For further examples and discussion of non-Markovian couplings of Markov processes, see (Ellis, 1976, 1978, 1980b,a).

A joint process $(\tilde{X}, \tilde{Y})$ arising from a non-stationary or non-Markovian coupling of $\mathbb{P}$ and $\mathbb{Q}$ has a very different stochastic structure than the processes X and Y themselves, and will be difficult to work with computationally. Thus we wish to exclude such couplings from the feasible set of an optimal transport problem. An obvious fix is to consider the family $\Pi_{\mathrm{M}}(\mathbb{P}, \mathbb{Q})$, defined as the set of couplings $\mathbb{P}$ and $\mathbb{Q}$ that are stationary and Markovian. Viewed as processes, elements of $\Pi_{\mathrm{M}}(\mathbb{P}, \mathbb{Q})$ correspond to joint processes $(\tilde{X}, \tilde{Y})$ that are stationary, Markov, and satisfy $\tilde{X} \sim X$ and $\tilde{Y} \sim Y$. While this is a natural choice, the optimal transport cost associated with Π_{M} may violate the triangle inequality, even when the underlying cost function c is itself a metric, see (Ellis, 1976, 1978). Moreover, the family $\Pi_{\mathrm{M}}(\mathbb{P}, \mathbb{Q})$ is not characterized by a simple set of constraints (Boyle and Petersen, 2009). Motivated by the need for ready interpretation and tractable computation, we consider the set of stationary Markov chains on $\mathcal{X} \times \mathcal{Y}$ whose transition distributions are couplings of those of X and Y. A formal definition is given below. The resulting set of couplings, called transition couplings, is characterized by a simple set of linear constraints involving P and Q, and one may show (see Appendix A.1) that the resulting OT cost does satisfy the triangle inequality as long as the underlying cost c does.

In order to reduce notation when considering vectors and matrices indexed by elements of $\mathcal{X} \times \mathcal{Y}$, we will indicate only the cardinality of the index set and adopt an indexing convention whereby a vector $u \in \mathbb{R}^{d^2}$ is indexed as $u(x, y)$ and a matrix $R \in [0, 1]^{d^2 \times d^2}$ is indexed as $R((x, y), (x', y'))$

for $(x,y),(x',y') \in \mathcal{X} \times \mathcal{Y}$. Note also that vectors of the form $R((x,y),\cdot)$ will be regarded as row vectors.

Definition 3.2. *Let P and Q be transition matrices on finite state spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. A transition matrix $R \in [0,1]^{d^2 \times d^2}$ is a **transition coupling** of P and Q if for every paired-state $(x,y) \in \mathcal{X} \times \mathcal{Y}$, the distribution $R((x,y),\cdot)$ is a coupling of the distributions $P(x,\cdot)$ and $Q(y,\cdot)$, formally $R((x,y),\cdot) \in \Pi(P(x,\cdot),Q(y,\cdot))$. Let $\Pi_{TC}(P,Q)$ denote the set of all transition couplings of P and Q.*

Standard results in Markov chain theory ensure that each transition coupling $R \in \Pi_{\mathrm{TC}}(P,Q)$ admits at least one stationary distribution $r \in \Delta_{d^2}$. Using r and R, one may construct a stationary Markov chain $(\tilde{X},\tilde{Y}) = \{(\tilde{X}_i,\tilde{Y}_i)\}_{i\geq 0}$ taking values in $\mathcal{X} \times \mathcal{Y}$. We will also refer to couplings constructed in this way as transition couplings, as stated in the following definition.

Definition 3.3. *Let X and Y be stationary Markov chains with transition matrices P and Q on the finite state spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. A stationary Markov chain $(\tilde{X},\tilde{Y}) = \{(\tilde{X}_i,\tilde{Y}_i)\}_{i\geq 0}$ taking values in $\mathcal{X} \times \mathcal{Y}$ with transition matrix $R \in [0,1]^{d^2 \times d^2}$ is a **transition coupling** of X and Y if $(\tilde{X},\tilde{Y})$ is a coupling of X and Y and $R \in \Pi_{TC}(P,Q)$.*

Each transition coupling of X and Y may be associated with a process measure $\pi \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$; let $\Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})$ denote the set of all such measures induced by transition couplings of X and Y. As the notation suggests, one may readily show that the process measure π induced by a transition coupling of X and Y is itself a coupling of the process measures $\mathbb{P}$ and $\mathbb{Q}$ associated with X and Y, respectively. As all elements of $\Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})$ are also stationary and Markovian, it follows that $\Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q}) \subset \Pi_{\mathrm{M}}(\mathbb{P},\mathbb{Q})$.

The couplings defined in Definition 3.3 are sometimes referred to as "Markovian couplings" in the literature (Levin and Peres, 2017), and they have been used, for example, to study diffusions (Banerjee and Kendall, 2018, 2016, 2017). We refer to such couplings as "transition couplings" in order to distinguish them from elements of $\Pi_{\mathrm{M}}(\mathbb{P},\mathbb{Q})$. Note that $\Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q}) \neq \emptyset$ since it contains the independent coupling, namely, the stationary Markov chain on $\mathcal{X} \times \mathcal{Y}$ with transition matrix $P \otimes Q((x,y),(x',y')) = P(x,x')\,Q(y,y')$ for all (x,y) and (x',y'). The independent coupling corresponds to a paired chain $(\tilde{X},\tilde{Y}) = \{(\tilde{X}_i,\tilde{Y}_i)\}_{i\geq 0}$ where $\tilde{X}$ and $\tilde{Y}$ are equal in distribution to X and Y, respectively, and evolve independently of one another.

A key advantage of considering $\Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})$ over $\Pi_{\mathrm{M}}(\mathbb{P},\mathbb{Q})$ is that the constraints defining $\Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})$ are linear and thus computationally tractable (the constraints defining $\Pi_{\mathrm{M}}(\mathbb{P},\mathbb{Q})$ are not). As we prove in Proposition 3.4 below, the set $\Pi_{\mathrm{TC}}(P,Q)$ of transition matrices actually characterizes the set $\Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})$ of transition couplings if X and Y are irreducible. Stated differently, the condition $R \in \Pi_{\mathrm{TC}}(P,Q)$ is sufficient to ensure that a chain $(\tilde{X},\tilde{Y})$ with transition matrix R is a transition coupling of X and Y. On the other hand, if X or Y is reducible, a stationary Markov chain with a transition matrix in $\Pi_{\mathrm{TC}}(P,Q)$ need not be a coupling of X and Y as the stationary distributions of P and Q are not unique. This follows from the fact that a transition coupling of reducible chains may admit as marginals any of the chains with transition matrices P or Q. So in order to solve the OTC problem by optimizing over $\Pi_{\mathrm{TC}}(P,Q)$ instead of $\Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})$, we must be careful to avoid this situation. Proposition 3.4 ensures that this cannot occur if X and Y are irreducible.

Proposition 3.4. *Let X and Y be irreducible stationary Markov chains with transition matrices P and Q, respectively. Then any stationary Markov chain with a transition matrix contained in $\Pi_{TC}(P,Q)$ is a transition coupling of X and Y.*

As a result of Proposition 3.4, we may avoid working explicitly with transition couplings of X and Y and work instead with the set of matrices $\Pi_{\mathrm{TC}}(P,Q)$.

Letting $c : \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to \mathbb{R}$ be a cost function defined on sample sequences of X and Y, we define the *optimal transition coupling (OTC) problem* for X and Y with cost c to be the program

$$\begin{aligned} \text{minimize} \quad & \int c \, d\pi \\ \text{subject to} \quad & \pi \in \Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q}). \end{aligned} \tag{3.2}$$

The minimum in (3.2), referred to as the OTC cost, assesses the degree to which the two chains may be "synced up" with respect to c. Any solution to (3.2) describes the joint distribution of the synchronized chains. Moreover, as a consequence of the pointwise ergodic theorem, any optimal transition coupling $\pi \in \Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})$ in Problem (3.2) is also optimal with respect to the averaged cost $((x_0, x_1, ...), (y_0, y_1, ...)) \mapsto \limsup_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} c((x_i, x_{i+1}, ...), (y_i, y_{i+1}, ...))$. In this sense, the quality of an alignment (equivalently, transition coupling) of the two chains X and Y is assessed based on its long-term average cost.

In the remainder of the chapter we assume that c is a single-letter cost, i.e., $c((x_0, x_1, ...), (y_0, y_1, ...)) = \tilde{c}(x_0, y_0)$ for some cost function $\tilde{c} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. In most of what follows we identify c and $\tilde{c}$, regarding c as a function on $\mathcal{X} \times \mathcal{Y}$ and writing $c(x_0, y_0)$ when no confusion will arise. The consideration of single-letter costs is motivated by our focus on computation and reflects existing work on computational OT, where a cost or metric is defined *a priori* on static observations. Single letter costs have also been the focus of previous work on optimal transport problems for stationary processes (Ornstein, 1973; Gray et al., 1975). Our arguments may be easily adapted to the case when the cost depends on a finite number of coordinates. In particular, any k-letter cost $c : \mathcal{X}^k \times \mathcal{Y}^k \to \mathbb{R}_+$ may be regarded as single-letter for the chains $\tilde{X} = (X_0^{k-1}, X_1^k, ...)$ and $\tilde{Y} = (Y_0^{k-1}, Y_1^k, ...)$ on $\mathcal{X}^k$ and $\mathcal{Y}^k$, respectively. For single-letter costs, we show in Appendix A.1 that optimal transition couplings exist, and that the OTC cost satisfies the triangle inequality whenever c does. Note that c is necessarily bounded, as $\mathcal{X}$ and $\mathcal{Y}$ are finite. Moreover, there is no loss in generality in assuming that c is non-negative since our results also hold after adding a constant to c.

A primary contribution of this chapter, and the focus of Sections 3.4 and 3.5, is the development of efficient algorithms for computing solutions to the OTC problem (3.2). Note that this problem involves the minimization of a linear objective over the non-convex set $\Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q})$, which makes it difficult to find a solution with off-the-shelf methods. Proposition 3.4 shows that one may optimize instead over the convex polyhedron $\Pi_{\text{TC}}(P, Q)$: informally, the program (3.2) can be reformulated as minimizing $\mathbb{E}c(\tilde{X}_0, \tilde{Y}_0)$ over $R \in \Pi_{\text{TC}}(P, Q)$, where $(\tilde{X}, \tilde{Y})$ is a stationary Markov chain generated by R. However, this reformulation has a non-convex objective, so some care is needed in order to obtain global solutions.

3.2.1 Related Work

For general references related to computational and stationary optimal transport, we refer the reader to Chapter 2. In the context of Markov chains, coupling methods have been widely used as a tool to establish rates of convergence (see for instance (Griffeath, 1976) or (Lindvall, 2002)). Examples of optimal Markovian couplings of Markov processes are studied in (Ellis, 1976, 1978, 1980a,b). Another line of work has explored total variation-type distances for models with Markovian structure. For example, (Chen and Kiefer, 2014) and (Kiefer, 2018) develop algorithms for and consider the computability of the total variation distance between hidden Markov models and labeled Markov chains. Similarly, (Daca et al., 2016) studies the inestimability of the total variation distance between Markov chains. More recent work has proposed direct adaptations of the optimal transport problem for processes with Markovian structure. (Moulos, 2021) studied the bicausal optimal transport problem for Markov chains and its connection to Markov decision processes. Unlike the OTC problem, in the bicausal transport problem, couplings are not required to be stationary or Markov themselves. We also remark that the optimal transition coupling problem appears in the unpublished manuscript (Aldous and Diaconis, 2009).

Recall from Chapter 2 that some existing work (Song et al., 2016; Zhang, 2000) has studied a modified form of the OTC problem that we refer to as the *one-step optimal transition coupling problem.* In the one-step OTC problem the expected cost is measured with respect to the one-step transition probabilities rather than the stationary distribution of the transition coupling. In particular, a transition coupling $R \in \Pi_{\mathrm{TC}}(P, Q)$ is one-step optimal if for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$R((x,y),\cdot) \in \underset{r \in \Pi(P(x,\cdot),Q(y,\cdot))}{\operatorname{argmin}} \sum_{x',y'} r(x',y')\, c(x',y').$$

Loosely, one can view the OTC problem (3.2) as an infinite-step version of the one-step OTC problem, wherein a transition coupling is chosen that minimizes the expected cost averaged over an infinite number of steps. The one-step transition coupling problem appears in (Song et al., 2016) where it is used to assess the distance between Markov decision processes. In another direction, (Zhang, 2000) show that solutions to the one-step transition coupling problem exist for Markov processes on Polish state spaces and lower semicontinuous cost functions. While the one-step problem is computationally convenient, in some situations it will yield poor alignments of the two chains of interest. We provide an example to illustrate this in Appendix A.3, showing that the one-step approach can yield a transition coupling with arbitrarily high expected cost over time.

3.3 Extension of OTC to Hidden Markov Models

Markov models are often employed as components of more complex models for sequential observations. Hidden Markov models (HMMs) are a widely used variant of the Markov model in which observations are modeled as conditionally independent random emissions arising from a latent Markov chain. HMMs have been applied successfully to a variety of problems including speech recognition (Bahl et al., 1986; Varga and Moore, 1990), text segmentation (Yamron et al., 1998), and modeling disease progression (Williams et al., 2020). For a detailed overview, we refer the reader to the text (Zucchini et al., 2017).

Formally, a HMM may be characterized by a pair (X, ϕ) where $X = (X_0, X_1, ...)$ is an unobserved Markov chain taking values in a finite set $\mathcal{X}$, and a function $\phi : \mathcal{X} \to \mathcal{M}(\mathcal{U})$ that maps each state $x \in \mathcal{X}$ to a distribution on a fixed observation space $\mathcal{U}$. The pair (X, ϕ) gives rise to a stationary process $U = (U_0, U_1, ...)$ where $U_0, U_1, \ldots \in \mathcal{U}$ are conditionally independent given X with $U_i \sim \phi(X_i)$ for $i \geq 0$. Note that the process U may exhibit long-range dependence. In this way, HMMs provide a simple means of modeling sequences with more complex dependence structures.

The OTC problem may be extended to processes with hidden Markov structure as follows. Let (X, ϕ) and (Y, ψ) be a pair of HMMs with observation spaces $\mathcal{U}$ and $\mathcal{V}$, respectively, and let $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$. Note that the cost c is specified on the observed spaces $\mathcal{U}$ and $\mathcal{V}$ rather than the state spaces of the unobserved Markov chains X and Y. However, one may extend c to a cost on $\mathcal{X} \times \mathcal{Y}$ by optimally coupling the emission distributions $\phi(x)$ and $\psi(y)$ for every pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. In more detail, let $\theta : \mathcal{X} \times \mathcal{Y} \to \mathcal{M}(\mathcal{U} \times \mathcal{V})$ and $c' : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be defined by

$$\theta(x, y) \in \operatorname*{argmin}_{\pi \in \Pi(\phi(x), \psi(y))} \int c \, d\pi \qquad \text{and} \qquad c'(x, y) = \min_{\pi \in \Pi(\phi(x), \psi(y))} \int c \, d\pi.$$

In other words, we define the functions $\theta : \mathcal{X} \times \mathcal{Y} \to \mathcal{M}(\mathcal{U} \times \mathcal{V})$ and $c' : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ such that for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\theta(x, y)$ is an optimal coupling and $c'(x, y)$ is the OT cost of the emission distributions $\phi(x)$ and $\psi(y)$ with respect to c. One may then find an optimal transition coupling (X', Y') of X and Y with respect to c' as in problem (3.2). The expected cost of this transition coupling corresponds to a cost between the HMMs (X, ϕ) and (Y, ψ) taking the original cost c into account. Moreover, the pair $((X', Y'), \theta)$ defines an optimal joint HMM of (X, ϕ) and (Y, ψ) from which samples in $\mathcal{U} \times \mathcal{V}$ may be drawn.

Leveraging the intuition from the standard OTC problem, the optimal transition coupling $((X', Y'), \theta)$ may be thought of as an alignment of the two HMMs (X, ϕ) and (Y, ψ) with respect to c. In this way, we may apply the OTC problem to any processes that can be embedded as or are well-approximated by HMMs. Before proceeding, we remark that (Chen et al., 2019) also proposes an OT problem for HMMs based on coupling the emission distributions of the two HMMs of interest. However, the latent Markov chains of either HMM are coupled using standard OT after a registration step. Our approach captures the Markovity of the latent sequences more directly and allows one to generate new samples from the coupled HMM.

3.4 Computing Optimal Transition Couplings

In this section, we turn our attention toward our primary goal of developing tractable algorithms for solving the OTC problem (3.2). As discussed in Section 3.2, the OTC problem is a non-convex, constrained optimization problem and thus there is little hope of obtaining global solutions via generic optimization algorithms. Adopting a more tailored approach, we draw a connection between the OTC problem and Markov decision processes (MDP). Having established this connection, we may leverage the wealth of algorithms for obtaining global solutions to MDPs to solve the OTC problem. As we will show, the framework of policy iteration naturally lends itself to our problem and leads to a computationally tractable algorithm combining standard MDP techniques with OT solvers.

3.4.1 Connection to Markov Decision Processes

A Markov decision process is characterized by a 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, c')$ consisting of a state space $\mathcal{S}$, an action space $\mathcal{A} = \bigcup_s \mathcal{A}_s$ where $\mathcal{A}_s$ is the set of allowable actions in state s, a set of transition distributions $\mathcal{P} = \{p(\cdot|s,a) : s \in \mathcal{S}, a \in \mathcal{A}\}$ on $\mathcal{S}$, and a cost function $c' : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. At each time step the process occupies a state $s \in \mathcal{S}$ and an agent chooses an action $a \in \mathcal{A}_s$; the process incurs a cost $c'(s,a)$ and then moves to a new state according to the distribution $p(\cdot|s,a)$. Informally, the goal of the agent is to choose actions to minimize her average cost. The behavior of an agent is described by a family $\gamma = \{\gamma_s(\cdot) : s \in \mathcal{S}\}$ of distributions $\gamma_s(\cdot) \in \mathcal{M}(\mathcal{A}_s)$ on the set of admissible actions, which is known as a *policy.* An agent following policy γ chooses her next action according to $\gamma_s(\cdot)$ whenever the system is in state s, independently of her previous actions.

It is easy to see that, in conjunction with the transition distributions $\mathcal{P}$, every policy γ induces a collection of Markov chains on the state space $\mathcal{S}$ indexed by initial states $s \in \mathcal{S}$. In the average-cost

MDP problem the goal is to identify a policy for which the induced Markov chain minimizes the limiting average cost, namely a policy γ minimizing

$$\bar{c}_\gamma(s) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_\gamma \left[c'(s_t, a_t) \middle| s_0 = s \right], \tag{3.3}$$

for each $s \in \mathcal{S}$. Note that the expectation in (3.3) is taken with respect to the Markov chain induced by γ. In general, the limiting average cost $\bar{c}_\gamma(s)$ will depend on the initial state s, but if γ induces an ergodic chain then the average cost will be constant. If all policies induce ergodic Markov chains, the MDP is referred to as "unichain"; otherwise the MDP is classified as "multichain". We refer the reader to (Puterman, 2005) for more details on MDPs.

The OTC problem (3.2) may readily be recast as an MDP. In detail, let the state space $\mathcal{S} = \mathcal{X} \times \mathcal{Y}$, and let $s = (x, y)$ denote an element of $\mathcal{S}$. Define the set of admissible actions in state s to be the corresponding set of row couplings $\mathcal{A}_s = \Pi(P(x, \cdot), Q(y, \cdot))$. For each state s and action $r_s \in \mathcal{A}_s$ define the transition distribution $p(\cdot|s, r_s) := r_s(\cdot)$, and the cost function $c'(s, r_s) = c(s) = c(x, y)$. Note that c' is independent of the action r_s. We refer to this MDP as TC-MDP.

Any policy γ for TC-MDP specifies distributions over $\Pi(P(x, \cdot), Q(y, \cdot))$ for each $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and thus corresponds to a single distribution over $\Pi_{\mathrm{TC}}(P, Q)$ that governs the random actions of the agent. In TC-MDP it suffices to consider only deterministic policies γ, namely policies such that for each state $s = (x, y)$ the distribution $\gamma_s(\cdot)$ is a point mass at unique element of $\mathcal{A}_s = \Pi(P(x, \cdot), Q(y, \cdot))$.

Proposition 3.5. *Let γ be a policy for TC-MDP. Then there exists a deterministic policy $\tilde{\gamma}$ such that $\bar{c}_\gamma(s) = \bar{c}_{\tilde{\gamma}}(s)$ for every $s \in \mathcal{S}$.*

Thus optimization over $\Pi_{\mathrm{TC}}(P, Q)$ is equivalent to optimization over deterministic policies. Importantly, *a deterministic policy corresponds to a fixed transition coupling matrix $R \in \Pi_{\mathrm{TC}}(P, Q)$*. Going forward, we refer to $R \in \Pi_{\mathrm{TC}}(P, Q)$ directly instead of the equivalent deterministic policy $\tilde{\gamma}$ in our notation. We note that, even when X and Y are ergodic, the same may not be true of the stationary Markov chain induced by a transition coupling matrix $R \in \Pi_{\mathrm{TC}}(P, Q)$ (see Appendix A.2). Specifically, a single element of $\Pi_{\mathrm{TC}}(P, Q)$ may have multiple stationary distributions and thus give rise to multiple stationary Markov chains depending on the initial state $s \in \mathcal{S}$. Thus

TC-MDP is classified as multichain. Finally, we may formalize the relationship between the OTC problem and TC-MDP.

Proposition 3.6. *If X and Y are irreducible, then any $R \in \Pi(P, Q)$ that is an optimal policy for TC-MDP corresponds to an optimal coupling $\pi_R \in \Pi_{TC}(\mathbb{P}, \mathbb{Q})$ with expected cost* $\min_{s \in \mathcal{S}} \overline{c}_R(s)$.

3.4.2 Policy Iteration

Now that we have shown that the OTC problem can be viewed as an MDP, we can leverage existing algorithms for MDPs to obtain solutions. To this end, we propose to adapt the framework of policy iteration (Howard, 1960). To facilitate our discussion, in what follows, we regard the cost function c and limiting average cost $\overline{c}_R$ as vectors in $\mathbb{R}_+^{d^2}$. For each $R \in \Pi_{\text{TC}}(P, Q)$, standard results (Puterman, 2005) guarantee that the limit $\overline{R} := \lim_{T\to\infty} T^{-1} \sum_{t=0}^{T-1} R^t$ exists. When R is aperiodic and irreducible, the Perron-Frobenius theorem ensures that $\overline{R} = \lim_{T\to\infty} R^T$ and the rows of $\overline{R}$ are equal to the stationary distributions of R.

In policy iteration, one repeatedly evaluates and improves policies. In the context of TC-MDP, for a given transition coupling matrix $R \in \Pi_{\text{TC}}(P, Q)$ the evaluation step computes the average cost (*gain*) vector $g = \overline{R}\, c$ and the total extra cost (*bias*) vector $h = \sum_{t=0}^{\infty} R^t(c - g)$. In practice, g and h may be obtained by solving a linear system of equations rather than evaluating infinite sums (see Algorithm 1a) (Puterman, 2005). The improvement step selects a new transition coupling matrix R' that minimizes $R'\, g$ or, if no improvement is possible, $R'\, h$ in an element-wise fashion (see Algorithm 1b). In more detail, we may select a transition coupling R' such that for each (x, y) the corresponding row $r = R'((x, y), \cdot)$ minimizes rg (or rh) over couplings $r \in \Pi(P(x, \cdot), Q(y, \cdot))$. To denote the element-wise argmin, we write $\text{elem-argmin}_{R \in \Pi_{\text{TC}}(P,Q)}\, R\, g$ (or $R\, h$). The improved matrix R' is obtained by solving d^2 OT problems with marginals $P(x, \cdot)$ and $Q(y, \cdot)$ and cost g (or h). This special feature of TC-MDP enables us to find improved transition coupling matrices in a computationally efficient manner despite working with an infinite action space. Once a fixed point in the evaluation and improvement process is reached, the procedure terminates. The resulting algorithm will be referred to as `ExactOTC` (see Algorithm 1). We initialize Algorithm 1 to the independent transition coupling $P \otimes Q$, defined in Section 3.2.

```
Algorithm 1: ExactOTC
$R_0 \leftarrow P \otimes Q$, $n \leftarrow 0$
while not converged do
    /* transition coupling evaluation */
    $(g_n, h_n) \leftarrow$ ExactTCE($R_n$)
    /* transition coupling improvement */
    $R_{n+1} \leftarrow$ ExactTCI($g_n, h_n, R_n, \Pi_{\mathrm{TC}}(P, Q)$)
    $n \leftarrow n + 1$
return $R_n$
```

```
Algorithm 1a: ExactTCE
input: $R$
Solve for $(g, h, w)$ such that
    $\begin{bmatrix} I-R & 0 & 0 \\ I & I-R & 0 \\ 0 & I & I-R \end{bmatrix} \begin{bmatrix} g \\ h \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ c \\ 0 \end{bmatrix}$
return $(g, h)$
```

```
Algorithm 1b: ExactTCI
input: $g, h, R_0, \Pi$
/* element-wise argmin
    */
$R' \leftarrow \text{elem-argmin}_{R \in \Pi} Rg$
if $R'g = R_0 g$ then
    $R' \leftarrow \text{elem-argmin}_{R \in \Pi} Rh$
    if $R'h = R_0 h$ then
        return $R_0$
    else
        return $R'$
else
    return $R'$
```

For finite state and action spaces, policy iteration is known to yield an optimal policy for the average-cost MDP in a finite number of steps (Puterman, 2005). While policy iteration may fail to converge for general compact action spaces (Dekker, 1987; Schweitzer, 1985; Puterman, 2005), as is the case for TC-MDP, we may exploit the polyhedral structure of $\Pi_{\mathrm{TC}}(P, Q)$ to establish the following convergence result.

Theorem 3.7. *Algorithm 1 converges to a solution (g^*, h^*, R^*) of TC-MDP in a finite number of iterations. Moreover, if X and Y are irreducible, R^* is the transition matrix of an optimal transition coupling of X and Y.*

Recall from Proposition 3.6 that an optimal solution to TC-MDP necessarily yields an optimal solution to (3.2). Thus Theorem 3.7 ensures that a solution to the OTC problem can be obtained from Algorithm 1 in a finite number of iterations. A proof of this result can be found in Section 3.9.3.3.

Remark 3.8. One may in principle adapt other MDP algorithms to solve the OTC problem. However, the standard alternatives to policy iteration either do not admit a computationally tractable implementation (e.g. linear programming) or are not as conducive to a convergence analysis (e.g. value iteration). We choose policy iteration because it balances both of these features, admitting a practical implementation while also enabling a theoretical convergence analysis. We acknowledge that OTC solvers based on policy iteration may not be preferable in every scenario and leave a detailed exploration of other MDP algorithms for the OTC problem to future work.

3.5 Fast Approximate Policy Iteration

The simplicity of Algorithm 1 in conjunction with the theoretical guarantee of Theorem 3.7 make it an appealing method for solving the OTC problem when the cardinality d of the state spaces of X and Y is small. However, each call to Algorithm 1a involves solving a system of $3d^2$ linear equations, requiring a total of $\mathcal{O}(d^6)$ operations. Furthermore, each call to Algorithm 1b entails solving d^2 linear programs each with $\mathcal{O}(d)$ constraints, which can be accomplished in a total time of $\mathcal{O}(d^5 \log d)$. We note that a similar dependence on the dimension of each coupling is observed in exact OT algorithms, such as the network simplex algorithm in (Peyré and Cuturi, 2019). For even moderate values of d, this may be too slow for practical use.

To alleviate the poor scaling with the dimension of the couplings in the standard OT problem, one may use entropic regularization, whereby a negative entropy term is added to the OT objective. Briefly, we review some related work regarding entropic regularization in OT from Chapter 2. (Cuturi, 2013) showed that solutions to the entropy-regularized OT problem may be obtained efficiently via Sinkhorn's algorithm (Sinkhorn, 1967). More recently, (Altschuler et al., 2017) proved that Sinkhorn's algorithm yields an approximation of the OT cost with error bounded by ε in near-linear time with respect to the dimension of the couplings under consideration. Subsequent work (Dvurechensky et al., 2018; Lin et al., 2019; Guo et al., 2020) has proposed and studied alternative algorithms for approximating the optimal transport cost, each with runtime scaling at least linearly with the dimension of the couplings in the problem. One might hope that a similar dependence on the size of the elements of $\Pi_{\text{TC}}(P, Q)$ may be achievable for the OTC problem by employing regularization.

In this section, we extend entropic regularization techniques to the OTC problem. This extension leads to an approximate algorithm that runs in $\tilde{\mathcal{O}}(d^4)$ time per iteration, where $\tilde{\mathcal{O}}(\cdot)$ omits non-leading poly-logarithmic factors. This complexity is nearly-linear in the dimension d^4 of the transition couplings. We first propose a truncation-based approximation of the `ExactTCE` transition coupling evaluation algorithm, which we call `ApproxTCE`. When the transition coupling to be evaluated satisfies a simple regularity condition, we show that one can obtain approximations of the gain and bias from `ApproxTCE` with error bounded by ε in $\tilde{\mathcal{O}}(d^4 \log \varepsilon^{-1})$ time.

Mirroring the derivation of entropic OT, we then propose an entropy-regularized approximation of the `ExactTCI` transition coupling improvement algorithm, called `EntropicTCI`. We perform a new analysis of the Sinkhorn algorithm (described in Section 3.5.3) that is tailored to transition coupling improvement to show that `EntropicTCI` yields an improved transition coupling with error bounded by ε in $\tilde{\mathcal{O}}(d^4\varepsilon^{-4})$ time. Combining these two algorithms, we obtain the `EntropicOTC` algorithm, which runs in $\tilde{\mathcal{O}}(d^4\varepsilon^{-4})$ time per iteration. We provide empirical support for these theoretical results through a simulation study in Section 3.7. We find that the improved efficiency at each iteration of `EntropicOTC` leads to a much faster runtime in practice as compared to `ExactOTC`. Our experiments also show that `EntropicOTC` yields an expected cost that closely approximates the unregularized OTC cost.

3.5.1 Constrained Optimal Transition Coupling Problem

We begin by defining a constrained set of transition couplings. Let $\mathcal{K}(\cdot\|\cdot)$ be the Kullback-Leibler (KL) divergence defined for $u, v \in \Delta_{d^2}$ by $\mathcal{K}(u\|v) = \sum_s u(s)\log(u(s)/v(s))$ with the convention that $0\log(0/0) = 0$ and $\mathcal{K}(u\|v) = +\infty$ if $u(s) > 0$ and $v(s) = 0$ for some index s. For every $\eta > 0$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, define the set

$$\Pi_\eta(P(x,\cdot), Q(y,\cdot)) = \left\{r \in \Pi(P(x,\cdot), Q(y,\cdot)) : \mathcal{K}\big(r \| P \otimes Q((x,y),\cdot)\big) \leq \eta\right\},$$

and the subset of transition coupling matrices

$$\Pi^\eta_{\mathrm{TC}}(P, Q) = \{R \in \Pi_{\mathrm{TC}}(P, Q) : R((x,y),\cdot) \in \Pi_\eta(P(x,\cdot), Q(y,\cdot)),\ \forall (x,y) \in \mathcal{X} \times \mathcal{Y}\}.$$

Elements of $\Pi^\eta_{\mathrm{TC}}(P, Q)$ have rows that are close in KL-divergence to the rows of the independent transition coupling $P \otimes Q$. When P and Q are aperiodic and irreducible, the same is true of $P \otimes Q$.

Fix $\eta > 0$ and let $\Pi_{\text{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})$ be the set of transition couplings with transition matrices in $\Pi_{\text{TC}}^{\eta}(P, Q)$. The *entropic OTC problem* is

$$\begin{aligned} \text{minimize} \quad & \int c \, d\pi \\ \text{subject to} \quad & \pi \in \Pi_{\text{TC}}^{\eta}(\mathbb{P}, \mathbb{Q}). \end{aligned} \tag{3.4}$$

For completeness, we establish in Appendix A.1 that a solution to (3.4) exists. As the divergence $\mathcal{K}(r \| P \otimes Q(s, \cdot))$ is bounded for $r \in \Pi(P(x, \cdot), Q(y, \cdot))$ and $s = (x, y) \in \mathcal{X} \times \mathcal{Y}$ (Cuturi, 2013), the program (3.4) coincides with the unconstrained OTC problem for sufficiently large η. Finally, note that (3.4) corresponds to an MDP in the same way that (3.2) does but with a constrained set of policies. In the rest of the section, we develop computationally efficient alternatives to Algorithms 1a and 1b for this constrained MDP.

3.5.2 Fast Approximate Transition Coupling Evaluation

Next, we propose a fast approximation of Algorithm 1a. Recall from our previous discussion that the gain vector g corresponding to any aperiodic and irreducible $R \in \Pi_{\text{TC}}(P, Q)$ is constant and thus may be written as $g = g_0 \mathbb{1}$ for a scalar g_0. Fixing such an $R \in \Pi_{\text{TC}}(P, Q)$ and $L, T \geq 1$, we approximate the gain g by averaging the cost over L steps of the Markov chain corresponding to R from each possible starting point in $\mathcal{X} \times \mathcal{Y}$. Moreover, we approximate the bias h by summing the total extra cost over T steps with respect to the approximate gain $\tilde{g}$. Formally, let $\tilde{g} := (d^{-2}(R^L c)^\top \mathbb{1})\mathbb{1}$ and $\tilde{h} := \sum_{t=0}^{T} R^t (c - \tilde{g})$. The resulting algorithm, which we refer to as `ApproxTCE`, is detailed in Algorithm 2a.

```
Algorithm 2a: ApproxTCE
  input: $R$, $L$, $T$
  $\tilde{g} \leftarrow (d^{-2}(R^L c)^\top \mathbb{1})\mathbb{1}$
  $\tilde{h} \leftarrow \sum_{t=0}^{T} R^t (c - \tilde{g})$
  return $(\tilde{g}, \tilde{h})$
```

The approximations $\tilde{g}$ and $\tilde{h}$ can be computed in $\mathcal{O}(Ld^4)$ and $\mathcal{O}(Td^4)$ time, respectively. Since g and h are equal to the limits of $\tilde{g}$ and $\tilde{h}$ as $L, T \to \infty$, we expect that larger L and T will yield better approximations. One must ensure that the L and T that are required for a good approximation do not grow too quickly with d. We show that this is the case in Theorem 3.9 below. We will say that a transition matrix $R \in [0,1]^{d^2 \times d^2}$ with stationary distribution $\lambda \in \Delta_{d^2}$ is mixing with coefficients $M \in \mathbb{R}_+$ and $\alpha \in [0,1)$ if for every $t \in \mathbb{N}$, $\max_{s \in \mathcal{X} \times \mathcal{Y}} \|R^t(s,\cdot) - \lambda\|_1 \leq M\alpha^t$. Recall that R is mixing whenever it is aperiodic and irreducible.

Theorem 3.9. *Let $R \in \Pi_{TC}(P,Q)$ be aperiodic and irreducible with mixing coefficients $M \in \mathbb{R}_+$ and $\alpha \in [0,1)$ and gain and bias vectors $g \in \mathbb{R}^{d^2}$ and $h \in \mathbb{R}^{d^2}$, respectively. Then for any $\varepsilon > 0$, there exist $L, T \in \mathbb{N}$ such that* `ApproxTCE`(R, L, T) *yields $(\tilde{g}, \tilde{h})$ satisfying $\|\tilde{g} - g\|_\infty \leq \varepsilon$ and $\|\tilde{h} - h\|_1 \leq \varepsilon$ in $\tilde{\mathcal{O}}\left(\frac{d^4}{\log \alpha^{-1}} \log\left(\frac{M}{\varepsilon(1-\alpha)}\right)\right)$ time.*

In particular, `ApproxTCE` does approximate `ExactTCE` in time scaling like $\tilde{\mathcal{O}}(d^4)$. Explicit choices of L and T are given in the proof of Theorem 3.9, which may be found in Section 3.9.4.

Remark 3.10. In practice, values of L and T satisfying the conclusion of Theorem 3.9 are unknown. In our experiments, we found that running Algorithm 2a with large, fixed values of L and T yields a high approximation accuracy while still running significantly more quickly than Algorithm 1a. Alternatively, L and T may be chosen adaptively by computing vectors $\tilde{g}$ and $\tilde{h}$ iteratively for larger and larger values of L and T until some convergence criterion is satisfied or L and T hit some prespecified thresholds. For example, letting $\tilde{g}^L$ and $\tilde{h}^T$ be the iterates of this procedure, one may iterate until $\|\tilde{g}^L - \tilde{g}^{L-1}\|_\infty < \varepsilon$ and $\|\tilde{h}^T - \tilde{h}^{T-1}\|_\infty < \varepsilon$. This approach achieves the same worst-case complexity as Algorithm 2a but allows for time-savings when the chain R mixes quickly.

For a set $\mathcal{U} \subset \mathbb{R}^n$, let $B_\varepsilon(u) \subset \mathbb{R}^n$ be the open ball of radius $\varepsilon > 0$ centered at $u \in \mathcal{U}$, and let $\mathrm{aff}(\mathcal{U})$ denote the affine hull, defined as $\mathrm{aff}(\mathcal{U}) = \{\sum_{i=1}^k \alpha_i u_i : k \in \mathbb{N}, u_1, ..., u_k \in \mathcal{U}, \sum_{i=1}^k \alpha_i = 1\}$. Let $\mathrm{ri}(\cdot)$ denote the relative interior, defined as $\mathrm{ri}(\mathcal{U}) = \{u \in \mathcal{U} : \exists \varepsilon > 0 \text{ s.t. } B_\varepsilon(u) \cap \mathrm{aff}(\mathcal{U}) \subset \mathcal{U}\}$.

Proposition 3.11. *If P and Q are aperiodic and irreducible then every $R \in ri(\Pi_{TC}(P,Q))$ is also aperiodic and irreducible, and thus mixing.*

As a consequence of Proposition 3.11, we need only verify that $R \in \mathrm{ri}(\Pi_{\mathrm{TC}}(P,Q))$ to ensure that Theorem 3.9 holds and that we may perform fast transition coupling evaluation via `ApproxTCE`. As we show in Theorem 3.12, this condition is naturally guaranteed when employing entropic OT techniques for speeding up the transition coupling improvement step.

3.5.3 Entropic Transition Coupling Improvement

Next we describe a means of speeding up Algorithm 1b. For the MDP corresponding to the entropic OTC problem, exact policy improvement can be performed by calling `ExactTCI` with $\Pi = \Pi_{\mathrm{TC}}^{\eta}(P, Q)$. However, no computation time is saved by doing this. Instead, we settle for an algorithm that yields approximately improved transition couplings with better computational efficiency. To find such an approximation, we reconsider the linear optimization problems that comprise the transition coupling improvement step. Namely, for each $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned} &\text{minimize} \quad \langle r, h\rangle \\ &\text{subject to} \quad r \in \Pi_{\eta}(P(x, \cdot), Q(y, \cdot)). \end{aligned} \tag{3.5}$$

By standard arguments, (3.5) is equivalent to

$$\begin{aligned} &\text{minimize} \quad \langle r, h\rangle + \frac{1}{\xi}\sum_{s'} r(s')\log r(s') \\ &\text{subject to} \quad r \in \Pi(P(x, \cdot), Q(y, \cdot)), \end{aligned} \tag{3.6}$$

for some $\xi \in [0, \infty]$ depending on (x, y), η and h. The reformulation (3.6) suggests that one use computational techniques for entropic OT in the place of linear programming to perform transition coupling improvement for the constrained OTC problem. In particular, we use the `ApproxOT` algorithm of (Altschuler et al., 2017), detailed in Appendix 3.9.4. Using `ApproxOT` instead of solving (3.6) exactly, we obtain the `EntropicTCI` algorithm detailed in Algorithm 2b.

Algorithm 2b: `EntropicTCI`

```
input: $h, \xi, \varepsilon$
for $(x, y) \in \mathcal{X} \times \mathcal{Y}$ do
    $R(s, \cdot) \leftarrow$ ApproxOT($P(x, \cdot)^{\top}, Q(y, \cdot)^{\top}, h, \xi, \varepsilon$)
return $R$
```

To provide further intuition for Algorithm 2b, it is helpful to consider the constrained OTC problem from an alternate perspective. For a probability measure $r \in \Delta_{d^2}$, let $H(r) = -\sum_s r(s)\log r(s)$ be its entropy. Then by duality theory, the constrained OTC problem (3.4) may be written as the finite-dimensional optimization problem

$$\begin{aligned} \text{minimize} \quad & \langle c, \lambda \rangle - \sum_s \frac{1}{\xi(s)} H(R(s, \cdot)) \\ \text{subject to} \quad & R \in \Pi_{\text{TC}}(P, Q) \\ & \lambda R = \lambda \\ & \langle \mathbb{1}, \lambda \rangle = 1, \end{aligned} \tag{3.7}$$

for some $\xi \in [0, \infty]^{d^2}$. In order to solve the problem above, we study its Lagrangian. Let $\alpha, \beta \in \mathbb{R}^{d^3}$, $\gamma \in \mathbb{R}^{d^2}$, and $\delta \in \mathbb{R}$ be Lagrange multipliers. The Lagrangian may be written as

$$\begin{aligned} \mathcal{L}(R, \lambda, \alpha, \beta, \gamma, \delta) = \langle c, \lambda \rangle & - \sum_{x,y} \frac{1}{\xi(x,y)} H(R((x,y), \cdot)) \\ & + \sum_{x,y,x'} \alpha(x,y,x') \left(\sum_{y'} R((x,y),(x',y')) - P(x,x') \right) \\ & + \sum_{x,y,y'} \beta(x,y,y') \left(\sum_{x'} R((x,y),(x',y')) - Q(y,y') \right) \\ & + \sum_{x',y'} \gamma(x',y') \left(\sum_{x,y} \lambda(x,y) R((x,y),(x',y')) - \lambda(x',y') \right) \\ & + \delta \left(\sum_{x,y} \lambda(x,y) - 1 \right). \end{aligned}$$

Taking the partial derivative of $\mathcal{L}$ with respect to $R((x,y),(x',y'))$ and setting it equal to zero, we find that

$$R(s,(x',y')) = \exp\left\{ -\xi(s)\alpha(s,x') - \frac{1}{2} \right\} \exp\left\{ -\xi(s)\lambda(s)\gamma(x',y') \right\} \exp\left\{ -\xi(s)\beta(s,y') - \frac{1}{2} \right\},$$

where we have used $s = (x,y)$ to reduce notation. When viewed as a $d \times d$ matrix, $R((x,y), \cdot)$ can be written as UKV where U and V are both non-negative diagonal matrices. Note that when $\xi(x,y) < \infty$, this implies that R is aperiodic and irreducible since R lies in the relative interior of $\Pi_{\text{TC}}(P,Q)$ (see Theorem 3.12).

A similar matrix form appears in the analysis of (Cuturi, 2013). An important difference is the matrix $K \in \mathbb{R}_+^{d \times d}$, which satisfies

$$K(x', y') = \exp\left\{-\xi(x, y)\lambda(x, y)\gamma(x', y')\right\}.$$

In (Cuturi, 2013), one finds that $K = e^{-\xi C}$ where C is the cost matrix. To better understand this difference, it is helpful to look at the partial derivative of the Lagrangian with respect to $\lambda(x, y)$. Evaluating this partial derivative and setting it equal to zero, we find

$$\gamma(x, y) = c(x, y) + \sum_{x', y'} R((x, y), (x', y'))\gamma(x', y') + \delta.$$

Absorbing the scalar δ into c to obtain an augmented cost $\tilde{c} = c + \delta$, we have

$$\gamma(x, y) = \tilde{c}(x, y) + \sum_{x', y'} R((x, y), (x', y'))\gamma(x', y').$$

Letting g be the gain of the policy R with respect to $\tilde{c}$, we recognize that the equation above is the Bellman recursion for the bias of R with respect to the cost $\tilde{c}+g$. As R is aperiodic and irreducible, g is a constant vector. Moreover, as the bias is invariant under constant shifts in cost, γ is exactly the bias h that appears in `EntropicTCI`. Returning to the form of $R((x, y), \cdot)$ established earlier, we find that

$$R((x, y), \cdot) = U \exp\left\{-\xi(x, y)\lambda(x, y)h\right\} V = U \exp\{-\tilde{\xi}(x, y)h\}V,$$

for non-negative diagonal matrices U and V and the constant $\tilde{\xi}(x, y) := \xi(x, y)\lambda(x, y)$. In this way, the bias h plays the role of the cost matrix C of (Cuturi, 2013).

In order to solve the program (3.7), one must grapple with the interdependence between the bias h and the policy R. A natural approach for doing so is to consider an alternating optimization algorithm in which one repeatedly solves for the bias h from a given policy R, then solves for a new policy R given the bias h. Indeed, this is the procedure one follows in `ExactOTC`. In practice, given a policy R, one approximately computes the bias h (`ApproxTCE`) in order to save time. Given a bias vector h, one solves for a new policy R by performing Sinkhorn iterations with the bias h as a cost matrix for each $R((x, y), \cdot)$ (`EntropicTCI`).

It was shown in (Altschuler et al., 2017) that ApproxOT yields an approximation of the OT cost in near-linear time with respect to the size of the couplings of interest. However, in order to control the approximation error of EntropicTCI, we rely on a different analysis showing that one can obtain an approximation of the entropic optimal coupling in near-linear time (see Lemma 3.18) . To the best of our knowledge, this result does not exist in the literature, so we provide a proof in Section 3.9.4. Using this result, we show the complexity bound below.

Theorem 3.12. *Let P and Q be aperiodic and irreducible, $h \in \mathbb{R}^{d^2}$, $\xi > 0$, and $\varepsilon > 0$. Then* EntropicTCI(h, ξ, ε) *returns $\hat{R} \in ri(\Pi_{TC}(P, Q))$ with $\max_s \|\hat{R}(s, \cdot) - R^*(s, \cdot)\|_1 \leq \varepsilon$ for some $R^* \in \operatorname{argmin}_{R' \in \Pi_{TC}(P,Q)} R'h - {}^1\!/\!_\xi H(R')$ in $\tilde{\mathcal{O}}(d^4\varepsilon^{-4})$ time.*

To summarize, this result states that EntropicTCI yields an approximately improved transition coupling in $\tilde{\mathcal{O}}(d^4)$ time rather than $\tilde{\mathcal{O}}(d^5)$ as previously discussed. In practice, further speedups are possible by utilizing the fact that the d^2 entropic OT problems to be solved are decoupled and thus may be computed in parallel.

3.5.4 EntropicOTC

Finally, using Algorithms 2a and 2b, we define the EntropicOTC algorithm, detailed in Algorithm 2. Essentially, EntropicOTC is defined by replacing ExactTCE and ExactTCI by the efficient alternatives, ApproxTCE and EntropicTCI. As stated in Theorem 3.12, EntropicTCI returns transition couplings in the relative interior of $\Pi_{\text{TC}}(P, Q)$, so the iterates of EntropicOTC are not restricted to the finite set of extreme points of $\Pi_{\text{TC}}(P, Q)$. Thus, convergence for Algorithm 2 must be assessed differently than in Algorithm 1. In our simulations we found that the element-wise inequality $\tilde{g}_{n+1} \geq \tilde{g}_n$ works well as an indicator of convergence.

```
Algorithm 2: EntropicOTC
  input: $L, T, \xi, \varepsilon$
  $n \leftarrow 0$
  while $n = 0$ or $\tilde{g}_{n+1} < \tilde{g}_n$ do
      /* transition coupling evaluation */
      $(\tilde{g}_n, \tilde{h}_n) \leftarrow$ ApproxTCE($R_n, L, T$)
      /* transition coupling improvement */
      $R_{n+1} \leftarrow$ EntropicTCI($\tilde{h}_n, \xi, \varepsilon$)
      $n \leftarrow n + 1$
  return $R_{n+1}$
```

3.6 Consistency

The computational and theoretical results presented above assume that one has complete knowledge of the transition matrices P and Q of the Markov chains X and Y under study. In practice, one may not have direct access to P and Q, but may instead have estimates $\hat{P}_n$ and $\hat{Q}_n$ derived from n observations of the chains X and Y. In the simplest case, $\hat{P}_n$ and $\hat{Q}_n$ may be obtained from the observed relative frequencies of each transition between states. In Theorem 3.13 below, we show that the cost and solution sets of the standard and regularized optimal transition coupling problems possess natural stability properties with respect to the marginal transition matrices. As a corollary, we obtain a consistency result for the OTC problem applied to the estimates $\hat{P}_n$ and $\hat{Q}_n$.

Recall that we use Δ_d to denote the probability simplex in $\mathbb{R}^d$ and note that the set of $d \times d$-dimensional transition matrices may be written as Δ_d^d. Likewise, the set of $d^2 \times d^2$-dimensional transition matrices may be written as $\Delta_{d^2}^{d^2}$. Note that we endow the sets of $d \times d$- and $d^2 \times d^2$-dimensional transition matrices with the topologies they inherit as subsets of $\mathbb{R}^{d \times d}$ and $\mathbb{R}^{d^2 \times d^2}$, respectively, and adopt the same convention for the set $\Delta_{d^2} \times \Delta_{d^2}^{d^2}$. Now, we may reformulate Problems (3.2) and (3.4) as follows:

$$\begin{aligned} \text{minimize} \quad & \langle c, \lambda \rangle \\ \text{subject to} \quad & R \in \Pi(P, Q) \\ & \lambda R = \lambda \\ & \lambda \in \Delta_{d^2}. \end{aligned} \qquad \text{(I)} \qquad\qquad \begin{aligned} \text{minimize} \quad & \langle c, \lambda \rangle \\ \text{subject to} \quad & R \in \Pi_\eta(P, Q) \\ & \lambda R = \lambda \\ & \lambda \in \Delta_{d^2}. \end{aligned} \qquad \text{(II)}$$

Let $\rho(P,Q)$ and $\rho_\eta(P,Q)$ denote the optimal values of Problems (I) and (II), respectively, and let $\Phi^*(P,Q)$ and $\Phi^*_\eta(P,Q)$ denote the associated sets of optimal solutions $(\lambda, R) \in \Delta_{d^2} \times \Delta_{d^2}^{d^2}$ to Problems (I) and (II), respectively. For metric spaces $\mathcal{U}$ and $\mathcal{Z}$, we will say that a function $F : \mathcal{U} \to 2^{\mathcal{Z}}$ is upper semicontinuous at a point $u_0 \in \mathcal{U}$ if for any neighborhood V of $F(u_0)$, there exists a neighborhood U of u_0 such that $F(u) \subset V$ for every $u \in U$.

Theorem 3.13. *Let $P, Q \in \Delta_d^d$ be irreducible transition matrices. Then the following hold:*

- *$\rho(\cdot,\cdot)$ is continuous and $\Phi^*(\cdot,\cdot)$ is upper semicontinuous at (P,Q)*
- *For any $\eta > 0$, $\rho_\eta(\cdot,\cdot)$ is continuous and $\Phi^*_\eta(\cdot,\cdot)$ is upper semicontinuous at (P,Q)*

Theorem 3.13 states that the optimal values and optimal solution sets of the OTC and entropic OTC problems are stable in the marginal transition matrices P and Q. We may use this result to prove a consistency result for either problem when applied to estimates $\hat{P}_n$ and $\hat{Q}_n$ derived from data. In stating the following result, we make use of the following definition: For a sequence of sets $\{A_n\}_{n\geq 0}$ in a topological space $\mathcal{A}$, let $\limsup_{n\to\infty} A_n = \bigcap_{n=0}^\infty \mathrm{cl}\,(\bigcup_{m=n}^\infty A_m)$, where $\mathrm{cl}(\cdot)$ denotes the closure with respect to topology of $\mathcal{A}$. Note that the presence of $\mathrm{cl}(\cdot)$ in our definition of limit superior of a sequence of sets differs from that commonly used in probability but is consistent with the definition appearing, for example, in (Rockafellar and Wets, 2009).

Corollary 3.14. *Let $X = \{X_i\}_{i\geq 0}$ and $Y = \{Y_i\}_{i\geq 0}$ be stationary, ergodic processes taking values in $\mathcal{X}$ and $\mathcal{Y}$, and defined on a common Borel probability space. Suppose further that X and Y have marginal, one-step transition matrices P and Q, respectively. Let $\hat{P}_n$ and $\hat{Q}_n$ be the one-step transition matrices estimated via relative frequencies from the sequences $X_0, ..., X_{n-1}$ and $Y_0, ..., Y_{n-1}$. Then with probability one, the following hold:*

- $\rho(\hat{P}_n, \hat{Q}_n) \to \rho(P,Q)$ *and* $\limsup_{n\to\infty} \Phi^*(\hat{P}_n, \hat{Q}_n) \subseteq \Phi^*(P,Q)$
- *For any $\eta > 0$,* $\rho_\eta(\hat{P}_n, \hat{Q}_n) \to \rho_\eta(P,Q)$ *and* $\limsup_{n\to\infty} \Phi^*_\eta(\hat{P}_n, \hat{Q}_n) \subseteq \Phi^*_\eta(P,Q)$

Corollary 3.14 allows us to apply the computational tools described above to real data in a principled manner. In particular, when the marginal transition matrices P and Q are unknown, we may use $\hat{P}_n$ and $\hat{Q}_n$ as proxies in the OTC problem to estimate the set of optimal transition couplings and their expected cost when n is large. Note that we do not require the generating processes themselves to be Markov: they need only be stationary and ergodic, so that the estimates $\hat{P}_n$ and $\hat{Q}_n$ converge to the true one-step transition matrices P and Q as n tends to infinity.

3.7 Experiments

In this section, we validate the proposed algorithms empirically by applying them to stationary Markov chains derived from both synthetic and real data. We begin by comparing the runtime of the proposed algorithms and approximation error of `EntropicOTC` via a simulation study. Subsequently, we illustrate the potential use of the OTC problem in practice through an application to computer-generated music.

We remark that an application of the OTC problem to graphs is studied in Chapter 4. In particular, a weighted graph may be associated with a stationary Markov chain by means of a simple random walk on its nodes with transition probabilities proportional to its edge weights. Leveraging this perspective, we propose to perform OT on the graphs of interest by applying the OTC problem to their associated Markov chains. In the aforementioned work, we demonstrate that this approach performs on par with state-of-the-art graph OT methods in a variety of graph comparison and alignment tasks on real and synthetic data.

`Matlab` implementations of `ExactOTC` and `EntropicOTC` as well as code for reproducing the experimental results to follow are available at `https://github.com/oconnor-kevin/OTC`. For `ApproxOT` and related OT algorithms, we used the implementation found at `https://github.com/JasonAltschuler/OptimalTransportNIPS17`.

3.7.1 Simulation Study

In order to validate the use of Algorithm 2 as a fast alternative to Algorithm 1, we performed a simulation study to compare their runtimes and the error of the entropic OTC cost as an approximation of the OTC cost. For each choice of the marginal state space size $d \in \{10, 20, ..., 100\}$, we perform five simulations, obtaining estimates of the runtimes and approximation error in each. In each simulation, we generate transition matrices $P \in [0,1]^{d\times d}$ and $Q \in [0,1]^{d\times d}$ and a cost matrix $c \in \mathbb{R}_+^{d\times d}$ by drawing each element of the matrix of interest independently from a standard normal distribution and then applying an appropriate normalization to the matrix. In the case of the

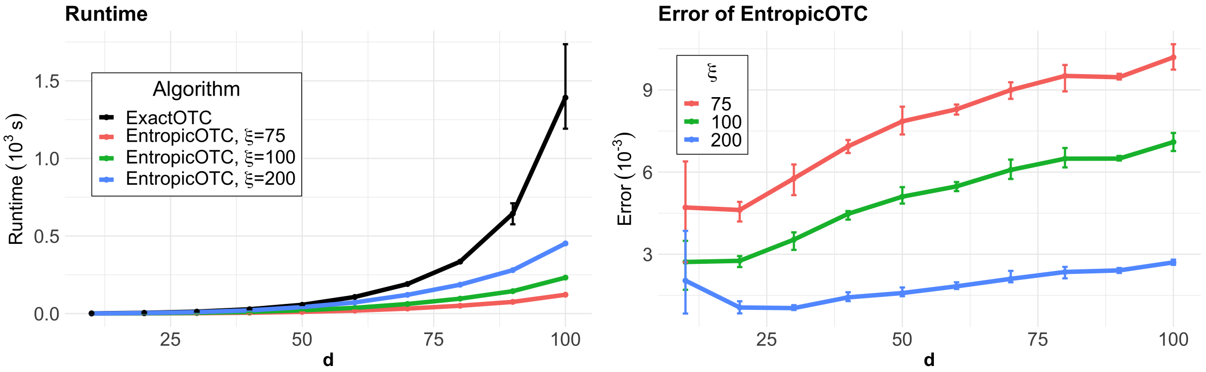


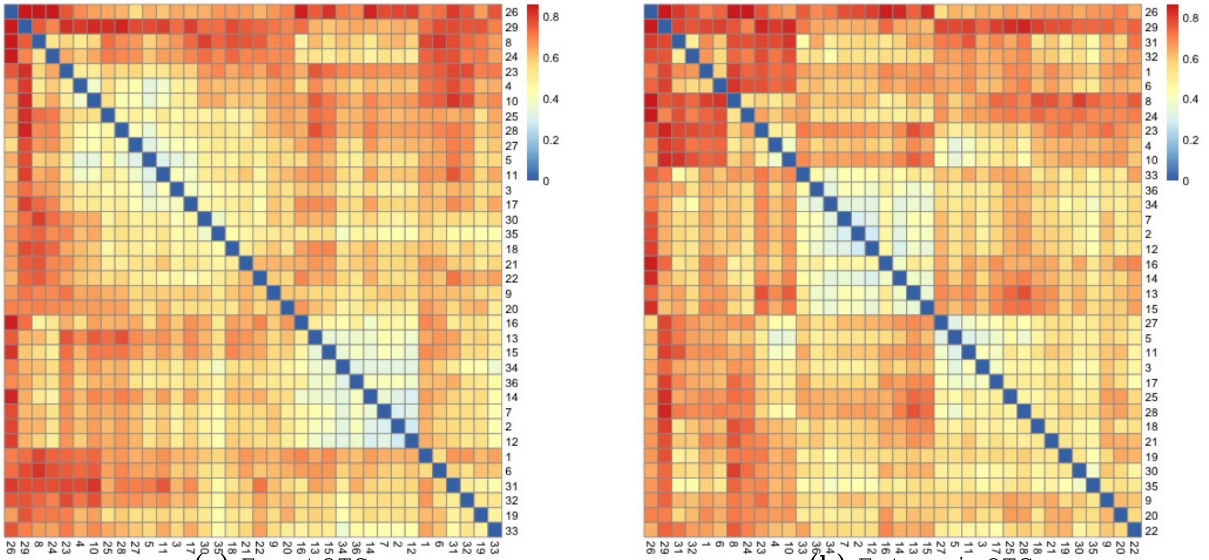
Figure 3.1: A comparison of total runtimes between ExactOTC and EntropicOTC and approximation errors of EntropicOTC for a range of d and ξ via simulation. Error bars show the minimum and maximum values observed over five simulations. Note that the error bars for the runtimes of EntropicOTC are not visible because little variation in runtime was observed over the simulations performed. Runtime is reported in units of 10^3 seconds while error is reported in units of 10^{-3} relative to the maximum value of the cost function c.

transition matrices, we apply a softmax normalization with weight 0.1 to each row of P and Q:

$$P(x, x') \mapsto \frac{e^{0.1P(x,x')}}{\sum_{\tilde{x}} e^{0.1P(x,\tilde{x})}}, \qquad Q(y, y') \mapsto \frac{e^{0.1Q(y,y')}}{\sum_{\tilde{y}} e^{0.1Q(y,\tilde{y})}}.$$

For the cost matrix, we apply an absolute value element-wise so that $c \in \mathbb{R}_+^{d\times d}$ and then divide each element by the maximum element in the matrix so that $\|c\|_\infty = 1$. After generating both the transition matrices and cost matrix, we run both ExactOTC and EntropicOTC for each $\xi \in \{75, 100, 200\}$ until convergence. In all runs of EntropicOTC, we choose L and T adaptively as described in Remark 3.10 with tolerance (ε) equal to 10^{-12} and upper bounds of 100 and 1000, respectively. For each choice of $\xi \in \{75, 100, 200\}$, we use 50, 100, and 200 Sinkhorn iterations, respectively. Runtimes of ExactOTC and EntropicOTC in a given iteration are measured from the start to convergence and thus correspond to *total* runtime rather than the runtime of individual iterations. The approximation error of EntropicOTC in a given iteration is measured by taking the absolute difference between the expected cost returned by EntropicOTC and that returned by ExactOTC. Note that after randomization, the cost function c is scaled to $\|c\|_\infty = 1$ and the error is reported on that scale.

The results of the simulation study are shown in Figure 3.1. The error bars in either plot denote the maximum and minimum values observed for each choice of parameters over the five repeated simulations. In our simulations, we found that the time savings in each iteration of EntropicOTC



(a) `ExactOTC` **(b)** `EntropicOTC`

Figure 3.2: Heatmap of costs for all pairs of pieces as computed by `ExactOTC` and `EntropicOTC`. Lower cost (indicated by blue) indicates a better correspondence between the two pieces. The list of pieces and composers considered may be found in Table 3.1.

resulted in substantial time savings over the entire runtime of the algorithm without substantial loss of accuracy. For example, when $d = 100$ and $\xi = 100$, we observed that `EntropicOTC` yielded a time savings of roughly 80% compared to `ExactOTC`. Moreover, weakening the regularization by increasing ξ reduces the error of `EntropicOTC` with little additional runtime. This supports our theoretical findings, indicating that `EntropicOTC` is a good alternative to `ExactOTC` when d is large.

3.7.2 Application to Computer-Generated Music

Next we illustrate the OTC problem in practice through an application to aligning and comparing computer-generated music. HMMs and other state-space models have been explored as a tool for modeling musical arrangements (Ames, 1989; Liu and Selfridge-Field, 2002; Weiland et al., 2005; Allan and Williams, 2005; Pikrakis et al., 2006; Ren et al., 2010; Bell, 2011; Yanchenko and Mukherjee, 2017; Das et al., 2018). In this line of work, sequences of notes are commonly modeled as a stationary processes with latent Markovian structure. As described in Section 3.2, the OTC problem easily extends to this setting, allowing one to apply OT methods to analyzing generative models for music. We utilize the computational tools developed above for two tasks: comparing pieces based on the sequences of notes they contain and generating paired sequences of notes based on existing pieces.

We analyzed a dataset of 36 pieces of classical music from 3 different classical composers (Bach, Beethoven and Mozart) downloaded from https://www.mfiles.co.uk/classical-midi.htm. The pieces considered along with the composer, musical key, and reference number between 1 and 36 may be found in Table 3.1. For each piece, a 3-layer HMMs with 5 hidden states was trained using the code provided in (Yanchenko and Mukherjee, 2017). We refer the reader to (Yanchenko and Mukherjee, 2017) and (Oliver et al., 2004) for details on layered HMMs but note that once a layered HMM is trained it may be recast as a standard HMM and thus the extension of OTC to HMMs described in Section 3.3 still applies. We considered two different cost functions between notes. The first cost function equal to 0 if the two notes are equal or some number of octaves (intervals of 12 semitones) apart, and 1 otherwise. The second cost function is 0 when the first cost function is 0, 1 when the two notes are 5 or 7 semitones apart (perfect consonance), 2 when the two notes are 4 or 9 semitones apart (imperfect consonance) and 10 otherwise. This tiered cost function incorporates a preference for unison over perfect consonance, perfect consonance over imperfect consonance, and imperfect consonance over dissonance.

In the first task, we computed the OTC cost for every pair of pieces, obtaining a pairwise cost matrix. Note that when running `EntropicOTC`, we use $L = 100$, $T = 1000$, $\xi = 50$, and 20 Sinkhorn iterations. The cost matrices obtained using `ExactOTC` and `EntropicOTC` are both depicted in Figure 3.2. The correspondence between rows and columns of the two heatmaps and the musical pieces considered can be found in Table 3.1. We remark that pieces in the same key tended to have lower OTC cost. For example, Bach's Fugue 2 from Book 1 (2 in Figure 3.2) and Fugue 2 from Book 2 (12 in Figure 3.2), both in C minor, had the lowest OTC and entropic OTC costs among all pairs considered. We observe that pairwise costs obtained by either algorithm only differ by 8×10^{-3} on average. In other words, `EntropicOTC` approximates the result of `ExactOTC` with high accuracy.

In the second task, we explored the samples generated from the optimal transition coupling of each pair of fitted HMMs. The optimal transition coupling maximizes the probability of generating consonant pairs of notes while preserving the distributions of the two sequences. This results in sequences that sound harmonious together more frequently. In Figure 3.3, we provide a paired

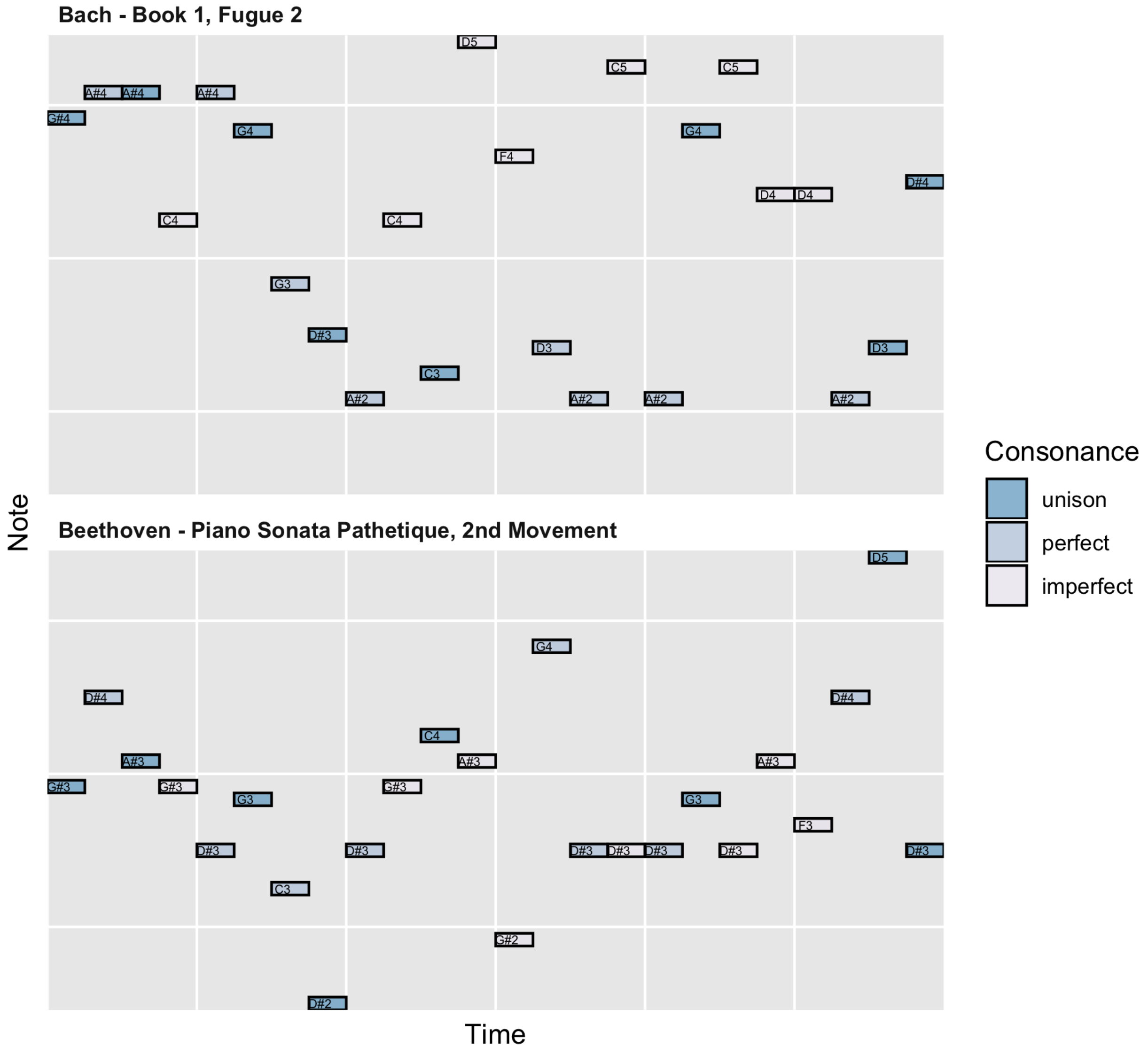


Figure 3.3: An illustration of samples drawn from an optimal transition coupling of Bach's Book 1, Fugue 2 and Beethoven's Sonata Pathétique, Movement 2, both in C minor. The color of each sampled note denotes its consonance with the other note played at the same time.

sequence drawn from the output of `ExactOTC` applied to pieces from Bach and Beethoven. Note that no dissonant pairs of notes were sampled in this sequence. Audio files for this sequence and sequences drawn from other pairings may be found in the accompanying supplemental materials.

3.8 Discussion

In this chapter, we introduced an optimal transport problem for stationary Markov chains that takes the Markovian dynamics into account called the optimal transition coupling (OTC) problem.

Intuitively, the OTC problem aims to synchronize the Markov chains of interest so as to minimize long-term average cost. We demonstrated how this problem may be easily extended to formulate an OT problem for HMMs. In the interest of computation, we recast this problem as a Markov decision process and leveraged this connection to prove that solutions can be obtained via an adaptation of the policy iteration algorithm, referred to as `ExactOTC`. Mirroring the development of entropic OT in (Cuturi, 2013), we also proposed an entropic OTC problem and an associated approximate algorithm, `EntropicOTC`, which scales better with dimension. For cases when the marginal Markov chains must be estimated from data, we showed that the plug-in estimates for either problem are consistent. We showed empirically that `EntropicOTC` approximates the OTC cost with high accuracy and substantially faster runtime than `ExactOTC` in large state space regimes. Finally, we illustrated the use of the OTC problem and the proposed algorithms in practice via an application to computer-generated music.

Future work may consider extending the ideas of the OTC problem to processes with more flexible structure such as Gibbs processes or dynamical linear models. We expect that the extension of our work to processes with richer temporal structure will present interesting computational challenges. Alternatively, future work may explore further applications of the OTC problem in practice. Our approach to analyzing computer-generated music may be easily transferred to any data that may be modeled by an HMM. HMMs and other sequence models with hidden Markov structure are commonly used in a variety of fields including genomics, speech recognition, protein folding, and natural language processing.

3.9 Proofs

3.9.1 Overview of Proofs

In what follows, we detail the proofs of our results. We begin by introducing some additional notation, covering some preliminaries on Markov chains, and remarking on some technical aspects relating to our results.

Additional Notation. We adopt the following additional notation: For a finite set $\mathcal{U} \subset \mathbb{R}$, we define $\min_{>0} \mathcal{U} = \min\{u \in \mathcal{U} : u > 0\}$. We define the inner product $\langle \cdot, \cdot \rangle$ for matrices $U, V \in \mathbb{R}^{n \times n}$ by $\langle U, V \rangle := \sum_{i,j} U_{ij} V_{ij}$. All vector and matrix equations and inequalities should be understood

	Composer	**Piece**	**Key**
1	Bach	Toccata and Fugue	D minor
2	Bach	Book 1, Fugue 2	C minor
3	Bach	Book 1, Fugue 10	E minor
4	Bach	Book 1, Fugue 14	F# minor
5	Bach	Book 1, Fugue 24	B minor
6	Bach	Book 1, Prelude 1	C major
7	Bach	Book 1, Prelude 2	C minor
8	Bach	Book 1, Prelude 3	C# major
9	Bach	Book 1, Prelude 6	D minor
10	Bach	Book 1, Prelude 14	F# minor
11	Bach	Book 1, Prelude 24	B minor
12	Bach	Book 2, Fugue 2	C minor
13	Bach	Book 2, Fugue 7	D# major
14	Bach	Book 2, Prelude 2	C minor
15	Bach	Book 2, Prelude 7	D# major
16	Bach	Book 2, Prelude 12	F minor
17	Bach	Bourrée in E minor	E minor
18	Bach	2 Part Invention, No. 13	A minor
19	Bach	2 Part Invention, No. 4	D minor
20	Bach	Prelude in C major	C major
21	Beethoven	Für Elise	A minor
22	Beethoven	Minuet in G	G major
23	Beethoven	Moonlight Sonata, Movement 1	C# minor
24	Beethoven	Sonata Pathétique, Movement 2	C minor
25	Beethoven	Symphony No. 7, Movement 2	A minor
26	Beethoven	Symphony No. 9, Movement 4	D minor
27	Beethoven	Violin Sonata 1, Movement 1	D major
28	Mozart	Piano Sonata No. 11, Movement 3	A major
29	Mozart	Horn Concerto 4, Movement 3	D# major
30	Mozart	Minuet and Trio, K.1	G major
31	Mozart	Minuet in F major, K.2	F major
32	Mozart	Österreichische Bundeshymne	D# major
33	Mozart	Piano Concerto No. 21, Movement 2	C major
34	Mozart	Piano Sonata No. 13, Movement 1	A# major
35	Mozart	Piano Sonata No. 16	C major
36	Mozart	Symphony No. 40, Movement 1	G minor

Table 3.1: Pieces considered in the application of OTC to computer-generated music.

to hold element-wise. For $i \leq j$, we let $u_i^j = (u_i, ..., u_j)$ and we will denote infinite sequences by boldface, lowercase letters such as $\mathbf{u} = (u_0, u_1, ...)$. For a collection of sets $\mathcal{U}_s \subset \mathbb{R}^{d^2}$ indexed by $s \in \mathcal{X} \times \mathcal{Y}$, we define $\bigotimes_s \mathcal{U}_s$ to be the set of matrices $U \in \mathbb{R}^{d^2 \times d^2}$ such that for every $s \in \mathcal{X} \times \mathcal{Y}$, $U(s, \cdot) \in \mathcal{U}_s$. In particular, we write $\Pi_{\mathrm{TC}}(P, Q) = \bigotimes_{(x,y)} \Pi(P(x, \cdot), Q(y, \cdot))$.

Preliminaries on Markov Chains. For a finite metric space $\mathcal{U}$, we say that a measure $\mu \in \mathcal{M}(\mathcal{U}^{\mathbb{N}})$ is *Markov* or *corresponds to a Markov chain taking values in* $\mathcal{U}$ if for any cylinder set $[u_0 \cdots u_k] \subset \mathcal{U}^{\mathbb{N}}$, $\mu([u_0 \cdots u_k])/\mu([u_0 \cdots u_{k-1}]) = \mu([u_{k-1}u_k])/\mu([u_{k-1}])$, where we let $0/0 = 0$. We say that μ is *stationary* if $\mu = \mu \circ \sigma^{-1}$, where $\sigma : \mathcal{U}^{\mathbb{N}} \to \mathcal{U}^{\mathbb{N}}$ is the left-shift map defined such that for any $\mathbf{u} \in \mathcal{U}^{\mathbb{N}}$, $\sigma(\mathbf{u})_i = u_{i+1}$. When $\mathcal{U}$ has cardinality $n \geq 1$, we define the transition matrix $U \in \mathbb{R}^{n \times n}$ of μ such that for every $u_{k-1}, u_k \in \mathcal{U}$, $U(u_{k-1}, u_k) = \mu([u_{k-1}u_k])/\mu([u_{k-1}])$. If μ is also stationary, its stationary distribution $\lambda_U \in \Delta_n$ is defined such that $\lambda_U(u) = \mu([u])$ for any $u \in \mathcal{U}$. We say that μ or U is *irreducible* if for every $u, u' \in \mathcal{U}$, there exists $k \geq 1$, possibly depending on u and u', such that $U^k(u, u') > 0$. We call μ or U *aperiodic* if $\gcd\{k \geq 1 : U^t(u, u') > 0\} = 1$ for every $u, u' \in \mathcal{U}$. Note that if μ is irreducible, its stationary distribution λ_U is unique. Furthermore, if μ is also aperiodic, there exists $M < \infty$ and $\alpha \in (0, 1)$ such that for any $t \geq 1$, $\max_u \|U^t(u, \cdot) - \lambda_U\|_1 \leq M\alpha^t$. For more details on basic Markov chain theory, we refer the reader to (Levin and Peres, 2017).

Technical Considerations. We endow the finite set $\mathcal{X} \times \mathcal{Y}$ with the discrete topology and $\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ with the corresponding product topology. For each $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\eta > 0$, we endow both $\Pi(P(x, \cdot), Q(y, \cdot))$ and $\Pi_\eta(P(x, \cdot), Q(y, \cdot))$ with the subspace topology inherited from the Euclidean topology on $\mathbb{R}^{d^2}$. Similarly, we endow $\Pi_{\text{TC}}(P, Q)$ and $\Pi_{\text{TC}}^{\eta}(P, Q)$ with the subspace topologies inherited from the Euclidean topology on $\mathbb{R}^{d^2 \times d^2}$. Unless stated otherwise, continuity of any function will be understood to mean with respect to the corresponding topology above.

3.9.2 Proofs from Section 3.2

Proposition 3.4. *Let X and Y be irreducible stationary Markov chains with transition matrices P and Q, respectively. Then any stationary Markov chain with a transition matrix contained in $\Pi_{TC}(P, Q)$ is a transition coupling of X and Y.*

Proof. Let $\pi \in \mathcal{M}((\mathcal{X} \times \mathcal{Y})^{\mathbb{N}})$ be the distribution of a stationary Markov chain with transition matrix $R \in \Pi_{\text{TC}}(P, Q)$ and stationary distribution $r \in \Delta_{d^2}$. Furthermore, let $r_{\mathcal{X}}$ and $r_{\mathcal{Y}} \in \Delta_d$ be the $\mathcal{X}$ and $\mathcal{Y}$ marginals of r, respectively. For a metric space $\mathcal{U}$ and a probability measure $\mu \in \mathcal{M}(\mathcal{U}^{\mathbb{N}})$, we define $\mu_k \in \mathcal{M}(\mathcal{U}^k)$ as the k-dimensional marginal distribution of μ. Formally, for any cylinder set $[a_0^{k-1}] = \{\mathbf{u} \in \mathcal{U}^{\mathbb{N}} : u_j = a_j, 0 \leq j \leq k-1\}$, $\mu_k(a_0^{k-1}) := \mu([a_0^{k-1}])$.

We wish to show that $\pi \in \Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q})$. Since π corresponds to a stationary Markov chain and $R \in \Pi_{\text{TC}}(P, Q)$ by assumption, it suffices to show that $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$. We will do this by showing that $\pi_k \in \Pi(\mathbb{P}_k, \mathbb{Q}_k)$ for every $k \geq 1$. Starting with $k = 1$, for any $y \in \mathcal{Y}$,

$$\begin{aligned}
r_{\mathcal{Y}}(y) &= \sum_x r(x, y) \\
&= \sum_x \sum_{x', y'} r(x', y') R((x', y'), (x, y)) \\
&= \sum_{x', y'} r(x', y') \sum_x R((x', y'), (x, y)) \\
&= \sum_{x', y'} r(x', y') Q(y', y) \\
&= \sum_{y'} r_{\mathcal{Y}}(y') Q(y', y).
\end{aligned}$$

We have proven that $r_{\mathcal{Y}}$ is invariant with respect to Q. Since Q is irreducible, the stationary distribution q of Q is unique. Thus, $r_{\mathcal{Y}} = q$. A similar argument will show that $r_{\mathcal{X}} = p$. Thus, $r \in \Pi(p, q)$ and therefore, $\pi_1 \in \Pi(\mathbb{P}_1, \mathbb{Q}_1)$.

Now suppose that $\pi_k \in \Pi(\mathbb{P}_k, \mathbb{Q}_k)$ for some $k \geq 1$. Fixing $y_0^k \in \mathcal{Y}^{k+1}$, it follows that

$$\begin{aligned}
\sum_{x_0^k} \pi_{k+1}(x_0^k, y_0^k) &= \sum_{x_0^k} \pi_k(x_0^{k-1}, y_0^{k-1}) R((x_{k-1}, y_{k-1}), (x_k, y_k)) \\
&= \sum_{x_0^{k-1}} \pi_k(x_0^{k-1}, y_0^{k-1}) Q(y_{k-1}, y_k) \\
&= \mathbb{Q}_k(y_0^{k-1}) Q(y_{k-1}, y_k) \\
&= \mathbb{Q}_{k+1}(y_0^k).
\end{aligned}$$

Again the proof for the other marginal is identical. So we find that $\pi_{k+1} \in \Pi(\mathbb{P}_{k+1}, \mathbb{Q}_{k+1})$ and since $k \geq 1$ was arbitrary, we conclude that $\pi \in \Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q})$. □

3.9.3 Proofs from Section 3.4

3.9.3.1 Existence of a Deterministic Policy

Proposition 3.5. *Let γ be a policy for TC-MDP. Then there exists a deterministic policy $\tilde{\gamma}$ such that $\bar{c}_{\gamma}(s) = \bar{c}_{\tilde{\gamma}}(s)$ for every $s \in \mathcal{S}$.*

Proof. Before proving the result, it will be helpful to fix some additional notation. Let $\gamma = \{\gamma_s(\cdot) : s \in \mathcal{X} \times \mathcal{Y}\}$ be a policy for TC-MDP. Recall that for each $s = (x, y)$, $\gamma_s(\cdot)$ describes a distribution on $\mathcal{A}_s = \Pi(P(x, \cdot), Q(y, \cdot))$. Define the deterministic policy $\tilde{\gamma} = \{\tilde{\gamma}_s(\cdot) : s \in \mathcal{X} \times \mathcal{Y}\}$ such that for every s, $\tilde{\gamma}_s(\cdot)$ assigns probability one to

$$\tilde{r}_s := \int_{\mathcal{A}_s} r_s \gamma_s(dr_s).$$

Here, $\tilde{r}_s$ is the expected action taken by the agent while occupying a state s and following the policy γ. Note that $\tilde{r}_s \in \mathcal{A}_s$ due to the convexity of $\mathcal{A}_s$. As such, we may collect the row vectors $\{\tilde{r}_s : s \in \mathcal{X} \times \mathcal{Y}\}$ into a single transition matrix $\tilde{R} \in \Pi_{\text{TC}}(P, Q)$ where $\tilde{R}(s, \cdot) = \tilde{r}_s(\cdot)$ for every $s \in \mathcal{X} \times \mathcal{Y}$. In what follows, let $\text{Prob}_{\gamma}(\cdot|s_0)$ and $\text{Prob}_{\tilde{\gamma}}(\cdot|s_0) \in \mathcal{M}(\{\mathcal{A} \times (\mathcal{X} \times \mathcal{Y})\}^{\mathbb{N}})$ be the probability measures corresponding to the action-state processes with initial state s_0 induced by γ and $\tilde{\gamma}$, respectively. In particular,

$$\text{Prob}_{\gamma}(dr_{s_0}, s_1, ..., dr_{s_{t-1}}, s_t|s_0) = \gamma_{s_0}(dr_{s_0})r_{s_0}(s_1) \cdots \gamma_{s_{t-1}}(dr_{s_{t-1}})r_{s_{t-1}}(s_t)$$

and the analogous statement holds for $\text{Prob}_{\tilde{\gamma}}(\cdot|s_0)$. In the case of $\tilde{\gamma}$, one may also show that $\text{Prob}_{\tilde{\gamma}}(s_t|s_0) = \tilde{R}^t(s_0, s_t)$. Finally, let $\mathbb{E}_{\gamma}[\cdot|s_0]$ and $\mathbb{E}_{\tilde{\gamma}}[\cdot|s_0]$ denote expectation with respect to $\text{Prob}_{\gamma}(\cdot|s_0)$ and $\text{Prob}_{\tilde{\gamma}}(\cdot|s_0)$, respectively.

Now, we can prove the result. For any $s_0 \in \mathcal{X} \times \mathcal{Y}$ and $t \geq 1$,

$$\begin{aligned}
\mathbb{E}_{\gamma}\left[c(s_t)|s_0\right] &= \sum_{s_t} c(s_t)\,\text{Prob}_{\gamma}(s_t|s_0) \\
&= \sum_{s_t} c(s_t) \int_{\mathcal{A}_{s_0}} \sum_{s_1} \cdots \int_{\mathcal{A}_{s_{t-1}}} \text{Prob}_{\gamma}(dr_{s_0}, s_1, ..., dr_{s_{t-1}}, s_t|s_0) \\
&= \sum_{s_t} c(s_t) \int_{\mathcal{A}_{s_0}} \sum_{s_1} \cdots \int_{\mathcal{A}_{s_{t-1}}} \gamma_{s_0}(dr_{s_0})\, r_{s_0}(s_1) \cdots \gamma_{s_{t-1}}(dr_{s_{t-1}})\, r_{s_{t-1}}(s_t)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s_1^t} c(s_t) \int_{\mathcal{A}_{s_0}} \cdots \int_{\mathcal{A}_{s_{t-1}}} \gamma_{s_0}(dr_s)\, r_{s_0}(s_1) \cdots \gamma_{s_{t-1}}(dr_{s_{t-1}})\, r_{s_{t-1}}(s_t) \\
&= \sum_{s_1^t} c(s_t) \tilde{r}_{s_0}(s_1) \cdots \tilde{r}_{s_{t-1}}(s_t) \\
&= \sum_{s_1^t} c(s_t) \tilde{R}(s_0, s_1) \cdots \tilde{R}(s_{t-1}, s_t) \\
&= \sum_{s_t} c(s_t) \tilde{R}^t(s_0, s_t) \\
&= \sum_{s_t} c(s_t) \operatorname{Prob}_{\tilde{\gamma}}(s_t | s_0) \\
&= \mathbb{E}_{\tilde{\gamma}} \left[c(s_t) | s_0 \right].
\end{aligned}
$$

Thus, for every $s \in \mathcal{X} \times \mathcal{Y}$,

$$
\overline{c}_{\gamma}(s) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\gamma} \left[c(s_t) | s_0 = s \right] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\tilde{\gamma}} \left[c(s_t) | s_0 = s \right] = \overline{c}_{\tilde{\gamma}}(s).
$$

□

3.9.3.2 Correspondence between TC-MDP and the OTC problem

Next, we prove Proposition 3.6 showing that optimal solutions to TC-MDP necessarily provide optimal solutions to the OTC problem. We rely on the basic idea of recurrent classes of states for finite-state Markov chains. For details on recurrence for Markov chains, we refer the reader to (Levin and Peres, 2017). For any $R \in \Pi_{\mathrm{TC}}(P, Q)$, let $\Lambda(R) := \{\lambda \in \mathcal{M}(\mathcal{X}) : \lambda R = \lambda\}$ denote the set of stationary distributions for R and let $\bigsqcup$ denote a disjoint union. Before proving the proposition, we require a lemma stating that for a given transition coupling matrix $R \in \Pi_{\mathrm{TC}}(P, Q)$, the stationary distribution of R that incurs the least expected cost may be chosen to be the unique stationary distribution of one of R's recurrent classes.

Lemma 3.15. *Let $R \in \Pi_{TC}(P, Q)$ and let $\mathcal{S}_r$ be the set of states belonging to some recurrent class of R. Moreover, for every $s \in \mathcal{S}_r$, let $\lambda_{R,s} \in \Lambda(R)$ denote the stationary distribution of R*

corresponding to the recurrent class in which s lies. Then $\lambda_{R,s}$ is uniquely defined and

$$\min_{s \in \mathcal{S}_r} \langle c, \lambda_{R,s} \rangle = \min_{\lambda \in \Lambda(R)} \langle c, \lambda \rangle.$$

Proof. The uniqueness of $\lambda_{R,s}$ follows from the fact that the chain obtained by restricting R to the recurrent class of s is necessarily irreducible. Now suppose that R has m recurrent classes $\{S_r^i\}_{i=1}^m$ and thus $\mathcal{S}_r = \bigsqcup_{i=1}^m S_r^i$. Then by (Puterman, 2005, Theorem A.5), there exist m linearly independent stationary distributions of R. Note that necessarily, the unique stationary distributions $\{\lambda_i\}_{i=1}^m$ corresponding to the m recurrent classes of R are linearly independent and constitute such a choice. Moreover, it is straightforward to show that $\Lambda(R)$ is equal to the convex hull of $\{\lambda_i\}_{i=1}^m$ and is thus compact. Then since minima of linear functions over a compact, convex set occur at the extreme points of the feasible set,

$$\min_{s \in \mathcal{S}_r} \langle c, \lambda_{R,s} \rangle = \min_{i=1,\ldots,m} \langle c, \lambda_i \rangle = \min_{\lambda \in \Lambda(R)} \langle c, \lambda \rangle.$$

□

Proposition 3.6. *If X and Y are irreducible, then any $R \in \Pi(P, Q)$ that is an optimal policy for TC-MDP corresponds to an optimal coupling $\pi_R \in \Pi_{TC}(\mathbb{P}, \mathbb{Q})$ with expected cost $\min_{s \in \mathcal{S}} \bar{c}_R(s)$.*

Proof. For every $R \in \Pi_{\mathrm{TC}}(P, Q)$ and $s \in \mathcal{S}$, let $\lambda_{R,s} \in \Lambda(R)$ be the stationary distribution of R defined by

$$\lambda_{R,s} := \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} R^t(s, \cdot).$$

Note that $\lambda_{R,s}$ is well-defined by (Puterman, 2005, Theorem A.5). Moreover, we will use $\mathcal{S}_r(R)$ to refer to the set of all states in $\mathcal{S}$ that belong to a recurrent class of R. Since the space $\mathcal{S}$ is finite, $\mathcal{S}_r(R)$ is necessarily non-empty for every $R \in \Pi_{\mathrm{TC}}(P, Q)$. Finally, note that whenever $s \in \mathcal{S}_r(R)$, $\lambda_{R,s}$ is the unique stationary distribution of R associated with the recurrent class in which s lies.

Now let $R^* \in \Pi_{\mathrm{TC}}(P, Q)$ be optimal for TC-MDP. We will construct a transition coupling $\pi_* \in \Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$ from R^* that is optimal in the OTC problem. Note that by definition, $\bar{c}_R(s) = \langle c, \lambda_{R,s} \rangle$. Then by the optimality of R^* in TC-MDP, $\langle c, \lambda_{R^*,s} \rangle = \min_{R \in \Pi_{\mathrm{TC}}(P,Q)} \langle c, \lambda_{R,s} \rangle$ for every $s \in \mathcal{S}$. So by Lemma 3.15,

$$\min_{s \in \mathcal{S}_r} \langle c, \lambda_{R^*,s} \rangle = \min_{s \in \mathcal{S}} \langle c, \lambda_{R^*,s} \rangle = \min_{R \in \Pi_{\mathrm{TC}}(P,Q)} \min_{s \in \mathcal{S}} \langle c, \lambda_{R,s} \rangle = \min_{R \in \Pi_{\mathrm{TC}}(P,Q)} \min_{\lambda \in \Lambda(R)} \langle c, \lambda \rangle. \tag{3.8}$$

Let $s^* \in \operatorname{argmin}_{s \in \mathcal{S}_r} \langle c, \lambda_{R^*,s} \rangle$ and define $\pi_* \in \Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$ to be the transition coupling with transition matrix R^* and stationary distribution λ_{s^*}. Then by (3.8),

$$\int c \, d\pi_* = \langle c, \lambda_{R^*,s^*} \rangle = \min_{R \in \Pi_{\mathrm{TC}}(P,Q)} \min_{\lambda \in \Lambda(R)} \langle c, \lambda \rangle.$$

But at this point, we recognize that the quantity on the right is exactly the OTC cost. To see this, note that by Proposition 3.4 every $\pi \in \Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$ is uniquely characterized by a transition matrix $R \in \Pi_{\mathrm{TC}}(P, Q)$ and a stationary distribution $\lambda \in \Lambda(R)$, and $\int c \, d\pi = \langle c, \lambda \rangle$. Thus

$$\int c \, d\pi_* = \min_{R \in \Pi_{\mathrm{TC}}(P,Q)} \min_{\lambda \in \Lambda(R)} \langle c, \lambda \rangle = \min_{\pi \in \Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})} \int c \, d\pi,$$

and we conclude that π_* is optimal for the OTC problem. Finally, by construction, $\int c \, d\pi_* = \min_{s \in \mathcal{S}} \langle c, \lambda_{R^*,s} \rangle = \min_{s \in \mathcal{S}} \overline{c}_{R^*}(s)$. □

3.9.3.3 Convergence of `ExactOTC`

Next, we prove the convergence of Algorithm 1 to a solution of TC-MDP. For any polyhedron $\mathcal{P} \in \mathbb{R}^{n \times n}$, let $\mathcal{E}(\mathcal{P})$ denote the extreme points of $\mathcal{P}$. Recall that if $\mathcal{P}$ is bounded, a linear function on $\mathcal{P}$ achieves its minimum on $\mathcal{E}(\mathcal{P})$ (Bertsimas and Tsitsiklis, 1997). Note that for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, since $\Pi(P(x, \cdot), Q(y, \cdot))$ is a bounded subset of $\mathbb{R}^{d^2}$ defined by a finite set of linear equality and inequality constraints, it is a bounded polyhedron.

Theorem 3.7. *Algorithm 1 converges to a solution (g^*, h^*, R^*) of TC-MDP in a finite number of iterations. Moreover, if X and Y are irreducible, R^* is the transition matrix of an optimal transition coupling of X and Y.*

Proof. We will first show that Algorithm 1 converges to some (g^*, h^*, R^*) and then argue that this is a solution to TC-MDP. Recall that for every $s = (x, y)$, $\mathcal{A}_s = \Pi(P(x, \cdot), Q(y, \cdot))$ and $\mathcal{A} = \bigcup_s \mathcal{A}_s$. In this proof, it is most convenient to consider the concatenatation of the state-action spaces instead of the union $\bigcup_s \mathcal{A}_s$. Abusing notation, we let $\mathcal{A} = \bigotimes_s \mathcal{A}_s$ for the remainder of the proof. Furthermore, let $\mathcal{A}'_s = \mathcal{E}(\mathcal{A}_s)$ be the set of extreme points of $\mathcal{A}_s$. As $\mathcal{A}_s$ is a bounded polyhedron, $\mathcal{A}'_s$ is finite. For every $n \geq 1$, let (g_n, h_n, R_n) be the n'th iterate of Algorithm 1. Since the rows of R_n are solutions of the linear programs in Algorithm 1b,$R_n(s,\cdot) \in \mathcal{E}(\mathcal{A}'_s)$ for every s. Thus the iterates of Algorithm 1 are the same as the iterates of the policy iteration algorithm for the restricted MDP $(\mathcal{X} \times \mathcal{Y}, \bigcup_s \mathcal{A}'_s, \{p(\cdot|s,a)\}, c)$ constructed by restricting the state-action spaces $\mathcal{A}_s$ of TC-MDP to $\mathcal{A}'_s$ for each s. Since $\mathcal{A}'_s$ is finite for every s, standard results (Puterman, 2005, Theorem 9.2.3) ensure that the iterates $\{(g_n, h_n, R_n)\}$ of Algorithm 1 will converge to a solution (g^*, h^*, R^*) in a finite number of iterations. Thus, we need only show that any stationary point of Algorithm 1 is necessarily a solution to TC-MDP.

Let (g^*, h^*, R^*) be a stationary point of Algorithm 1. Then $R^* = \texttt{ExactTCI}(g^*, h^*, R^*, \bigotimes_s \mathcal{A}'_s)$ and consequently, $R^*(s,\cdot) \in \operatorname{argmin}_{r \in \mathcal{A}'_s} rh^*$ for every s. Since $\mathcal{A}_s$ is a bounded polyhedron, $\min_{r \in \mathcal{A}_s} rh^* = \min_{r \in \mathcal{A}'_s} rh^*$ and we find that $R^*(s,\cdot) \in \operatorname{argmin}_{r \in \mathcal{A}_s} rh^*$. Since $\mathcal{A} = \bigotimes_s \mathcal{A}_s$, we may write $R^* \in \operatorname{argmin}_{R \in \mathcal{A}} Rh^*$ where the minimum is understood to be element-wise. Using the assumption that (g^*, h^*, R^*) is a stationary point of Algorithm 1 again, $(g^*, h^*) = \texttt{ExactTCE}(R^*)$. It follows that

$$g^* + h^* = R^* h^* + c. \tag{3.9}$$

Since $R^* \in \operatorname{argmin}_{R \in \mathcal{A}} Rh^*$, we obtain

$$g^* + h^* = \min_{R \in \mathcal{A}} Rh^* + c.$$

Then by (Puterman, 2005, Theorem 9.1.2 (c)), g^* is the optimal expected cost for TC-MDP. Moreover, by (3.9) and (Puterman, 2005, Theorem 8.2.6 (b)), $g^* = \overline{R}^* c = \overline{c}_{R^*}$, where we remind the reader that $\overline{R}^* = \lim_{T \to \infty} {}^1\!/\!{}_T \sum_{t=0}^{T-1} R^{*t}$. Thus R^* has optimal expected cost among policies for TC-MDP and we conclude that (g^*, h^*, R^*) is a solution to TC-MDP.

If X and Y are irreducible, then by Proposition 3.4, every transition coupling matrix in $\Pi_{\text{TC}}(P, Q)$ induces a transition coupling in $\Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q})$. Since R^* has minimal expected cost over all elements of $\Pi_{\text{TC}}(P, Q)$, it attains the minimum in Problem (3.2) and is thus an optimal transition coupling. □

3.9.4 Proofs from Section 3.5

3.9.4.1 Complexity of Approximate Transition Coupling Evaluation

Theorem 3.9. *Let $R \in \Pi_{TC}(P, Q)$ be aperiodic and irreducible with mixing coefficients $M \in \mathbb{R}_+$ and $\alpha \in [0, 1)$ and gain and bias vectors $g \in \mathbb{R}^{d^2}$ and $h \in \mathbb{R}^{d^2}$, respectively. Then for any $\varepsilon > 0$, there exist $L, T \in \mathbb{N}$ such that* `ApproxTCE`(R, L, T) *yields $(\tilde{g}, \tilde{h})$ satisfying $\|\tilde{g} - g\|_\infty \leq \varepsilon$ and $\|\tilde{h} - h\|_1 \leq \varepsilon$ in $\tilde{\mathcal{O}}\left(\frac{d^4}{\log \alpha^{-1}} \log\left(\frac{M}{\varepsilon(1-\alpha)}\right)\right)$ time.*

Proof. Briefly, we remind the reader that $g = \overline{R}c$ and $h = \sum_{t=0}^{\infty} R^t(c - g)$, and that for integers $L, T \geq 1$ to be chosen later,

$$\tilde{g} = \langle {}^1\!/\!{}_{d^2} R^L c, \mathbb{1}\rangle \mathbb{1} \quad \text{and} \quad \tilde{h} = \sum_{t=0}^{T} R^t(c - \tilde{g}).$$

Note that the expression for $\tilde{g}$ may also be written as

$$\tilde{g} = \left(\frac{1}{d^2} \sum_s R^L(s, \cdot)c\right) \mathbb{1}.$$

We begin by studying the approximation error for $\tilde{h}$ by first considering the intermediate quantity $h' := \sum_{t=0}^{T} R^t(c - g)$. By the triangle inequality,

$$\|\tilde{h} - h\|_1 \leq \|\tilde{h} - h'\|_1 + \|h' - h\|_1, \tag{3.10}$$

so it suffices to control the two terms on the right hand side. Using Hölder's inequality, it follows that

$$\begin{aligned}
\|\tilde{h} - h'\|_1 &= \left\| \sum_{t=0}^{T} R^t(\tilde{g} - g) \right\|_1 \\
&\leq \sum_{t=0}^{T} \left\| R^t(\tilde{g} - g) \right\|_1 \\
&\leq d^2 \sum_{t=0}^{T} \max_s \left| R^t(s, \cdot)(\tilde{g} - g) \right| \\
&\overset{(*)}{\leq} d^2 \sum_{t=0}^{T} \|\tilde{g} - g\|_\infty
\end{aligned}$$

$$= (T+1)d^2\|\tilde{g} - g\|_\infty,$$

where (*) uses the fact that $\|R^t(s,\cdot)\|_1 = 1$ for every $t \geq 1$ and $s \in \mathcal{X} \times \mathcal{Y}$. Next we wish to bound $\|h' - h\|_1$. Since $R^t\overline{R} = \overline{R}$ for any $t \geq 1$, we may write h and h' as

$$h = \sum_{t=0}^{\infty}(R^t - \overline{R})c \quad \text{and} \quad h' = \sum_{t=0}^{T}(R^t - \overline{R})c.$$

Moreover, since R is aperiodic and irreducible, the Perron-Frobenius theorem implies that $\overline{R}(s,\cdot) = \lambda_R$ for every $s \in \mathcal{X} \times \mathcal{Y}$, where $\lambda_R \in \Delta_{d^2}$ is the unique stationary distribution of R. Now by Hölder's inequality and the mixing assumption on R,

$$\begin{aligned}
\|h' - h\|_1 &= \left\| \sum_{t=T+1}^{\infty} (R^t - \overline{R})c \right\|_1 \\
&\leq \sum_{t=T+1}^{\infty} \|(R^t - \overline{R})c\|_1 \\
&\leq d^2 \sum_{t=T+1}^{\infty} \max_s \left|(R^t(s,\cdot) - \lambda_R)c\right| \\
&\leq \|c\|_\infty d^2 \sum_{t=T+1}^{\infty} \max_s \left\|R^t(s,\cdot) - \lambda_R\right\|_1 \\
&\leq \|c\|_\infty d^2 \sum_{t=T+1}^{\infty} M\alpha^t \\
&= M\|c\|_\infty \frac{\alpha^{T+1}}{1-\alpha} d^2.
\end{aligned}$$

Thus by (3.10),

$$\|\tilde{h} - h\|_1 \leq (T+1)\|\tilde{g} - g\|_\infty d^2 + M\|c\|_\infty \frac{\alpha^{T+1}}{1-\alpha} d^2. \tag{3.11}$$

So in order to bound $\|\tilde{h} - h\|_1$, we require a bound on $\|\tilde{g} - g\|_\infty$. Using the fact that $\tilde{g}$ and g are constant vectors, Hölder's inequality and the mixing assumption on R,

$$\begin{aligned}
\|\tilde{g} - g\|_\infty &= \left\| \left(\frac{1}{d^2} \sum_s R^L(s,\cdot)c \right) \mathbb{1} - \overline{R}c \right\|_\infty \\
&= \left| \frac{1}{d^2} \sum_s R^L(s,\cdot)c - \lambda_R c \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{d^2}\sum_s \left|(R^L(s,\cdot)-\lambda_R)c\right| \\
&\leq \frac{1}{d^2}\sum_s \|c\|_\infty \|R^L(s,\cdot)-\lambda_R\|_1 \\
&\leq \frac{1}{d^2}\sum_s M\alpha^L \|c\|_\infty \\
&\leq M\alpha^L\|c\|_\infty.
\end{aligned}
$$

Plugging this into (3.11),

$$\|\tilde{h}-h\|_1 \leq M\alpha^L\|c\|_\infty(T+1)d^2 + M\|c\|_\infty \frac{\alpha^{T+1}}{1-\alpha}d^2.$$

Then choosing

$$T+1 \geq \frac{1}{\log\alpha^{-1}}\log\left(\frac{2M\|c\|_\infty d^2\varepsilon^{-1}}{(1-\alpha)}\right) = \tilde{\mathcal{O}}\left(\frac{1}{\log\alpha^{-1}}\log\left(\frac{M}{\varepsilon(1-\alpha)}\right)\right) \tag{3.12}$$

and

$$L \geq \frac{\log\left(2(T+1)M\|c\|_\infty d^2\varepsilon^{-1}\right)}{\log\alpha^{-1}} = \tilde{\mathcal{O}}\left(\frac{1}{\log\alpha^{-1}}\log\left(\frac{M}{\varepsilon}\right)\right), \tag{3.13}$$

we obtain $\|\tilde{h}-h\|_1 \leq \varepsilon$. Note that for this choice of L, $\|\tilde{g}-g\|_\infty \leq \varepsilon/2(T+1)$. Since $T+1 \geq 1$, this implies that $\|\tilde{g}-g\|_\infty \leq \varepsilon$. So the error for $\tilde{g}$ is controlled at the desired level as well.

Now consider the cost of computing $\tilde{g}$ and $\tilde{h}$. Computing $\tilde{g}$ requires L multiplications of a vector in $\mathbb{R}^{d^2}$ by $R \in \mathbb{R}^{d^2\times d^2}$, which takes $\mathcal{O}(Ld^4)$ time, followed by an inner product with $\mathbb{1} \in \mathbb{R}^{d^2}$, multiplication with $\mathbb{1} \in \mathbb{R}^{d^2}$ and multiplication by $1/d^2$, each in $\mathcal{O}(d^2)$ time. This requires $\mathcal{O}(Ld^4)+\mathcal{O}(d^2)+\mathcal{O}(d^2)+\mathcal{O}(d^2) = \mathcal{O}(Ld^4)$ time. Letting L be the minimum integer satisfying (3.13), this takes time

$$\mathcal{O}(Ld^4) = \tilde{\mathcal{O}}\left(\frac{d^4}{\log\alpha^{-1}}\log\left(\frac{M}{\varepsilon}\right)\right).$$

On the other hand, given $\tilde{g}$, computing $\tilde{h}$ requires computing $c-\tilde{g} \in \mathbb{R}^{d^2}$ in $\mathcal{O}(d^2)$ operations then multiplying by $R \in \mathbb{R}^{d^2\times d^2}$ $T+1$ times in $\mathcal{O}(Td^4)$ time. Finally, the sum may also be evaluated in $\mathcal{O}(Td^4)$, requiring a total time of $\mathcal{O}(d^2)+\mathcal{O}(Td^4)+\mathcal{O}(Td^4) = \mathcal{O}(Td^4)$. Letting T be the minimum integer satisfying (3.12), this takes time

$$\mathcal{O}(Td^4) = \tilde{\mathcal{O}}\left(\frac{d^4}{\log \alpha^{-1}} \log\left(\frac{M}{\varepsilon(1-\alpha)}\right)\right). \tag{3.14}$$

In total, we find that `ApproxTCE`(R, L, T) takes time

$$\tilde{\mathcal{O}}\left(\frac{d^4}{\log \alpha^{-1}} \log\left(\frac{M}{\varepsilon}\right)\right) + \tilde{\mathcal{O}}\left(\frac{d^4}{\log \alpha^{-1}} \log\left(\frac{M}{\varepsilon(1-\alpha)}\right)\right) = \tilde{\mathcal{O}}\left(\frac{d^4}{\log \alpha^{-1}} \log\left(\frac{M}{\varepsilon(1-\alpha)}\right)\right).$$

□

3.9.4.2 Aperiodicity and Irreducibility of Elements of ri($\Pi_{\mathrm{TC}}(P, Q)$)

Next we prove Proposition 3.11 regarding the aperiodicity and irreducibility of elements of ri($\Pi_{\mathrm{TC}}(P, Q)$). We begin with two elementary lemmas about the independent transition coupling.

Lemma 3.16. *For any $k \geq 1$, $(P \otimes Q)^k = P^k \otimes Q^k$.*

Proof. The result clearly holds for $k = 1$, so assume that it holds for some $k \geq 1$. For any (x, y), $(x', y') \in \mathcal{X} \times \mathcal{Y}$, we can show

$$\begin{aligned}
(P \otimes Q)^{k+1}((x,y),(x',y')) &= \sum_{\tilde{x},\tilde{y}} (P \otimes Q)^k((x,y),(\tilde{x},\tilde{y}))\, P \otimes Q((\tilde{x},\tilde{y}),(x',y')) \\
&= \sum_{\tilde{x},\tilde{y}} P^k(x,\tilde{x})\, Q^k(y,\tilde{y})\, P(\tilde{x},x')\, Q(\tilde{y},y') \\
&= \sum_{\tilde{x}} P^k(x,\tilde{x})\, P(\tilde{x},x') \sum_{\tilde{y}} Q^k(y,\tilde{y})\, Q(\tilde{y},y') \\
&= P^{k+1}(x,x')\, Q^{k+1}(y,y') \\
&= P^{k+1} \otimes Q^{k+1}((x,y),(x',y')).
\end{aligned}$$

By induction, the lemma is proven. □

Lemma 3.17. *If P and Q are aperiodic and irreducible, then the independent transition coupling $P \otimes Q$ is aperiodic and irreducible.*

Proof. Since P and Q are aperiodic and irreducible, there exist $\ell_0, m_0 \geq 1$ such that for any $\ell \geq \ell_0$ and $m \geq m_0$, $P^\ell > 0$ and $Q^m > 0$ (Levin and Peres, 2017, Proposition 1.7). Defining $k_0 := \ell_0 \vee m_0$, for every $k \geq k_0$, $P^k, Q^k > 0$. By Lemma 3.16, it follows that $(P \otimes Q)^k = P^k \otimes Q^k > 0$ for all $k \geq k_0$. Thus $P \otimes Q$ is irreducible. Furthermore, for every $s \in \mathcal{X} \times \mathcal{Y}$, $\gcd\{k \geq 1 : (P \otimes Q)^k(s, s) > 0\} = \gcd\{..., k_0, k_0 + 1, ...\} = 1$ and we conclude that $P \otimes Q$ is also aperiodic. □

Next we prove Proposition 3.11. Recall that for a set $\mathcal{U} \subset \mathbb{R}^n$, $B_\varepsilon(u) \subset \mathbb{R}^n$ denotes the open ball of radius $\varepsilon > 0$ centered at $u \in \mathcal{U}$, $\text{aff}(\mathcal{U})$ denotes the affine hull, defined as $\text{aff}(\mathcal{U}) = \{\sum_{i=1}^k \alpha_i u_i : k \in \mathbb{N}, u_1, ..., u_k \in \mathcal{U}, \sum_{i=1}^k \alpha_i = 1\}$, and $\text{ri}(\mathcal{U})$ denotes the relative interior, defined as $\text{ri}(\mathcal{U}) = \{u \in \mathcal{U} : \exists \varepsilon > 0 \text{ s.t. } B_\varepsilon(u) \cap \text{aff}(\mathcal{U}) \subset \mathcal{U}\}$.

Proposition 3.11. *If P and Q are aperiodic and irreducible then every $R \in ri(\Pi_{TC}(P, Q))$ is also aperiodic and irreducible, and thus mixing.*

Proof. First we establish that $P \otimes Q(s, s') > 0$ implies that $R(s, s') > 0$ for every $s, s' \in \mathcal{X} \times \mathcal{Y}$. Suppose for the sake of contradiction that there exist $s, s' \in \mathcal{X} \times \mathcal{Y}$ such that $P \otimes Q(s, s') > 0$ and $R(s, s') = 0$. By definition, there is some $\varepsilon > 0$ such that $B_\varepsilon(R) \cap \text{aff}(\Pi_{\text{TC}}(P, Q)) \subset \Pi_{\text{TC}}(P, Q)$. Defining $R' = R + \frac{\varepsilon}{2} d$ where $d = (R - P \otimes Q)/\|R - P \otimes Q\|_2$, one may verify that $R' \in B_\varepsilon(R) \cap \text{aff}(\Pi_{\text{TC}}(P, Q))$. Thus by the choice of R, we have $R' \in \Pi_{\text{TC}}(P, Q)$. However, our assumptions imply that $R'(s, s') < 0$, a contradiction. This proves the preliminary claim.

By nature of the fact that $R((x, y), \cdot) \in \Pi(P(x, \cdot), Q(y, \cdot))$, one may easily establish that the reverse implication holds: $R(s, s') > 0$ implies that $P \otimes Q(s, s') > 0$ for every $s, s' \in \mathcal{X} \times \mathcal{Y}$. As such, one may find a positive constant $a > 0$ such that $aP \otimes Q \leq R$ where the inequality is understood to hold element-wise. Now, by Lemma 3.17, $P \otimes Q$ is aperiodic and irreducible. Thus there exists $k \geq 1$ such that $(P \otimes Q)^k > 0$. Thus, $R^k \geq a^k (P \otimes Q)^k > 0$ and it follows that R is aperiodic and irreducible as well. The mixing property of R follows from (Levin and Peres, 2017, Theorem 4.9). □

3.9.4.3 Complexity of Entropic Transition Coupling Improvement

Next we aim to prove Theorem 3.12, showing that `EntropicTCI` returns an improved transition coupling with error bounded by $\varepsilon > 0$ in $\tilde{\mathcal{O}}(d^4 \varepsilon^{-4})$ time. Recall that `EntropicTCI` improves policies by solving d^2 entropy-regularized OT transport problems, calling the `ApproxOT` algorithm (Altschuler et al., 2017) for each problem. Before we can prove Theorem 3.12, we must analyze the computational complexity of `ApproxOT`. In the following discussion as well as Lemma 3.18, we find it most convenient to adopt the notation of (Altschuler et al., 2017). Thus, we fix two probability vectors $r \in \Delta_m$ and $c \in \Delta_n$, a non-negative cost matrix $C \in \mathbb{R}_+^{m \times n}$, a regularization parameter $\xi > 0$, and an error tolerance $\varepsilon > 0$. For vectors in $\mathbb{R}^m$ or $\mathbb{R}^n$ and matrices in $\mathbb{R}^{m \times n}$, we temporarily drop the double-indexing convention, using subscripts instead to denote elements (i.e. u_i and X_{ij}). Finally, for a coupling $X \in \Pi(r, c)$, let $H(X) = -\sum_{ij} X_{ij} \log X_{ij}$ be the Shannon entropy.

Recall that the entropic OT problem is defined as,

$$\begin{aligned} \text{minimize} \quad & \langle X, C \rangle - \frac{1}{\xi} H(X) \\ \text{subject to} \quad & X \in \Pi(r, c). \end{aligned} \tag{3.15}$$

In (Cuturi, 2013), Cuturi showed that solutions to (3.15) have a computationally convenient form. Namely, if $X_\xi^* \in \Pi(r, c)$ is the solution to (3.15), then it is unique and can be written as $X_\xi^* = \operatorname{diag}(e^{u^*}) K \operatorname{diag}(e^{v^*})$ for some $u^* \in \mathbb{R}^m$ and $v^* \in \mathbb{R}^n$, where $K = e^{-\xi C}$. As a result, (3.15) can be formulated as a matrix scaling problem and solved using Sinkhorn's algorithm (Sinkhorn, 1967).

More recent work (Altschuler et al., 2017) introduced the ApproxOT algorithm (Algorithm 3), which combines Sinkhorn's algorithm with a rounding step to obtain an approximate solution to the OT problem. In particular, ApproxOT runs Sinkhorn (Algorithm 4) to obtain a coupling of the form $X' = \operatorname{diag}(e^{u'}) K \operatorname{diag}(e^{v'}) \in \Pi(r', c')$, where $\|r - r'\|_1 + \|c - c'\|_1 \leq \varepsilon$, then applies Round (Algorithm 5) to X' to obtain $\hat{X} \in \Pi(r, c)$. ApproxOT was originally intended for approximating the OT cost, but we use it to approximate the regularized optimal coupling $X_\xi^* \in \Pi(r, c)$. In particular, we wish to show that for appropriate choice of parameters, ApproxOT yields a coupling $\hat{X} \in \Pi(r, c)$ such that $\|\hat{X} - X_\xi^*\|_1 \leq \varepsilon$ in $\tilde{\mathcal{O}}(mn\varepsilon^{-4})$ time. To the best of our knowledge, this result has not appeared in the literature. So we state and prove it in Lemma 3.18.

Algorithm 3: ApproxOT

```
result: Optimal coupling
input: $r, c, C, \xi, \varepsilon$
/* Subset to positive elements */
$\mathcal{R} \leftarrow \{i : r_i > 0\}$, $\mathcal{C} \leftarrow \{j : c_j > 0\}$
$\mathcal{S} \leftarrow \mathcal{R} \times \mathcal{C}$, $\tilde{r} \leftarrow r_{\mathcal{R}}$, $c \leftarrow c_{\mathcal{C}}$
/* Set parameters */
$J \leftarrow 4 \log n \|C_{\mathcal{S}}\|_\infty / \varepsilon - \log \min_{ij}\{\tilde{r}_i, \tilde{c}_j\}$
$\varepsilon' \leftarrow \varepsilon^2 / 8J$
$K \leftarrow \exp(-\xi C_{\mathcal{S}})$
/* Approximate Sinkhorn projection */
$X' \leftarrow$ Sinkhorn$(K, \tilde{r}, \tilde{c}, \varepsilon')$
/* Round to feasible coupling */
$X' \leftarrow$ Round$(X', \Pi(\tilde{r}, \tilde{c}))$
/* Replace zeroes */
$\hat{X} \leftarrow 0_{d \times d}$, $\hat{X}_{\mathcal{S}} \leftarrow X'$
return $\hat{X}$
```

Algorithm 4: Sinkhorn

```
result: Approximate Sinkhorn projection
input: $K, r, c, \varepsilon'$
$k \leftarrow 0$
$X_0 \leftarrow K / \|K\|_1$, $u^0 \leftarrow 0$, $v^0 \leftarrow 0$
while $\|X_k \mathbb{1} - r\|_1 + \|X_k^\top \mathbb{1} - c\|_1 > \varepsilon'$ do
    $k \leftarrow k + 1$
    if $k$ odd then
        $r^k \leftarrow X_k \mathbb{1}$
        $u_i \leftarrow \log(r_i / r_i^k)$ for $i \in [n]$
        $u^k \leftarrow u^{k-1} + u$, $v^k \leftarrow v^{k-1}$
    else
        $c^k \leftarrow X_k^\top \mathbb{1}$
        $v_j \leftarrow \log(c_j / c_j^k)$ for $j \in [n]$
        $v^k \leftarrow v^{k-1} + v$, $u^k \leftarrow u^{k-1}$
    $X_k \leftarrow \operatorname{diag}(e^{u^k}) K \operatorname{diag}(e^{v^k})$
return $X_k$
```

Algorithm 5: Round

```
result: Feasible coupling
input: $F, \Pi(r, c)$
$r' \leftarrow F\mathbb{1}$
$X \leftarrow \operatorname{diag}(x)$ with $x_i = r_i / r'_i \wedge 1$
$F' \leftarrow XF$
$c' \leftarrow (F')^\top \mathbb{1}$
$Y \leftarrow \operatorname{diag}(y)$ with $y_j = c_j / c'_j \wedge 1$
$F'' \leftarrow F'Y$
$r'' \leftarrow F''\mathbb{1}$, $c'' \leftarrow (F'')^\top \mathbb{1}$
$\mathrm{err}_r \leftarrow r - r''$, $\mathrm{err}_c \leftarrow c - c''$
return $F'' + \mathrm{err}_r \, \mathrm{err}_c^\top / \| \mathrm{err}_r \|_1$
```

Note that ApproxOT was originally defined for fully-supported marginal probability vectors ($r, c > 0$). However, this will not always be the case in Algorithm 2b. In particular, transition couplings may be sparse, even when P and Q are strictly positive. Thus we add an extra step to ApproxOT that subsets the quantities of interest to their positive entries. For an index set $\mathcal{I}$ and a vector / matrix A we let $A_{\mathcal{I}}$ denote the subvector / matrix that retains only elements with indices contained in $\mathcal{I}$.

Lemma 3.18. *Let $r \in \Delta_m$ and $c \in \Delta_n$ have all positive entries, $C \in \mathbb{R}_+^{m \times n}$, $\xi > 0$ and $\varepsilon \in (0, 1)$. Then* ApproxOT$(r, c, C, \xi, \varepsilon)$ *(Algorithm 3) returns a coupling $\hat{X} \in \Pi(r, c)$ such that $\|\hat{X} - X_\xi^*\|_1 \leq \varepsilon$,* *where* $X^*_\xi \in \operatorname{argmin}_{X \in \Pi(r,c)} \langle X, C\rangle - 1/\xi H(X)$*, in time* $\tilde{\mathcal{O}}(mn\varepsilon^{-4}\xi\|C\|_\infty(\xi^2\|C\|_\infty^2 + (\log b^{-1})^2))$ *where* $b = \min_{ij}\{r_i, c_j\}$.

Proof. Let $\varepsilon' > 0$, $K = e^{-\xi C}$, $X' \in \Delta_{m\times n}$ be the output of $\texttt{Sinkhorn}(K, r, c, \varepsilon')$ and $\hat{X} \in \Pi(r, c)$ be the output of $\texttt{Round}(X', \Pi(r, c))$. By the triangle inequality,

$$\|\hat{X} - X^*_\xi\|_1 \leq \|\hat{X} - X'\|_1 + \|X' - X^*_\xi\|_1. \tag{3.16}$$

We will first describe how to control the second term on the right hand side. By Pinsker's inequality, $\|X' - X^*_\xi\|_1^2 \leq 2\mathcal{K}(X^*_\xi\|X')$, so it suffices to bound the KL-divergence between the two couplings. From Lemma 2 of (Cuturi, 2013) that $X^*_\xi = \operatorname{diag}(e^{u^*})K\operatorname{diag}(e^{v^*})$ for some $u^* \in \mathbb{R}^m$, $v^* \in \mathbb{R}^n$, and $K = e^{-\xi C}$. By construction we also have $X' = \operatorname{diag}(e^{u'})K\operatorname{diag}(e^{v'})$ for some $u' \in \mathbb{R}^m$ and $v' \in \mathbb{R}^n$. Now rewriting the KL-divergence,

$$\begin{aligned}
\mathcal{K}(X^*_\xi\|X') &= \sum_{ij} X^*_{\xi,ij}\log X^*_{\xi,ij} - \sum_{ij} X^*_{\xi,ij}\log X'_{ij} \\
&= \sum_{ij} X^*_{\xi,ij}\left(u^*_i + v^*_j - \xi C_{ij}\right) - \sum_{ij} X^*_{\xi,ij}\left(u'_i + v'_j - \xi C_{ij}\right) \\
&= \sum_{ij} X^*_{\xi,ij}(u^*_i - u'_i) + \sum_{ij} X^*_{\xi,ij}(v^*_j - v'_j) \\
&= \sum_i (u^*_i - u'_i)\sum_j X^*_{\xi,ij} + \sum_j (v^*_j - v'_j)\sum_i X^*_{\xi,ij} \\
&= \sum_i (u^*_i - u'_i)r_i + \sum_j (v^*_j - v'_i)c_j \\
&= \langle u^* - u', r\rangle + \langle v^* - v', c\rangle.
\end{aligned}$$

Writing $\psi(u, v) = \langle \mathbb{1}, \operatorname{diag}(e^u)K\operatorname{diag}(e^v)\mathbb{1}\rangle - \langle u, r\rangle - \langle v, c\rangle$ for the objective of the dual entropic OT problem (Dvurechensky et al., 2018), we immediately see that

$$\tilde{\psi}(u', v') := \psi(u', v') - \psi(u^*, v^*) = \langle u^* - u', r\rangle + \langle v^* - v', c\rangle.$$

Now let r' and c' be the row and column marginals of X', respectively. Using the two previous displays and applying the upper bound from (Dvurechensky et al., 2018, Lemma 2), we obtain

$$\mathcal{K}(X_\xi^* \| X') = \tilde{\psi}(u, v) \leq J \left(\|r' - r\|_1 + \|c' - c\|_1 \right),$$

where $J = \xi \|C\|_\infty - \log \min_{ij}\{r_i, c_j\}$. For ease of notation, we will let $b := \min_{ij}\{r_i, c_j\}$. Now by (Altschuler et al., 2017, Theorem 2) and the fact that each iteration of `Sinkhorn` takes $\mathcal{O}(mn)$ time, $\texttt{Sinkhorn}(K, r, c, \varepsilon')$ returns a coupling with $X' \in \Pi(r', c')$ satisfying $\|r' - r\|_1 + \|c' - c\|_1 \leq \varepsilon'$ in $\mathcal{O}(mn(\varepsilon')^{-2} \log(s/\ell))$ time where $s = \sum_{ij} K_{ij}$ and $\ell = \min_{ij} K_{ij}$. As C is non-negative, $s = \sum_{ij} e^{-\xi C_{ij}} \leq \sum_{ij} 1 = mn$. Furthermore, $\ell = e^{-\xi \|C\|_\infty}$ so we get a total runtime of $\mathcal{O}(mn(\varepsilon')^{-2}(\log mn + \xi\|C\|_\infty)) = \tilde{\mathcal{O}}(mn(\varepsilon')^{-2}\xi\|C\|_\infty)$. Now choosing $\varepsilon' = \varepsilon^2/8J$, we have

$$\|X' - X_\xi^*\|_1 \leq \sqrt{2J(\|r' - r\|_1 + \|c' - c\|_1)} \leq \sqrt{2J\varepsilon'} = \sqrt{2J\varepsilon^2/8J} = \varepsilon/2.$$

Since $\varepsilon' = \varepsilon^2/8J$, the runtime becomes

$$\begin{aligned}
\tilde{\mathcal{O}}(mn(\varepsilon')^{-2}\xi\|C\|_\infty) &= \tilde{\mathcal{O}}(mn(\varepsilon^2/8J)^{-2}\xi\|C\|_\infty) \\
&= \tilde{\mathcal{O}}(mn\varepsilon^{-4}\xi\|C\|_\infty J^2) \\
&= \tilde{\mathcal{O}}(mn\varepsilon^{-4}\xi\|C\|_\infty(\xi\|C\|_\infty - \log b)^2) \\
&= \tilde{\mathcal{O}}(mn\varepsilon^{-4}\xi\|C\|_\infty(\xi^2\|C\|_\infty^2 + (\log b^{-1})^2)).
\end{aligned}$$

Now we must bound $\|\hat{X} - X'\|_1$. By (Altschuler et al., 2017, Lemma 7), Algorithm 5 returns $\hat{X}$ satisfying

$$\|\hat{X} - X'\|_1 \leq 2(\|r' - r\|_1 + \|c' - c\|_1),$$

in $\mathcal{O}(mn)$ time. So it suffices to check that $\|r' - r\|_1 + \|c' - c\|_1 \leq \varepsilon' = \varepsilon^2/8J$ is enough to guarantee that $\|\hat{X} - X'\|_1 \leq \varepsilon/2$. This will follow immediately from $\|\hat{X} - X'\|_1 \leq 2\varepsilon' = \varepsilon^2/4J \leq \varepsilon/2J$ if we can establish that $J \geq 1$. To see this, first note that $b = \min_{i,j}\{r_i, c_j\} \leq 1/(m \vee n)$. This implies that $-\log b \geq \log(m \vee n)$ and since $\xi > 0$,

$$J = \xi\|C\|_\infty - \log b \geq -\log b \geq \log(m \vee n) \geq 1,$$

assuming that $m \vee n > 2$. If $m \vee n = 2$, then one can check that letting $\varepsilon' = \varepsilon^2 \log 2/8J$ is enough to obtain the desired bounds without affecting the computational complexity. Thus by (3.16), we obtain $\|\hat{X} - X^*_\xi\|_1 \leq \varepsilon$ in time $\tilde{\mathcal{O}}(mn\varepsilon^{-4}\xi\|C\|_\infty(\xi^2\|C\|_\infty^2 + (\log b^{-1})^2) + mn) = \tilde{\mathcal{O}}(mn\varepsilon^{-4}\xi\|C\|_\infty(\xi^2\|C\|_\infty^2 + (\log b^{-1})^2))$. □

Now we can proceed to the proof of Theorem 3.12.

Theorem 3.12. *Let P and Q be aperiodic and irreducible, $h \in \mathbb{R}^{d^2}$, $\xi > 0$, and $\varepsilon > 0$. Then* `EntropicTCI`(h, ξ, ε) *returns $\hat{R} \in ri(\Pi_{TC}(P, Q))$ with $\max_s \|\hat{R}(s, \cdot) - R^*(s, \cdot)\|_1 \leq \varepsilon$ for some $R^* \in \operatorname{argmin}_{R' \in \Pi_{TC}(P,Q)} R'h - {}^1\!/\!_\xi H(R')$ in $\tilde{\mathcal{O}}(d^4\varepsilon^{-4})$ time.*

Proof. Without loss of generality, we may assume that h is non-negative. Otherwise, one can consider the modified bias $h + \|h\|_\infty \mathbb{1}$. Since we are interested in optimal couplings with respect to h rather than expected cost and $\|h + \|h\|_\infty \mathbb{1}\|_\infty = \mathcal{O}(\|h\|_\infty)$, this has no effect on the output of `ApproxOT` or the computational complexity. Now, in order to analyze the complexity of `EntropicTCI`, we must first analyze the complexity of `ApproxOT`. Fix $s = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and, after removing points outside of the supports of $P(x, \cdot)$ and $Q(y, \cdot)$, consider the entropic OT problem for marginal probability measures $P(x, \cdot)$ and $Q(y, \cdot)$ and cost h,

$$\begin{aligned} &\text{minimize} \quad \langle r, h \rangle - \frac{1}{\xi} H(r) \\ &\text{subject to} \quad r \in \Pi(P(x, \cdot), Q(y, \cdot)). \end{aligned} \tag{3.17}$$

Then by (Cuturi, 2013, Lemma 2), there exists a unique solution $r^*_s \in \Pi(P(x, \cdot), Q(y, \cdot))$ to problem (3.17). Furthermore by Lemma 3.18, `ApproxOT`$(P(x, \cdot)^\top, Q(y, \cdot)^\top, h, \xi, \varepsilon)$ returns $\hat{r}_s \in \Pi(P(x, \cdot), Q(y, \cdot))$ such that $\|\hat{r}_s - r^*_s\|_1 \leq \varepsilon$ in $\tilde{\mathcal{O}}(d^2\varepsilon^{-4})$ time. One may also verify using arguments in (Altschuler et al., 2017) that $\hat{r}_s \in \text{ri}(\Pi(P(x, \cdot), Q(y, \cdot)))$.

Now we may analyze the error and computational complexity of `EntropicTCI`(h, ξ, ε). Calling `ApproxOT`$(P(x, \cdot)^\top, Q(y, \cdot)^\top, h, \xi, \varepsilon)$ for every $s = (x, y) \in \mathcal{X} \times \mathcal{Y}$, we obtain $\hat{R} \in \Pi_{\text{TC}}(P, Q)$, where $\hat{R}(s, \cdot) = \hat{r}_s(\cdot)$, in $d^2\tilde{\mathcal{O}}(d^2\varepsilon^{-4}) = \tilde{\mathcal{O}}(d^4\varepsilon^{-4})$ time. Note that since the relative interior commutes with cartesian products of convex sets, $\hat{R} \in \text{ri}(\Pi_{\text{TC}}(P, Q))$. Then defining $R^* \in \Pi_{\text{TC}}(P, Q)$ such that $R^*(s,\cdot) = r_s^*(\cdot)$, we have

$$\max_s \|\hat{R}(s,\cdot) - R^*(s,\cdot)\|_1 = \max_s \|\hat{r}_s - r_s^*\|_1 \leq \varepsilon,$$

by construction. This concludes the proof. □

3.9.5 Proofs from Section 3.6

Our proof of Theorem 3.13 relies on a well-known result regarding the stability of certain optimization problems. Before stating this result, fix spaces $\mathcal{Z}$ and $\mathcal{U}$ corresponding to the set of possible solutions and set of parameters for the optimization problem of interest, respectively. Now consider the following problem.

$$\begin{aligned} \text{minimize} \quad & f(z,u) \\ \text{subject to} \quad & z \in \Phi(u). \end{aligned} \tag{3.18}$$

Note that $f(\cdot,u) : \mathcal{Z} \to \mathbb{R}$ describes the objective to be minimized and $\Phi(u) \subset \mathcal{Z}$ represents the feasible set of Problem (4.5), both indexed by a parameter $u \in \mathcal{U}$. We will call a set $\mathcal{V} \subset \mathcal{Z}$ a neighborhood of a subset $\mathcal{W} \subset \mathcal{Z}$ if $\mathcal{W} \subset \operatorname{int} \mathcal{V}$. Neighborhoods in $\mathcal{U}$ will be defined similarly. Recall that a multifunction $F : \mathcal{U} \to 2^{\mathcal{Z}}$ is upper semicontinuous at a point $u_0 \in \mathcal{U}$ if for any neighborhood $\mathcal{V}_{\mathcal{Z}}$ of the set $F(u_0)$, there exists a neighborhood $\mathcal{V}_{\mathcal{U}}$ of u_0 such that for every $u \in \mathcal{V}_{\mathcal{U}}$, $F(u) \subset \mathcal{V}_{\mathcal{Z}}$.

Theorem A ((Bonnans and Shapiro, 2013, Proposition 4.4))**.** *Let u_0 be a given point in the parameter space $\mathcal{U}$. Suppose that (i) the function $f(z,u)$ is continuous on $\mathcal{Z} \times \mathcal{U}$, (ii) the graph of the multifunction $\Phi(\cdot)$ is a closed subset of $\mathcal{U} \times \mathcal{Z}$, (iii) there exists $\alpha \in \mathbb{R}$ and a compact set $C \subset \mathcal{Z}$ such that for every u in a neighborhood of u_0, the level set $\{z \in \Phi(u) : f(z,u) \leq \alpha\}$ is nonempty and contained in C, (iv) for any neighborhood $\mathcal{V}_{\mathcal{Z}}$ of the set $\operatorname{argmin}_{z\in\Phi(u_0)} f(z,u_0)$ there exists a neighborhood $\mathcal{V}_U$ of u_0 such that $\mathcal{V}_{\mathcal{Z}} \cap \Phi(u) \neq \emptyset$ for all $u \in \mathcal{V}_{\mathcal{U}}$. Then the optimal value function $u \mapsto \min_{z\in\Phi(u)} f(z,u)$ is continuous at $u = u_0$ and the multifunction $u \mapsto \operatorname{argmin}_{z\in\Phi(u)} f(z,u)$ is upper semicontinuous at u_0.*

Both Problems (I) and (II) may be recast in the form of Problem (4.5). Let

$$\mathcal{Z} = \left\{(\lambda, R) \in \Delta_{d^2} \times \Delta_{d^2}^{d^2} : R \in \Pi_{\mathrm{TC}}(P, Q) \text{ for some } P, Q \in \Delta_d^d, \lambda R = \lambda\right\}$$

and $\mathcal{U} = \Delta_d^d \times \Delta_d^d$ be the set of all valid pairs of transition matrices in $\mathbb{R}^{d\times d}$. It is straightforward to verify that $\mathcal{Z}$ and $\mathcal{U}$ are in fact compact subsets of $\mathbb{R}^{d^2} \times \mathbb{R}^{d^2\times d^2}$ and $\mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d}$, respectively. The objective function $f(\cdot)$ is identified with the map $(\lambda, R) \mapsto \langle c, \lambda\rangle$ and does not depend on the parameter $u = (P, Q)$. We will refer to the constraint functions for Problems (I) and (II) by $\Phi : \mathcal{U} \to 2^{\mathcal{Z}}$ and $\Phi_\eta : \mathcal{U} \to 2^{\mathcal{Z}}$, and their optimal solution functions by $\Phi^* : \mathcal{U} \to 2^{\mathcal{Z}}$ and $\Phi^*_\eta : \mathcal{U} \to 2^{\mathcal{Z}}$, respectively.

Theorem 3.13. *Let $P, Q \in \Delta_d^d$ be irreducible transition matrices. Then the following hold:*

- *$\rho(\cdot,\cdot)$ is continuous and $\Phi^*(\cdot,\cdot)$ is upper semicontinuous at (P, Q)*
- *For any $\eta > 0$, $\rho_\eta(\cdot,\cdot)$ is continuous and $\Phi^*_\eta(\cdot,\cdot)$ is upper semicontinuous at (P, Q)*

Proof. We will prove the result for Problem (3.2) as the proof for Problem (3.4) is similar. As the two problems are equivalent, it suffices to check the conditions of Theorem A for Problem (I) at the point $u_0 = (P, Q) \in \mathcal{U}$. First, (i) is vacuously true since the objective $f(\cdot)$ does not depend on u. Next, we will show that the graph of $\Phi(\cdot)$ is a closed subset of $\mathcal{U} \times \mathcal{Z}$. Fix a sequence $\{(P_n, Q_n, \lambda_n, R_n)\}_{n\geq 1} \subset \operatorname{graph}\Phi(\cdot)$. As a subset of the compact set $\Delta_d^d \times \Delta_d^d \times \Delta_{d^2} \times \Delta_{d^2}^{d^2}$, it has a subsequence, which we also label as $\{(P_n, Q_n, \lambda_n, R_n)\}_{n\geq 1}$ converging to some $(P', Q', \lambda', R') \in \Delta_d^d \times \Delta_d^d \times \Delta_{d^2} \times \Delta_{d^2}^{d^2}$. Taking limits of the linear equations $R_n \in \Pi_{\mathrm{TC}}(P_n, Q_n)$ and $\lambda_n R_n = \lambda_n$, we conclude that $R' \in \Pi_{\mathrm{TC}}(P', Q')$ and $\lambda' R' = \lambda'$. Thus $(P', Q', \lambda', R') \in \operatorname{graph}\Phi(\cdot)$ and (ii) holds. To show that (iii) is satisfied, note that one may let $\alpha = \|c\|_\infty$ and use the fact that the entire set $\mathcal{Z}$ is compact. Finally, we will show that (iv) is satisfied. Let $\mathcal{V}_{\mathcal{Z}} \subset \mathcal{Z}$ be a neighborhood of $\operatorname{argmin}_{z\in\Phi(u_0)} f(z, u_0)$. Then define the neighborhood $\mathcal{V}_{\mathcal{U}}$ of $u_0 = (P, Q)$ as

$$\mathcal{V}_{\mathcal{U}} := \{(P, Q) \in \Delta_d^d \times \Delta_d^d : R \in \Pi_{\mathrm{TC}}(P, Q) \text{ for some } (\lambda, R) \in \mathcal{V}_{\mathcal{Z}}\}.$$

Note that $\mathcal{V}_{\mathcal{U}}$ is nonempty by the non-emptiness of $\mathcal{V}_{\mathcal{Z}}$ and the definition of $\mathcal{Z}$. Moreover, $\mathcal{V}_{\mathcal{Z}} \cap \Phi(u) \neq \emptyset$ for all $u \in \mathcal{V}_{\mathcal{U}}$ by construction. Thus all the conditions of Theorem A are satisfied and the desired convergence holds. □

CHAPTER 4
Graph Optimal Transport via Optimal Transition Coupling

In this chapter, we present a novel approach to optimal transport between graphs from the perspective of stationary Markov chains. A weighted graph may be associated with a stationary Markov chain by means of a random walk on the vertex set with transition distributions depending on the edge weights of the graph. After drawing this connection, we describe how optimal transport techniques for stationary Markov chains may be used in order to perform comparison and alignment of the graphs under study. In particular, we propose the graph optimal transition coupling problem, referred to as GraphOTC, in which the Markov chains associated to two given graphs are optimally synchronized to minimize an expected cost. The joint synchronized chain yields an alignment of the vertices and edges in the two graphs, and the expected cost of the synchronized chain acts as a measure of distance or dissimilarity between the two graphs. We demonstrate that GraphOTC performs equal to or better than existing state-of-the-art techniques in graph optimal transport for several tasks and datasets. Finally, we also describe a generalization of the GraphOTC problem, called the FusedOTC problem, from which we recover the GraphOTC and OT costs as special cases.

4.1 Introduction

In graph comparison tasks, one aims to assess the similarity or dissimilarity of two graphs by means of their topologies and vertex characteristics. In graph alignment or matching tasks, one aims to associate the vertices and edges in a graph with similar vertices and edges in another graph. Both comparison and alignment are of fundamental importance in the study of graphs in machine learning and data science. Examples of applications include image captioning (Chen et al., 2020), object recognition (Yan et al., 2018), domain adaptation (Malka et al.), aligning single-cell multi-omics data (Demetci et al., 2020), and language comparison (Alvarez-Melis and Jaakkola, 2018).

Some recent work (Peyré et al., 2016; Titouan et al., 2019; Maretic et al., 2019, 2020; Dong and Sawin, 2020) addresses the problems of graph comparison and alignment using techniques from optimal transport. In the optimal transport (OT) problem, one seeks a plan for transporting mass between two probability measures of interest that minimizes expected cost. When each graph under study is associated with a probability distribution and a cost or distance between the vertices in each pair of graphs is available, graph comparison and alignment both fit naturally into the framework of OT. In particular, optimal transport plans between the probability distributions correspond to alignments between the graphs, while the expected cost of transportation serves as a measure of dissimilarity or distance between graphs of interest.

To date, most OT-based approaches to graph comparison and alignment fall into one of two categories: spectral methods or methods based on the Gromov-Wasserstein distance. In the spectral approaches, each graph is associated with a zero-mean, multivariate normal distribution whose covariance matrix is a function of the graph Laplacian. Graph OT is then performed via OT between the associated normal distributions of the graphs of interest. In theory and example, one finds that these methods emphasize differences in global structure, i.e., graph structure that is robust to small changes in vertices and edges. On the other hand, the Gromov-Wasserstein approach associates each graph with a discrete distribution over its vertex set and aims to couple these probability measures so as to minimize changes in the edge weights of either graph. The resulting graph OT problem emphasizes differences in local structure over global structure.

In this chapter, we describe a novel approach to OT for graph comparison and alignment that balances differences in both global and local structure. We first propose to associate each graph with a stationary Markov chain on its vertex set, with transition probabilities depending on the edge weights of the graph. After drawing this connection, we define an OT problem for graphs by means of an OT problem between the associated Markov chains of the graphs of interest. Leveraging recent work in OT for Markov chains, we define the GraphOTC problem, which aims to find a product graph with an associated Markov chain that minimizes the expectation of the pre-specified cost with respect to the chain's stationary distribution. This new approach can incorporate local structure by means of the cost function between the vertices of either graph, and it accounts for global structure via the stationary distributions of the associated Markov chains. GraphOTC can easily accommodate many distinct types of cost functions, including cost functions that depend on the intrinsic local structure of the graphs and cost functions that involve external features of the nodes. Furthermore, GraphOTC is easily interpretable and does not rely on selecting any free parameters. Finally, drawing inspiration from the Fused Gromov-Wasserstein problem (Titouan et al., 2019), we situate the GraphOTC problem within a broader theoretical context via a graph OT framework that we call FusedOTC.

Contributions. The contributions of this chapter are as follows: (a) we define the GraphOTC problem for graphs that extends OT techniques for Markov chains to graphs; (b) we demonstrate that when the underlying cost satisfies the properties of a metric, the associated GraphOTC cost is a metric on certain equivalence classes of graphs; (c) we demonstrate that the performance of GraphOTC equals or surpasses the state-of-the-art in graph OT for several graph alignment and graph comparison tasks; (d) we describe a broader graph OT framework, known as FusedOTC, that includes the GraphOTC and OT costs as special cases.

4.1.1 Related work

Spectral methods. One line of work (Maretic et al., 2019, 2020; Dong and Sawin, 2020) uses graph spectral techniques to define OT problems for graphs. In particular, this approach associates to each graph a multivariate Gaussian with zero mean and covariance matrix equal to the pseudoinverse of the graph Laplacian. The Wasserstein distance between Gaussians in the same space may be computed analytically in terms of the respective covariance matrices. For graphs with different numbers of vertices, (Maretic et al., 2020) and (Dong and Sawin, 2020) propose to optimize this distance over soft many-to-one assignments between vertices in either graph. At present, this family of approaches is unable to incorporate available feature information or underlying cost functions, relying completely on intrinsic structure in their respective optimization problems.

Variants of Gromov-Wasserstein. Another line of work (Mémoli, 2011; Peyré et al., 2016; Titouan et al., 2019; Vayer et al., 2019, 2020) considers the Gromov-Wasserstein (GW) distance and related extensions. In this work, one tries to couple distributions on the nodes in each graph so as to minimize an expected transport cost between vertices while minimizing changes in edges between the two graphs. This approach allows one to capture differences in both features and structure between graphs. We refer the reader to (Dong and Sawin, 2020) for a discussion on the differences between spectral-based graph OT methods and GW distances. A number of variants of the GW distance have been proposed for a variety of tasks including cross-domain alignment (Chen et al., 2020), graph partitioning (Xu et al., 2019a), graph matching (Xu et al., 2019a,b), and node embedding (Xu et al., 2019b). The work (Barbe et al., 2020) proposes to incorporate global structure into the Wasserstein and Fused GW distances by applying a heat diffusion to the vertex features before computing the cost matrix.

4.2 Preliminaries

Notation. Let $\mathbb{R}_+$ be the set of non-negative real numbers and for every $n \geq 1$, let $\Delta_n = \{\mu \in \mathbb{R}_+^n \,|\, \sum_{i=1}^n \mu_i = 1\}$ be the probability simplex in $\mathbb{R}^n$. For a Polish space $\mathcal{U}$, we will use $\mathcal{M}(\mathcal{U})$ to denote the set of Borel probability measures on $\mathcal{U}$. Note that whenever the set $\mathcal{U}$ is finite, we will frequently regard probability measures in $\mathcal{M}(\mathcal{U})$ as vectors in $\Delta_{|\mathcal{U}|}$. We define the inner product $\langle \cdot, \cdot \rangle$ for matrices $U, V \in \mathbb{R}^{m \times n}$ by $\langle U, V \rangle = \sum_{i,j} U_{ij} V_{ij}$. For a vector $u \in \mathbb{R}^m$ and matrix $U \in \mathbb{R}^{m \times n}$, we will denote by $u \odot U$ the matrix satisfying $(u \odot U)_{ij} = u_i U_{ij}$. Graphs will be denoted by triples $\mathcal{G} = (V, E, w)$ where V is the vertex set, $E \subset V \times V$ is the edge set, and $w : E \to \mathbb{R}$ is a function that gives the weight of each edge. All graphs considered in this paper will be assumed to be undirected and connected. For unweighted graphs, we define $w(e) = 0$ for all $e \in E$ by convention.

4.2.1 Optimal Transport

Recall the setting of the optimal transport problem for finite spaces: let $\mu \in \mathcal{M}(\mathcal{U})$ and $\nu \in \mathcal{M}(\mathcal{V})$ be probability measures on finite spaces $\mathcal{U}$ and $\mathcal{V}$, respectively, and let $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$ be a non-negative function. Transport plans in the optimal transport (OT) problem are formalized mathematically as *couplings*: A probability measure $\pi \in \mathcal{M}(\mathcal{U} \times \mathcal{V})$ is a coupling of μ and ν if $\pi \mathbb{1} = \mu$ and $\pi^\top \mathbb{1} = \nu$. We will denote the set of couplings of μ and ν by $\Pi(\mu, \nu)$. The OT problem is to minimize the expectation of c over the set $\Pi(\mu, \nu)$:

$$\min \left\{ \langle \pi, c \rangle : \pi \in \Pi(\mu, \nu) \right\}. \tag{4.1}$$

The optimal value of Problem (4.1) is referred to as the optimal transport cost and optimal solutions are referred to as optimal transport plans or optimal couplings. For more details on the OT problem in general, the reader may consult (Peyré and Cuturi, 2019).

4.2.2 Optimal Transition Couplings of Markov chains

In Chapter 3, the OT problem (4.1) was adapted to the setting of Markov chains, resulting in an optimization problem called the optimal transition coupling (OTC) problem. Let $P \in [0,1]^{|\mathcal{U}|\times|\mathcal{U}|}$ and $Q \in [0,1]^{|\mathcal{V}|\times|\mathcal{V}|}$ be aperiodic and irreducible transition matrices on $\mathcal{U}$ and $\mathcal{V}$, respectively. A transition coupling of P and Q is a transition matrix $R \in [0,1]^{|\mathcal{U}||\mathcal{V}|\times|\mathcal{U}||\mathcal{V}|}$ satisfying

$$\sum_{\tilde{v}\in\mathcal{V}} R((u,v),(u',\tilde{v})) = P(u,u') \qquad \text{and} \qquad \sum_{\tilde{u}\in\mathcal{U}} R((u,v),(\tilde{u},v')) = Q(v,v'),$$

for every $(u,v),(u',v') \in \mathcal{U}\times\mathcal{V}$. In other words, the rows of the joint transition matrix R are couplings of the rows of the transition matrices P and Q. Every transition coupling corresponds to a stationary Markov chain taking values in $\mathcal{U}\times\mathcal{V}$. Furthermore, every such Markov chain is necessarily a coupling of the Markov chains corresponding to the transition matrices P and Q, and thus one may define an OT problem for Markov chains in terms of transition couplings (see Chapter 3). We will denote the set of transition couplings of P and Q by $\Pi_{\mathrm{TC}}(P,Q)$.

The OTC problem is to minimize the expectation of c over the set of stationary distributions of transition couplings of P and Q:

$$\min\left\{\langle\lambda,c\rangle : R \in \Pi_{\mathrm{TC}}(P,Q), \lambda R = \lambda, \lambda \in \Delta_{|\mathcal{U}|\times|\mathcal{V}|}\right\}. \tag{4.2}$$

As discussed in Chapter 3, one may view the OTC problem as trying to synchronize the Markov chains corresponding to P and Q with respect to the long-run average of the cost c. The optimal pair (λ,R) characterizes the synchronized chain, while the minimal expected cost $\langle\lambda,c\rangle$ provides a measure of dissimilarity between P and Q. Importantly, this dissimilarity measure emphasizes long-term average differences over short-term behavior. This feature is due to the stationarity constraint in the OTC problem and will be key in enabling our proposed approach to graph OT to capture differences in global structure despite being defined in terms of a cost between vertices.

4.3 GraphOTC

Relationships between graphs and Markov chains have been studied extensively in the literature (Lovász, 1993; Aldous and Fill, 2002; Levin and Peres, 2017). We propose to leverage this connection in order to define a new OT problem for graphs. In particular, we view each graph in terms of a stationary random walk on its vertex set and apply the OTC problem to the random walks associated with the graphs of interest. Formally, to any graph $\mathcal{G} = (V, E, w)$, we associate a transition matrix $P \in [0,1]^{|V|\times|V|}$, defined as follows: Let

$$P(u,u') = \frac{\exp\{w(u,u')\}}{\sum_{\tilde{u}:(u,\tilde{u})\in E} \exp\{w(u,\tilde{u})\}}, \qquad \forall (u,u') \in E,$$

and $P(u,u') = 0$ otherwise. If P is irreducible, then it defines a unique stationary Markov chain, which we also refer to as P when no confusion may arise. Using this construction, connected graphs will correspond to irreducible Markov chains. Moreover, connected graphs with at least one self loop will correspond to aperiodic Markov chains, and we note that any irreducible Markov chain can be made aperiodic by considering the "lazy" chain instead (see (Levin and Peres, 2017, p. 9)). For the rest of the chapter, we will associate the graphs of interest, $\mathcal{G}_1$ and $\mathcal{G}_2$, with the stationary Markov chains P and Q, respectively. We will also assume for convenience that P and Q are aperiodic and irreducible.

The random walk on a graph encodes important geometric properties of the graph, and it has tight connections to both local properties of the graph (such as edge weights and generalized node degrees) and global properties of the graph (such as the Laplacian and its spectrum). Indeed, the stationary distribution of the random walk gives a notion of the global importance of the vertices within the graph. For these reasons, coupling the random walks of two graphs provides a natural and informative setting for performing graph OT.

In light of this perspective, we now define the GraphOTC problem as

$$d(\mathcal{G}_1, \mathcal{G}_2) = \min\left\{\langle \lambda, c\rangle : R \in \Pi_{\mathrm{TC}}(P,Q), \lambda R = \lambda, \lambda \in \Delta_{|V_1|\times|V_2|}\right\}. \tag{4.3}$$

Note that in addition to a coupling cost of $\mathcal{G}_1$ and $\mathcal{G}_2$, we also obtain an optimal transition coupling of the Markov chains P and Q. In particular, if (λ, R) is an optimal solution to Problem (4.3), then $\lambda(u,v)$ describes the alignment probability of vertices $u \in V_1$ and $v \in V_2$, while $\lambda(u,v)R((u,v),(u',v'))$ describes the alignment probability of the edges $(u,u') \in E_1$ and $(v,v') \in E_2$. Note that the optimal transition coupling also provides an alignment of higher-order paths, but we do not explore this observation any further here.

The GraphOTC problem brings the tools of Markov chain OT to bear on graphs. Moreover, GraphOTC incorporates both local information (by means of a cost) and global structure (via transition couplings). As we will demonstrate empirically in Sections 4.4 and 4.5, this results in an approach to graph OT that automatically balances vertex and edge information with topological structure in performing alignment and comparison.

GraphOTC is a metric. Building upon previous results in stationary optimal transport, one may establish that the GraphOTC cost is a metric on a certain space of graphs when the cost c is a metric. We will say that $\mathcal{G}_1 \sim \mathcal{G}_2$ for two graphs $\mathcal{G}_1 = (V, E, w_1)$ and $\mathcal{G}_2 = (V, E, w_2)$ if there exists $C \in \mathbb{R}$ such that $w_1(u,v) = w_2(u,v) + C$ for every $(u,v) \in E$. We prove in the Section 4.8 that $\mathcal{G}_1 \sim \mathcal{G}_2$ if and only if their respective random walks are identical.

Theorem 4.1. *Suppose that the cost function $c : V \times V \to \mathbb{R}_+$ satisfies the properties of a metric on V. Then d is a metric on the equivalence classes defined by $\sim$.*

The proof of Theorem 4.1 is deferred to Section 4.8.

4.3.1 GraphOTC with Intrinsic Cost Functions

The GraphOTC problem, along with several other graph OT methods, relies on the specification of a cost function $c : V_1 \times V_2 \to \mathbb{R}_+$. Commonly, the cost c may be derived from features associated with vertices in the graphs or from distances between vertices when both graphs are embedded in a common metric space. For example, if one has access to label functions $\ell_1 : V_1 \to \mathcal{A}$ and $\ell_2 : V_2 \to \mathcal{A}$ associating each vertex in V_1 and V_2 with a label in a finite alphabet $\mathcal{A}$, then one may let $c(u,v) = \delta(\ell_1(u) \neq \ell_2(v))$. We refer to this function as the 0-1 cost for the label functions ℓ_1 and ℓ_2. Alternatively, if there exist maps $f_1 : V_1 \to \mathcal{U}$ and $f_2 : V_2 \to \mathcal{U}$ taking vertices in V_1 and V_2 to points in a metric space $(\mathcal{U}, \rho)$, then a cost may be defined as $c(u,v) = \rho(f_1(u), f_2(v))$ or $c(u,v) = \rho(f_1(u), f_2(v))^2$.

However, in some contexts such vertex features may unavailable in practice, and one may wish to define a cost function when one is not given. In this case, one may consider costs defined in terms of intrinsic properties of the graphs of interest. For example, letting $D_i : V_i \to \mathbb{N}$ be the degree function for graph $\mathcal{G}_i$, we might define the cost as $c_{\text{deg}}(u, v) = (D_1(u) - D_2(v))^2$. Alternatively, one may consider costs based on the degree distributions of the neighborhoods of u and v. Another approach is to embed the vertices in a Euclidean space *a priori* using graph embedding methods such as Laplacian eigenmaps (Belkin and Niyogi, 2003). We demonstrate in Section 4.4 that the squared-degree cost c_{deg} can adequately capture important graph structure when used with the GraphOTC problem.

4.3.2 Solving the GraphOTC Problem

Given the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, one may associate to each graph a stationary Markov chain, as described above. Once these Markov chains have been defined, solving the GraphOTC problem amounts to solving the OTC problem. Despite the non-convexity of the OTC problem, it was shown in Chapter 3 that one may obtain solutions via an adaptation of the policy iteration algorithm (Howard, 1960), referred to as `ExactOTC`. The algorithm `ExactOTC` exhibits a runtime scaling like $\mathcal{O}((|V_1||V_2|)^3)$ per iteration. In practice, convergence is typically observed after less than 5 iterations. A more efficient algorithm based on entropic regularization and Sinkhorn iterations was proposed in Chapter 3 and is known as `EntropicOTC`. This algorithm was observed to scale better with the sizes of the marginal state spaces, yielding a runtime of $\mathcal{O}((|V_1||V_2|)^2)$ per iteration (ignoring poly-logarithmic factors) in the case of GraphOTC. This runtime is nearly-linear in the dimension of couplings under consideration and in this sense is comparable to the state-of-the-art for entropic OT algorithms (Peyré and Cuturi, 2019).

4.4 Examples

In this section, we compare GraphOTC to existing approaches for graph OT in a few examples.

4.4.1 Stochastic Block Models

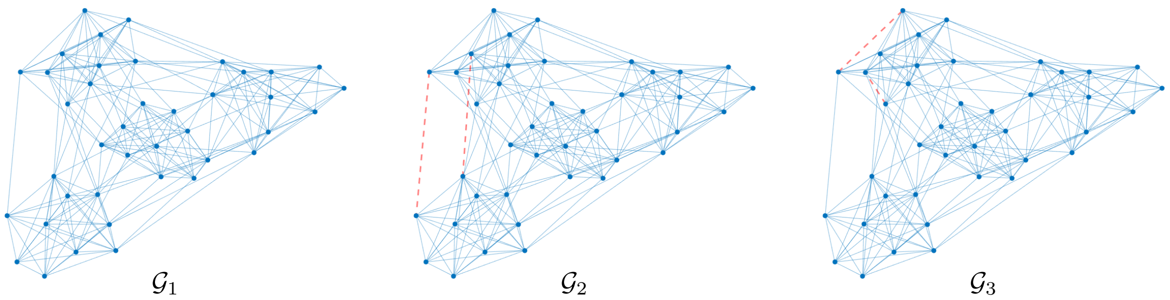


Figure 4.1: Three registered graphs with block structure. $\mathcal{G}_1$ was generated from a stochastic block model, $\mathcal{G}_2$ was obtained by removing two edges between blocks, and $\mathcal{G}_3$ was obtained by removing two edges within blocks. The removed edges of each graph are highlighted with red dashed lines.

Stochastic block models (Holland et al., 1983) (SBMs) are a common model for random graphs with group (or community) structure, and they have been used in a variety of applications, including community detection and graph clustering (Abbe, 2018; Lee and Wilkinson, 2019; Abbe and Sandon, 2015). In an SBM, nodes are grouped into *blocks* representing communities within the graph. Edges are drawn between nodes independently at random with probabilities depending on whether the two nodes are within the same block or not. Generally, connection probabilities are higher within blocks than between blocks, leading to more densely connected subgraphs.

Table 4.1: Costs for graphs in Figure 4.1.

Algorithm	Chosen Cost	$\mathcal{G}_1$ vs. $\mathcal{G}_2$	$\mathcal{G}_1$ vs. $\mathcal{G}_3$	Ratio
GraphOTC	0-1	0.0396	0.0265	1.49
GraphOTC	c_{deg}	0.1718	0.1315	1.31
GW	–	0.1150	0.1125	1.02
GOT	–	0.0039	0.0019	2.05

In order to develop some intuition about the behavior of various graph OT methods, we apply the GraphOTC distance along with existing graph OT methods to a selection of three graphs with block structure. The graphs of interest are depicted in Figure 4.1. $\mathcal{G}_1$ was drawn from an SBM with 40 nodes and 4 blocks, $\mathcal{G}_2$ was obtained by removing two edges between blocks from $\mathcal{G}_1$, and $\mathcal{G}_3$ was obtained by removing two edges within blocks from $\mathcal{G}_1$. Intuitively, the topology of a graph with block structure will depend more strongly on edges between blocks than within blocks and thus we regard $\mathcal{G}_2$ as more dissimilar to $\mathcal{G}_1$ than $\mathcal{G}_3$ is to $\mathcal{G}_1$. In Table 4.1, we provide the distances computed by GraphOTC for two different costs, as well as distances computed by two other graph OT methods known as GW (Maretic et al., 2020) and GOT (Maretic et al., 2019), respectively. We find that GraphOTC and GOT regard $\mathcal{G}_2$ as more dissimilar from $\mathcal{G}_1$ than $\mathcal{G}_3$ is. On the other hand, GW detects little difference in the dissimilarity of the graphs. Similar results were observed for other choices of removed edges.

4.4.2 Wheel Graphs

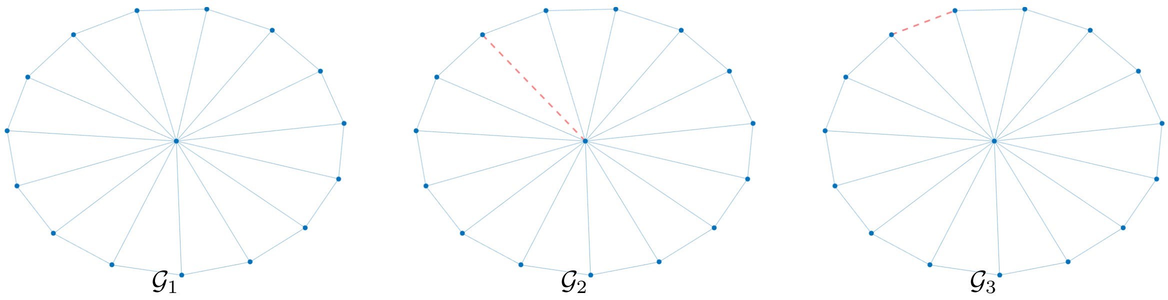


Figure 4.2: Three registered wheel graphs. $\mathcal{G}_1$ is a wheel graph of order 16, $\mathcal{G}_2$ was obtained by removing a spoke, and $\mathcal{G}_3$ was obtained by removing a wheel edge. The removed edges of each graph are highlighted with red dashed lines.

A wheel graph of order n is a graph containing a cycle of order $n-1$ such that every node in the cycle is connected to a central node called a *hub*. The edges contained in the cycle are called wheel edges, and the edges connected to the hub are called *spoke edges* or *spokes* for short. We denote a wheel graph with n nodes by W_n.

Table 4.2: Costs for graphs in Figure 4.2.

Algorithm	Chosen Cost	$\mathcal{G}_1$ vs. $\mathcal{G}_2$	$\mathcal{G}_1$ vs. $\mathcal{G}_3$	Ratio
GraphOTC	0-1	0.0838	0.0607	1.38
GraphOTC	c_{deg}	2.6552	2.5517	1.04
GW	–	0.0512	0.0547	0.93
GOT	–	0.0459	0.0736	0.62

In Figure 4.2, we apply several different graph OT methods to two pairs of wheel-type graphs. The graph $\mathcal{G}_1$ is a wheel graph W_{16}, and $\mathcal{G}_2$ and $\mathcal{G}_3$ are both copies of $\mathcal{G}_1$ with one edge removed. $\mathcal{G}_2$ was obtained by removing a spoke and $\mathcal{G}_3$ was generated by removing a wheel edge from $\mathcal{G}_1$. Since the hub node is connected with all other nodes, it is desirable for the topology of the wheel graph to depend more strongly on spokes than on wheel edges. Spokes contain the hub, while wheel edges do not. Thus, we expect $\mathcal{G}_3$ to be more similar to $\mathcal{G}_1$ than $\mathcal{G}_2$ is to $\mathcal{G}_1$. In Table 4.2, we see GraphOTC detects that spokes are more influential than wheel edges, while GW and GOT do not. This provides some evidence that our approach accounts for differences in global structure when comparing graphs.

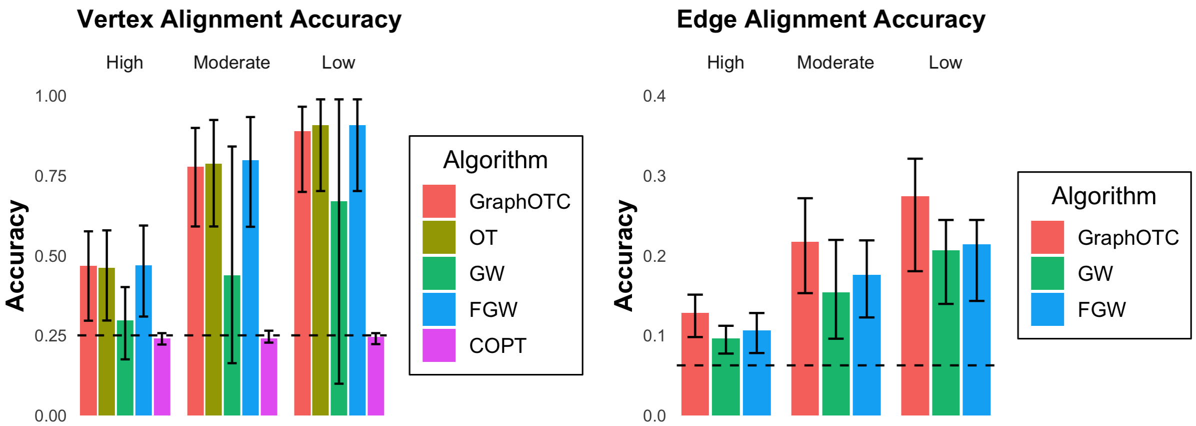


Figure 4.3: Point cloud alignment accuracies. We evaluate the alignment accuracies of each method in three different regimes: high overlap, moderate overlap, and low overlap (defined formally in Appendix B.1.1). Accuracies reported are the average observed over 5 random pairs of point clouds. The horizontal dashed line in each plot indicates the accuracy of random guessing.

4.5 Experiments

In this section, we demonstrate the performance of GraphOTC on point cloud alignment and graph classification tasks. Complete experimental details may be found in Appendix B.1. In both experiments, we compare GraphOTC to the following graph OT baselines: standard optimal transport cost (OT), Gromov-Wasserstein (GW) (Peyré et al., 2016), Fused Gromov-Wasserstein (FGW) (Titouan et al., 2019; Vayer et al., 2020), and Coordinated Optimal Transport (COPT) (Dong and Sawin, 2020). We remark that GOT (Maretic et al., 2020) solves an optimization problem which is nearly identical to COPT, so we omit it from our comparison. Code for reproducing the experiments may be found at `https://github.com/oconnor-kevin/GraphOTC`.

4.5.1 Point Cloud Alignment

In our first experiment, we consider the task of aligning graphs derived from point clouds. We consider randomly generated point clouds $D_1 = \{x_{1,1}, ..., x_{1,N_1}\}$ and $D_2 = \{x_{2,1}, ..., x_{2,N_2}\}$ in $\mathbb{R}^3$, independently drawn from a 4-component Gaussian mixture model as described in Appendix B.1.1. Given the point clouds D_1 and D_2, the graph $\mathcal{G}_1 = (V_1, E_1, w_1)$ is defined to be the graph with vertex set $V_1 = D_1$, with edges and edge weights chosen as follows. Let $\lambda \geq 0$ and $\tilde{w}_1 \in \mathbb{R}^{|V_1| \times |V_1|}$ be the matrix such that for all $v_j, v_k \in V_1$, let $\tilde{w}_1(v_j, v_k) = -\lambda \|x_{1,j} - x_{1,k}\|_2$. In particular, λ allows one to tune the relative importance of the pairwise distances in determining the edge weights. If $\lambda = 0$, then all edges are given equal weight, while as λ becomes large, only edges between points that are very close to one another are given a non-negligible weight. Next, let w_1 be the edge weights obtained by rescaling the weights $\tilde{w}_1$ to lie in the interval $[0, 1]$ and then setting all weights that fall below 0.1 to 0. Finally, we remove all edges with weight equal to 0. The graph $\mathcal{G}_2$ corresponding to the other point cloud is constructed in the same way with the same λ.

In each iteration, we generate two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ as described above and apply each of the aforementioned graph OT methods to align the two graphs. Each method returns a soft alignment of the vertex sets V_1 and V_2 in the form of a coupling $\pi \in \Pi(p, q)$. Similarly, all methods other than COPT return an alignment of edges in the form of a 2-step coupling $\pi_{\text{2-step}} \in \mathcal{M}((V_1 \times V_2)^2)$. In the case of GW and FGW, $\pi_{\text{2-step}} = \pi \otimes \pi \in \Pi(p \otimes p, q \otimes q)$ while for GraphOTC, $\pi_{\text{2-step}} = \lambda \odot R \in \Pi(p \odot P, q \odot Q)$, for some (λ, R) in the OTC constraint set.

In Figure 4.3, we plot the vertex and edge alignment accuracies for each of the methods tested in the case $\lambda = 10^{-2}$. Similar results were observed for other values of λ in $\{10^{-5}, 10^{-4}, ..., 10^{-1}\}$. Vertex alignment accuracy was assessed by summing the mass of the optimal coupling of the two graphs for pairs of vertices that were generated from the same mixture components. We find that OT, FGW, and GraphOTC perform roughly equivalently in the task of vertex alignment, while GW and COPT exhibit worse performance. In particular, COPT is only able to achieve an alignment accuracy roughly equivalent to random guessing. We suspect that this is because COPT is not able to take the geometric information into account in aligning the point cloud graphs.

We also compared the edge alignment accuracies of each algorithm in each of the three overlap regimes. Edge alignment accuracy was evaluated by summing the mass of the 2-step optimal coupling $\pi_{\text{2-step}}$ for pairs of edges connecting identical mixture components. We notice that GraphOTC outperforms FGW and GW from the standpoint of edge alignment. This provides further evidence that GraphOTC is able to balance local information (distances between nodes) with higher order structure in comparing the graphs of interest. Finally, we remark that the edge alignment accuracies were much lower than the vertex alignment accuracies observed for all algorithms. We suspect that this is due at least in part to the increased difficulty of edge alignment. In particular, random guessing for edge alignment yields an accuracy of 6.25% vs. 25% for vertex alignment.

4.5.2 Graph Classification

In our next experiment, we demonstrate the utility of GraphOTC in a graph classification task. We consider a selection of benchmark graph datasets from (Kersting et al., 2016) containing discrete vertex attributes as well as a class label for every graph. This collection of datasets includes AIDS (Riesen and Bunke, 2008), BZR (Sutherland et al., 2003), Cuneiform (Kriege et al., 2018), MCF-7 (Yan), MOLT-4 (Yan), MUTAG (Debnath et al., 1991), and Yeast (Yan). We obtain a cost function from the vertex attributes by letting the cost for a pair of vertices be equal to 0 if their labels are identical and 1 otherwise. Using this cost function, we fit a simple 5-nearest neighbor classifier using each of the graph OT costs to a randomly-sampled training set of graphs consisting of 80% of the data. In Table 4.3, we report the average classification accuracy observed on the held-out test set for each graph OT cost and dataset over 5 random samplings of the training and test sets.

Table 4.3: 5-nearest neighbor classification accuracies for graphs with discrete node attributes. Average accuracies observed over 5 random samplings of the training and test sets are reported along with their standard deviation.

Algorithm	AIDS	BZR	Cuneiform	MCF-7	MOLT-4	MUTAG	Yeast
GraphOTC	88.0 ± 4.9	**84.8 ± 6.6**	**73.2 ± 7.8**	92.8 ± 4.2	**92.0 ± 2.0**	**85.4 ± 7.1**	90.8 ± 6.4
OT	84.4 ± 6.1	76.4 ± 4.6	71.3 ± 7.7	**93.6 ± 3.3**	**92.0 ± 2.0**	63.2 ± 7.3	91.2 ± 7.0
GW	98.8 ± 1.8	78.0 ± 8.5	12.8 ± 4.6	**93.6 ± 3.3**	91.6 ± 2.6	81.6 ± 7.0	**91.6 ± 6.2**
FGW	**99.2 ± 1.1**	80.4 ± 7.4	71.7 ± 7.2	92.8 ± 4.2	91.2 ± 2.3	83.8 ± 8.3	90.0 ± 5.5
COPT	98.0 ± 1.4	73.6 ± 7.9	16.6 ± 3.1	92.4 ± 4.8	91.6 ± 2.6	80.0 ± 5.6	90.4 ± 6.7

In Table 4.3, we see that GraphOTC outperforms the baseline methods in several cases. We emphasize that GraphOTC is competitive with other graph OT methods without the need for tuning any hyperparameters. This suggests that GraphOTC sufficiently captures important differences among the graphs of interest by default and provides further support for the proposed approach. When taking standard deviations into account, we do not see significant differences in performance between the algorithms in most cases. Therefore, it would be helpful to perform a more detailed study in the future with a greater number of training-test set randomizations.

4.6 FusedOTC

In this section, we describe a more general framework of graph OT problems that includes both GraphOTC and OT of the stationary distributions as extremal cases. This more general framework is analogous to the Fused Gromov-Wasserstein distance proposed in (Titouan et al., 2019). Specifically, in order to more flexibly capture global and local differences in the graphs of interest, we augment the GraphOTC objective, adding a term that penalizes changes in edge weights (as in GW). We refer to the resulting graph OT problem as FusedOTC. We show in Theorem 4.2 that one may recover both the GraphOTC and standard OT costs as special cases of FusedOTC by taking appropriate limits.

Let $\alpha \in [0, 1]$ and let $p \in \Delta_{|V_1|}$ and $q \in \Delta_{|V_2|}$ be the stationary distributions of the chains P and Q. The Fused Gromov-Wasserstein (FGW) problem (Titouan et al., 2019) for graphs $\mathcal{G}_1 = (V_1, E_1, w_1)$ and $\mathcal{G}_2 = (V_2, E_2, w_2)$ may be written in a simplified form as

$$d_\alpha^{\mathrm{FGW}}(\mathcal{G}_1, \mathcal{G}_2) = \min \left\{ \alpha \langle \pi, c \rangle + (1 - \alpha) \langle \pi \otimes \pi, E \rangle : \pi \in \Pi(p, q) \right\}, \tag{4.4}$$

where $E \in \mathbb{R}^{|V_1||V_2| \times |V_1||V_2|}$ is the matrix satisfying $E((u, v), (u', v')) = |w_1(u, u') - w_2(v, v')|$. Note that compared to the standard OT distance between p and q, the objective in Problem (4.4) includes an additional term penalizing changes in edge weights. Depending on one's choice of α, the FGW problem will prioritize coupling similar vertices or similar edge weights. In the case that $\alpha = 1$, one recovers the standard OT problem for marginals p and q and cost c, while in the case that $\alpha = 0$, one recovers the Gromov-Wasserstein problem for graphs $\mathcal{G}_1$ and $\mathcal{G}_2$.

We may extend the flexibility of the FGW problem to GraphOTC in a straightforward manner, yielding a graph OT problem that can adaptively balance expected cost with correct edge coupling while taking global graph structure into account. We refer to this problem as the FusedOTC problem. Fixing parameters $\alpha \in [0, 1]$ and $\tau \in \mathbb{N}$, the FusedOTC problem is defined as

$$d_{\tau,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = \min \left\{ \alpha \langle \lambda, c \rangle + (1 - \alpha) \langle \lambda \odot R, E \rangle : R \in \Pi_{\mathrm{TC}}(P^\tau, Q^\tau), \lambda R = \lambda, \lambda \in \Delta_{|V_1| \times |V_2|} \right\}.$$

The FusedOTC problem is highly flexible and generalizes both the GraphOTC problem as well as the standard OT problem between the stationary distributions of the chains P and Q. In the case $\alpha = 1$ and $\tau = 1$, we recover the GraphOTC problem, while in the case $\alpha = 0$ and $\tau \in \mathbb{N}$, we obtain a new graph OT problem that is independent of the cost c. Relying on basic results in Markov chain theory, we expect that in the limit $\tau \to \infty$, the chains P^τ and Q^τ become IID. In this limit and with $\alpha = 1$, we recover the Wasserstein distance (d^{W}) between the stationary distributions $p \in \Delta_{|V_1|}$ and $q \in \Delta_{|V_2|}$ of the Markov chains P and Q. These properties are formalized in the following theorem, whose proof appears in Section 4.8.

Theorem 4.2. *The FusedOTC cost $d_{\tau,\alpha}$ satisfies the following:*

- $\lim_{\alpha \to 1} d_{1,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = d(\mathcal{G}_1, \mathcal{G}_2)$
- $\lim_{(\tau,\alpha)\to(\infty,1)} d_{\tau,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = d^W(\mathcal{G}_1, \mathcal{G}_2)$

4.7 Discussion

In this chapter, we proposed the GraphOTC problem for comparing and aligning graphs. This new approach to graph OT applies ideas from constrained OT for Markov chains to the random walks associated to the graphs. In theory and practice, we demonstrated that GraphOTC balances differences in both global and local structure. In synthetic and real data experiments, we showed that GraphOTC exhibits equal or better performance to state-of-the-art graph OT methods in both comparison and alignment tasks. We also described a more flexible framework for graph OT known as FusedOTC, from which one may recover both GraphOTC as well as standard OT as special cases. Future work may aim to develop computationally tractable algorithms for solving the FusedOTC problem and explore principled means of selecting the hyperparameters α and τ in practice.

Graph-structured data may be found in a wide variety of application areas. The GraphOTC problem offers a novel approach to studying this data from the perspective of Markov chains and optimal transport. We showed that GraphOTC achieves the state-of-the-art in graph OT for a 5-nearest neighbor classification task for several real datasets including BZR, MOLT-4, and MUTAG. Graphs in each of these datasets describe molecular structure for different compounds of interest. Consequently, GraphOTC may find application in this area, enabling practitioners to compare and align molecules. Potential applications beyond biochemistry might include the analysis of social networks or protein-protein interaction networks. While GraphOTC does not present any direct opportunities for negative societal impacts, indirect negative effects are possible in each of the potential applications mentioned.

This study has several limitations. While GraphOTC performs well in the experiments presented here, other methods may be better suited to particular tasks. Additionally, further experiments are necessary to characterize the performance of GraphOTC fully. Lastly, we note that all of the currently available graph OT methods, including GraphOTC, may present computational challenges for large graphs.

4.8 Proofs

In this section, we prove Theorems 4.1 and 4.2.

4.8.1 Proof of Theorem 4.1

We begin by proving Theorem 4.1, concluding that the GraphOTC cost is a metric on a certain set of equivalence classes of graphs. We will first prove the following lemma, which states that two graphs are equivalent if and only if they have identical associated transition matrices.

Lemma 4.3. *For two graphs $\mathcal{G}_1 = (V, E, w_1)$ and $\mathcal{G}_2 = (V, E, w_2)$ with associated Markov transition matrices P and Q, $\mathcal{G}_1 \sim \mathcal{G}_2$ if and only if P and Q are equal.*

Proof. Suppose first that $\mathcal{G}_1 \sim \mathcal{G}_2$ and thus there exists C such that $w_1(u, v) = w_2(u, v) + C$ for every $(u, v) \in E$. Then clearly the two transition matrices P and Q associated with $\mathcal{G}_1$ and $\mathcal{G}_2$ are equal. Now suppose that P and Q are equal. Then for every $(u, v) \in E$, we have

$$\frac{\exp\{w_1(u,v)\}}{\sum_{\tilde{v}:(u,\tilde{v})\in E} \exp\{w_1(u,\tilde{v})\}} = \frac{\exp\{w_2(u,v)\}}{\sum_{\tilde{v}:(u,\tilde{v})\in E} \exp\{w_2(u,\tilde{v})\}}.$$

Written another way, we have $w_1(u, v) = w_2(u, v) + C_u$, where

$$C_u = \ln\left(\frac{\sum_{\tilde{v}:(u,\tilde{v})\in E} \exp\{w_1(u,\tilde{v})\}}{\sum_{\tilde{v}:(u,\tilde{v})\in E} \exp\{w_2(u,\tilde{v})\}}\right).$$

Since $\mathcal{G}_1$ and $\mathcal{G}_2$ are undirected, $(v, u) \in E$ and a similar argument will establish that $w_1(v, u) = w_2(v, u) + C_v$. The undirectedness of $\mathcal{G}_1$ and $\mathcal{G}_2$ also implies that $w_1(u, v) = w_1(v, u)$ and $w_2(u, v) = w_2(v, u)$. It follows that $w_1(v, u) = w_2(v, u) + C_u$ and thus $C_u = C_v$. Since $\mathcal{G}_1$ and $\mathcal{G}_2$ are connected, there exists a sequence of edges $(u_1, u_2), (u_2, u_3), (u_{n-1}, u_n) \in E$ such that $u \in \{u_1, ..., u_n\}$ for every $u \in V$. Iterating the arguments above for all edges in this sequence we conclude that $C_u = C_v$ for every $u, v \in V$. It follows that $w_1(u, v) = w_2(u, v) + C$ for some constant C that is independent of u and v. Then by definition, $\mathcal{G}_1 \sim \mathcal{G}_2$ and the claim is proven. □

Before proceeding with the proof of Theorem 4.1, we introduce some necessary background. We will review the optimal joining distance $\mathcal{S}$, introduced in Chapter 1. Informally, for stationary processes X and Y taking values in finite sets $\mathcal{X}$ and $\mathcal{Y}$, the optimal joining distance $\mathcal{S}(c; X, Y)$ with respect to $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is obtained by minimizing $\mathbb{E}[c(\tilde{X}, \tilde{Y})]$ over the set of jointly stationary paired processes $(\tilde{X}, \tilde{Y})$ taking values in $\mathcal{X} \times \mathcal{Y}$ such that $\tilde{X}$ and $\tilde{Y}$ are equal in distribution to X and Y, respectively. Such processes are referred to as *joinings* of X and Y (Furstenberg, 1967; de la Rue, 2020). Since every transition coupling of stationary Markov chains P and Q is also a joining of P and Q, we have

$$\mathcal{S}(c; P, Q) \leq \min \left\{ \langle \lambda, c \rangle : R \in \Pi_{\mathrm{TC}}(P, Q), \lambda R = \lambda, \lambda \in \Delta_{|\mathcal{X}| \times |\mathcal{Y}|} \right\}.$$

In particular, if the chains P and Q are associated with graphs $\mathcal{G}_1 = (\mathcal{X}, E_1, w_1)$ and $\mathcal{G}_2 = (\mathcal{Y}, E_2, w_2)$, $\mathcal{S}(c; P, Q) \leq d(\mathcal{G}_1, \mathcal{G}_2)$. Finally, we remark that in the case that $\mathcal{X} = \mathcal{Y}$, $\mathcal{S}(c; \cdot, \cdot)$ is known to satisfy the properties of a metric when c does (Gray et al., 1975).

Theorem 4.1. *Suppose that the cost function $c : V \times V \to \mathbb{R}_+$ satisfies the properties of a metric on V. Then d is a metric on the equivalence classes defined by $\sim$.*

Proof. The symmetry of d is clear. Moreover, it was established in Chapter 3 that the optimal transition coupling cost satisfies the triangle inequality for Markov chains when the cost c does. Thus d satisfies the triangle inequality for graphs. So it suffices to show that $d(\mathcal{G}_1, \mathcal{G}_2) = 0$ if and only if $\mathcal{G}_1 \sim \mathcal{G}_2$. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be graphs satisfying $\mathcal{G}_1 \sim \mathcal{G}_2$ with associated transition matrices P and Q. By Lemma 4.3, P and Q are equal and clearly $d(\mathcal{G}_1, \mathcal{G}_2) = 0$ since $\langle \lambda, c \rangle = 0$ is achieved by λ satisfying $\lambda(u, v) = p(u)\delta(u = v)$, which is stationary for the transition coupling satisfying $R((u, v), (u', v')) = P(u, u')\delta(u = v)\delta(u' = v')$. Now suppose that $\mathcal{G}_1 \nsim \mathcal{G}_2$. Again by Lemma 4.3, the transition matrices P and Q are distinct and consequently so are their associated stationary Markov chains. Since it defines a distance on stationary processes, the optimal joining distance satisfies $\mathcal{S}(c; P, Q) > 0$. Since the optimal transition coupling cost is lower bounded by the optimal joining distance $\mathcal{S}(c; \cdot, \cdot)$, it follows that

$$0 < \mathcal{S}(c; P, Q) \leq \min\left\{\langle \lambda, c\rangle : R \in \Pi_{\text{TC}}(P, Q), \lambda R = \lambda, \lambda \in \Delta_{|V|\times|V|}\right\} = d(\mathcal{G}_1, \mathcal{G}_2).$$

Thus $d(\mathcal{G}_1, \mathcal{G}_2) = 0$ if and only if $\mathcal{G}_1 \sim \mathcal{G}_2$ and the proof is concluded. □

4.8.2 Proof of Theorem 4.2

Next we prove Theorem 4.2. Our proof follows an argument similar to the proof of Theorem 3.13 in Chapter 3, which establishes a stability result for the OTC problem. Our proof utilizes a well-known stability result, detailed in Theorem A below, for optimization problems satisfying certain conditions,. Before stating this result, we fix some notation and definitions. Let $\mathcal{Z}$ and $\mathcal{U}$ be Polish spaces corresponding to a set of possible solutions and a set of parameters, respectively. Then for a function $f : \mathcal{Z} \times \mathcal{U} \to \mathbb{R}$ and feasible set $\Phi : \mathcal{U} \to 2^{\mathcal{Z}}$, consider the optimization problem indexed by a parameter $u \in \mathcal{U}$,

$$\min\{f(z, u) : z \in \Phi(u)\}. \tag{4.5}$$

Note that $f(\cdot, u) : \mathcal{Z} \to \mathbb{R}$ describes the objective to be minimized and $\Phi(u) \subset \mathcal{Z}$ represents the feasible set of Problem (4.5). We will call a set $\mathcal{V} \subset \mathcal{Z}$ a neighborhood of a subset $\mathcal{W} \subset \mathcal{Z}$ if $\mathcal{W} \subset \operatorname{int} \mathcal{V}$. Neighborhoods in $\mathcal{U}$ will be defined similarly. Now we may state the key result for the proof of Theorem 4.2.

Theorem A ((Bonnans and Shapiro, 2013, Proposition 4.4))**.** *Let u_0 be a given point in the parameter space $\mathcal{U}$. Suppose that (i) the function $f(z, u)$ is continuous on $\mathcal{Z} \times \mathcal{U}$, (ii) the graph of the multifunction $\Phi(\cdot)$ is a closed subset of $\mathcal{U} \times \mathcal{Z}$, (iii) there exists $\alpha \in \mathbb{R}$ and a compact set $C \subset \mathcal{Z}$ such that for every u in a neighborhood of u_0, the level set $\{z \in \Phi(u) : f(z, u) \leq \alpha\}$ is nonempty and contained in C, (iv) for any neighborhood $\mathcal{V}_{\mathcal{Z}}$ of the set $\operatorname{argmin}_{z\in\Phi(u_0)} f(z, u_0)$ there exists a neighborhood $\mathcal{V}_U$ of u_0 such that $\mathcal{V}_{\mathcal{Z}} \cap \Phi(u) \neq \emptyset$ for all $u \in \mathcal{V}_{\mathcal{U}}$. Then the optimal value function $u \mapsto \min_{z\in\Phi(u)} f(z, u)$ is continuous at $u = u_0$ and the multifunction $u \mapsto \operatorname{argmin}_{z\in\Phi(u)} f(z, u)$ is upper semicontinuous at u_0.*

In order to simplify notation going forward, we will assume without loss of generality that the vertex sets V_1 and V_2 satisfy $|V_1| = |V_2| = d$ for some $d \in \mathbb{N}$. Moreover, we will make use of the fact that the set of $d \times d$ (resp. $d^2 \times d^2$) transition matrices may be identified with the set Δ_d^d (resp. $\Delta_{d^2}^{d^2}$). Now, let

$$\mathcal{Z} = \left\{(\lambda, R) \in \Delta_{d^2} \times \Delta_{d^2}^{d^2} : R \in \Pi_{\text{TC}}(P, Q) \text{ for some } P, Q \in \Delta_d^d, \lambda R = \lambda\right\}$$

be the union of all feasible sets for the FusedOTC problem and $\mathcal{U} = \text{cl}\{(P^\tau, Q^\tau) : \tau \in \mathbb{N}\} \times [0,1]$ be the set of all valid pairs (P^τ, Q^τ, α) including the limit $(\overline{P}, \overline{Q}, \alpha)$ where $\overline{P} = \lim_{\tau\to\infty} P^\tau$ and $\overline{Q} = \lim_{\tau\to\infty} Q^\tau$. Note that since P and Q are aperiodic and irreducible, $\overline{P}$ and $\overline{Q}$ exist and have rows equal to the stationary distributions p and $q \in \Delta_d$ of P and Q, respectively. One may easily verify that $\mathcal{Z}$ and $\mathcal{U}$ are compact subsets of $\mathbb{R}^{d^2} \times \mathbb{R}^{d^2 \times d^2}$ and $\mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \times \mathbb{R}$, respectively. The objective function $f(\cdot,\cdot)$ is identified with the map

$$((\lambda, R), (P^\tau, Q^\tau, \alpha)) \mapsto \alpha\langle c, \lambda\rangle + (1-\alpha)\langle E, \lambda \odot R\rangle,$$

which does not depend on P^τ or Q^τ. We will refer to the constraint functions for the FusedOTC problem by $\Phi : \mathcal{U} \to 2^{\mathcal{Z}}$. In particular, we let

$$\Phi((P^\tau, Q^\tau, \alpha)) = \left\{(\lambda, R) \in \Delta_{d^2} \times \Delta_{d^2}^{d^2} : R \in \Pi_{\text{TC}}(P^\tau, Q^\tau), \lambda R = \lambda, \lambda \in \Delta_{d^2}\right\}.$$

Now we may proceed with the proof of Theorem 4.2.

Theorem 4.2. *The FusedOTC cost $d_{\tau,\alpha}$ satisfies the following:*

- $\lim_{\alpha\to 1} d_{1,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = d(\mathcal{G}_1, \mathcal{G}_2)$
- $\lim_{(\tau,\alpha)\to(\infty,1)} d_{\tau,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = d^W(\mathcal{G}_1, \mathcal{G}_2)$

Proof. We begin by proving the second claim. It will suffice to check the conditions of Theorem A for the FusedOTC problem at the point $u_0 = (\overline{P}, \overline{Q}, 1) \in \mathcal{U}$. First, (i) clearly holds since the objective $f(\cdot,\cdot)$ is quadratic in (λ, R), linear in α, and does not depend on P^τ or Q^τ. Next, we will show that the graph of $\Phi(\cdot)$ is a closed subset of $\mathcal{U} \times \mathcal{Z}$. Fix a sequence $\{(P_n, Q_n, \alpha_n, \lambda_n, R_n)\}_{n\geq 1} \subset \text{graph}\,\Phi(\cdot)$. As a subset of the compact set $\Delta_d^d \times \Delta_d^d \times [0,1] \times \Delta_{d^2} \times \Delta_{d^2}^{d^2}$, it has a subsequence, which we also label as $\{(P_n, Q_n, \alpha_n, \lambda_n, R_n)\}_{n\geq 1}$ converging to some $(P', Q', \alpha', \lambda', R') \in \Delta_d^d \times \Delta_d^d \times [0,1] \times \Delta_{d^2} \times \Delta_{d^2}^{d^2}$. Taking limits of the linear equations $R_n \in \Pi_{\mathrm{TC}}(P_n, Q_n)$ and $\lambda_n R_n = \lambda_n$, we conclude that $R' \in \Pi_{\mathrm{TC}}(P', Q')$ and $\lambda' R' = \lambda'$. Thus $(P', Q', \alpha', \lambda', R') \in \operatorname{graph}\Phi(\cdot)$ and (ii) holds. To show that (iii) is satisfied, note that one may let $a = \|c\|_\infty + \|E\|_\infty$ and use the fact that the entire set $\mathcal{Z}$ is compact. Finally, we will show that (iv) is satisfied. Let $\mathcal{V}_{\mathcal{Z}} \subset \mathcal{Z}$ be a neighborhood of $\operatorname{argmin}_{z\in\Phi(u_0)} f(z, u_0)$. Then define the neighborhood $\mathcal{V}_{\mathcal{U}}$ of $u_0 = (\overline{P}, \overline{Q}, 1)$ as

$$\mathcal{V}_{\mathcal{U}} := \{(P, Q) \in \Delta_d^d \times \Delta_d^d : R \in \Pi_{\mathrm{TC}}(P, Q) \text{ for some } (\lambda, R) \in \mathcal{V}_{\mathcal{Z}}\} \times \{1\}.$$

Note that $\mathcal{V}_{\mathcal{U}}$ is nonempty by the non-emptiness of $\mathcal{V}_{\mathcal{Z}}$ and the definition of $\mathcal{Z}$. Moreover, $\mathcal{V}_{\mathcal{Z}} \cap \Phi(u) \neq \emptyset$ for all $u \in \mathcal{V}_{\mathcal{U}}$ by construction. Thus all the conditions of Theorem A are satisfied and we conclude that $u \mapsto \min_{z\in\Phi(u)} f(z, u)$ is continuous at $u_0 = (\overline{P}, \overline{Q}, 1)$. Letting $u(\tau, \alpha) = (P^\tau, Q^\tau, \alpha)$, the continuity of $u(\cdot, \cdot)$ implies that the map $(\tau, \alpha) \mapsto \min_{z\in\Phi(u(\tau,\alpha))} f(z, u(\tau, \alpha))$ satisfies

$$\lim_{(\tau,\alpha)\to(\infty,1)} d_{\tau,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = \lim_{(\tau,\alpha)\to(\infty,1)} \min_{z\in\Phi(u(\tau,\alpha))} f(z, u(\tau, \alpha)) = \min_{z\in\Phi(\overline{P},\overline{Q},1)} f(z, (\overline{P}, \overline{Q}, 1)).$$

In particular,

$$\lim_{(\tau,\alpha)\to(\infty,1)} d_{\tau,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = \min\left\{\langle \lambda, c\rangle : R \in \Pi_{\mathrm{TC}}(\overline{P}, \overline{Q}), \lambda R = \lambda, \lambda \in \Delta_{|V_1|\times|V_2|}\right\}.$$

Since the chains associated with $\overline{P}$ and $\overline{Q}$ are IID, the OTC cost between them is simply the standard OT cost between their respective stationary distributions p and q. As a consequence, we find that

$$\lim_{(\tau,\alpha)\to(\infty,1)} d_{\tau,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = d^{\mathrm{W}}(\mathcal{G}_1, \mathcal{G}_2).$$

Using the same line of reasoning as above, one may establish that the map $u \mapsto \min_{z\in\Phi(u)} f(z, u)$ is continuous at $u_0 = (P, Q, \alpha)$ for any $\alpha \in [0, 1]$. Letting $u(\alpha) = (P, Q, \alpha)$, the continuity of the $u(\cdot)$ implies that $\alpha \mapsto \min_{z\in\Phi(u(\alpha))} f(z, u(\alpha))$ is continuous at 1 and thus

$$\lim_{\alpha\to 1} d_{1,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = \lim_{\alpha\to 1} \min_{z\in\Phi(u(\alpha))} f(z, u(\alpha)) = \min_{z\in\Phi(u(1))} f(z, u(1)) = d(\mathcal{G}_1, \mathcal{G}_2).$$

This concludes the proof. □

CHAPTER 5
Consistent Estimation of Optimal Joinings

In this chapter, we consider the estimation of an optimal joining of two stationary processes from dependent observations. In the interest of computational tractability, we also introduce the entropic optimal joining problem, a solution of which can be estimated with improved computational efficiency. We show that these estimates are consistent in the large sample limit when the stationary processes of interest are ergodic. Finally, when the processes of interest satisfy a certain mixing condition, we prove a bound on the expected error of the estimated optimal joining cost.

5.1 Introduction

In this chapter, 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). Recall from Chapters 1 and 2 that 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 1.1 and 1.2 in Chapter 1). To address these, and other, issues arising in the general setting, we restrict our attention to *stationary* couplings of stationary processes, referred to as *joinings*. This seemingly mild restriction has far reaching consequences.

The primary focus of this chapter 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.

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. 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 the consistency of the resulting estimates. 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.

The rest of the chapter is organized as follows. Background on optimal transport, optimal joinings, some initial results, and related work are presented in the next section. In Section 5.3, 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 5.4. In Section 5.5, we introduce the entropic optimal joining problem and discuss how our estimation scheme and consistency result extend to this problem. We close with a discussion of our results in Section 5.6. Proofs of the main results are presented in Section 5.7.

5.2 Preliminaries and First Results

In this section, we review some background on optimal transport and notation. Let $\mathcal{U}$ and $\mathcal{V}$ be metric spaces let $\mu \in \mathcal{M}(\mathcal{U})$ and $\nu \in \mathcal{M}(\mathcal{V})$ be probability measures. Recall that the optimal transport problem for μ and ν with respect to a non-negative cost function $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$ is defined by

$$\mathcal{T}(c; \mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int c(u, v)\, d\pi(u, v),$$

where $\Pi(\mu, \nu)$ is the set of couplings of μ and ν. As noted previously, the optimal transport problem is very general, and makes no assumptions 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 (Tolstikhin et al., 2018; Arjovsky et al., 2017), graphs (Xu et al., 2019a,b; Titouan et al., 2019), trees (Wang et al., 2020), finite sets (Sommerfeld and Munk, 2018; Montrucchio and Pistone, 2021), and sequence spaces (Kolesnikov and Zaev, 2017).

5.2.1 Couplings of Stationary 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. Each $\mathbf{x} = (x_1, x_2, \ldots) \in \mathcal{U}$ is an infinite sequence with entries in $\mathcal{X}$, and each $\mathbf{y} = (y_1, y_2, \ldots) \in \mathcal{V}$ is an infinite sequence with entries in $\mathcal{Y}$. Let $\sigma : \mathcal{X}^{\mathbb{N}} \to \mathcal{X}^{\mathbb{N}}$ be the left-shift map on $\mathcal{X}^{\mathbb{N}}$ defined by $\sigma(x_1, x_2, ...) = x_2, x_3, \ldots$, and 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 $\mathcal{Y}^{\mathbb{N}}$.

Definition 5.1. *Let γ be a measure on the sequence space $\mathcal{U}^{\mathbb{N}}$ where $\mathcal{U}$ is finite. For $k \geq 1$, let γ_k be the distribution of the first k coordinates of $\mathbf{u} = (u_1, u_2, \ldots)$ under γ, that is, $\gamma_k(B) = \gamma(B \times \mathcal{U} \times \mathcal{U} \times \cdots)$ for each $B \subseteq \mathcal{U}^k$.*

It is helpful to recall the simple equivalence between stationary measures and stationary processes. Each 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 $\mathbb{P}(X_1^k \in B) = \mu(B \times \mathbb{R} \times \mathbb{R} \times \cdots)$ for all $B \subseteq \mathcal{X}^k$ and all $k \geq 1$. If μ is ergodic, so is the process X. In the same way, each measure $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ corresponds to a stationary process $Y = Y_1, Y_2, \ldots$ with values in $\mathcal{Y}$. If $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\nu \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ give rise to processes X and Y, respectively, then 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* require the joint process $(\tilde{X}, \tilde{Y})$ to be stationary.

In what follows we will consider a single letter cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ 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, we can readily extend our presentation to cost functions depending on any 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 that any continuous cost function $c_0 : \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to [0, \infty)$ defined on infinite sequences can be uniformly well approximated by a finite letter cost function.

Any single letter cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ can be extended to a cost function $c_0 : \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to [0, \infty)$ 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

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c_0 \, d\pi \;=\; \inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi_1 \;=\; \inf_{(\tilde{X},\tilde{Y}) \in \Pi(X,Y)} \mathbb{E} c(\tilde{X}_1, \tilde{Y}_1) \tag{5.1}$$

where X and Y are the stationary processes associated with μ and ν respectively.

5.2.2 Joinings of Stationary Processes

Recall Example 1.2 from Chapter 1, which showed that the optimal transport cost between an iid process and a deterministic process can be zero. This example illustrates an important feature of the optimal transport problem for stationary processes: for a single letter cost function c, an optimal coupling $(\tilde{X}, \tilde{Y})$ of processes X and Y need only align the first component of $\tilde{X}$ and $\tilde{Y}$; the joint behavior of $\tilde{X}$ and $\tilde{Y}$ at subsequent time points is not important. Analogous remarks apply to finite letter costs. If X and Y represent discrete-time audio and video sequences, an optimal coupling will sync these sequences at the initial time point, but will not be sensitive to differences at subsequent time points. While the finite memory cost functions play a role in this behavior, the examples above suggest that non-stationary couplings are also a problem. From a theoretical and practical point of view, it is natural to restrict attention to couplings that share the broad stochastic structure of the processes X and Y being coupled. This motivates the consideration of stationary couplings, also known as joinings.

Definition 5.2. *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. Joinings were first introduced by Furstenberg (Furstenberg, 1967) and have been studied extensively in the ergodic theory literature since that time; an overview and more details can be found in Chapter 2. Note that the independent coupling $\mu \otimes \nu$ is stationary and therefore $\mathcal{J}(\mu, \nu)$ is non-empty.

Let $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be a single letter cost function. Constraining the optimal transport problem (5.1) to the set of stationary couplings, we obtain the *optimal joining problem*,

$$\inf_{\lambda \in \mathcal{J}(\mu,\nu)} \int c \, d\lambda_1 \;=\; \inf_{(\tilde{X},\tilde{Y}) \in \mathcal{J}(X,Y)} \mathbb{E} c(\tilde{X}_1, \tilde{Y}_1) \tag{5.2}$$

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 (5.2) by $\mathcal{J}_{\min}(\mu, \nu)$, and the value of the infimum by $\mathcal{S}(c; \mu, \nu)$. Elements of $\mathcal{J}_{\min}(\mu, \nu)$ will be called optimal joinings, and $\mathcal{S}(c; \mu, \nu)$ will be called the optimal joining cost. The following proposition collects some standard properties of $\mathcal{J}(\mu, \nu)$ and $\mathcal{J}_{\min}(\mu, \nu)$; for more details see (McGoff and Nobel, 2021) and the references therein.

Proposition 5.3. *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.*

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 5.4. *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*

$$\mathcal{S}(c; \mu, \nu) = \lim_{k\to\infty} k^{-1} \mathcal{T}(c_k; \mu_k, \nu_k) = \mathcal{T}(\overline{c}; \mu, \nu).$$

The first equality in Proposition 5.4 was proven in (Gray et al., 1975) 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 5.7 using cyclical monotonicity of optimal couplings.

Proposition 5.4 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 average behavior, relative to the single letter cost, over the complete history of the joint process. This is a natural shift in objective when considering optimal transport for stationary processes. In comparison to the optimal transport problem $\mathcal{T}(\overline{c}; \mu, \nu)$, the optimal joining problem circumvents the need to work with the averaged cost $\overline{c}$, which may be highly irregular. On the other hand, the set of joinings $\mathcal{J}(\mu, \nu)$ has relatively simple structure (i.e. compactness, convexity) and leads to an optimization problem that is easier to study.

Remark 5.5. Proposition 5.4 implies that the optimal joining problem $\mathcal{S}(c; \mu, \nu)$ satisfies Kantorovich duality with respect to $\overline{c}$. In particular,

$$\mathcal{S}(c; \mu, \nu) = \sup_{(f,g)\in L^1(\mu)\times L^1(\nu)} \left\{ \int f \, d\mu + \int g \, d\nu : f \oplus g \leq \overline{c} \right\}. \tag{5.3}$$

The only difference between (5.3) and the standard Kantorovich dual problem is the use of $\overline{c}$ instead of c. More details on Kantorovich duality can be found in (Villani, 2008) and (Santambrogio, 2015). While we do not make use of the dual optimal joining problem in this chapter, we expect that it may be of use in future analyses.

5.2.3 Existing Work in Estimation for Stationary Optimal Transport

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 that are consistent for some class of ergodic processes. In terms of our main result, this is analogous to seeking a single choice of sequence $\{k(n)\}$ such that the desired convergence holds uniformly over all pairs μ and ν from some set. The estimates we propose are an extension of those proposed in (Ornstein and Weiss, 1990) for the $\overline{d}$-distance. In that previous work, it was shown that the scheme with $k(n) = o(\log n)$ is consistent whenever μ and ν are B-processes. Later work studied the limits of this estimation scheme (Marton and Shields, 1994) and the properties of processes for which the scheme is consistent (Ornstein and Shields, 1994). 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 $\{k(n)\}$. Furthermore, our finite-sample results allow one to identify explicit sequences $\{k(n)\}$ and corresponding error bounds under additional mixing conditions on μ and ν. We also extend our results to the setting with entropic penalization.

5.3 Estimation of an Optimal Joining and Its Expected Cost

Before describing the proposed estimation scheme, we first develop some intuition for our approach. In words, Proposition 5.4 states that the optimal transport cost between k-dimensional distributions of μ and ν converges to the optimal joining cost. Intuitively, we expect that for large k, a "good" estimate of $\frac{1}{k}\mathcal{T}(c_k; \mu_k, \nu_k)$ will be a "good" estimate of the optimal joining cost $\mathcal{S}(c; \mu, \nu)$. Furthermore, a stationary process measure with expected cost equal to the estimated optimal joining cost should be "close" to the set of optimal joinings $\mathcal{J}_{\min}(\mu, \nu)$.

Building off of this intuition, we propose an estimation scheme that is comprised of three steps. First, we construct k-block empirical measures from the observations. The resulting probability measures act as empirical estimates of μ_k and ν_k. Second, we select an optimal transport plan between the empirical k-block measures with respect to c_k. The expected cost of this coupling will be an empirical proxy for $\mathcal{T}(c_k; \mu_k, \nu_k)$. Finally, we construct a stationary process measure from the coupling in the second step. This is done via the k-block process construction, described formally in Definition 5.6.

Definition 5.6 ((Block process construction))**.** *Let $\mathcal{U}$ be a finite space 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 $\gamma_k \in \mathcal{M}(\mathcal{U}^k)$ to the unique probability measure on $\mathcal{U}^{\mathbb{N}}$ obtained by independently concatenating γ_k to itself infinitely many times. Formally, for any ℓk-dimensional cylinder set $C = C_1 \times \cdots \times C_{\ell k} \times \mathcal{U} \times \cdots \subset \mathcal{U}^{\mathbb{N}}$,*

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

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 $U \subset \mathcal{U}^{\mathbb{N}}$,

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

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

Remark 5.7. For simplicity, we will use the same notation for the block process construction defined in Definition 5.6 regardless of the alphabet of the processes under consideration. As such, $\Lambda^k[\mu_k]$ and $\Lambda^k[\nu_k]$ are well-defined.

The block process construction is standard in ergodic theory (Ornstein and Weiss, 1990; Shields, 1996) and ensures that the resulting process is stationary. Moreover, the block process constructed from a coupling is necessarily a joining of the block processes constructed from the coupling's marginal measures. Thus, we may use it to construct an estimate of an optimal joining of μ and ν. The details of the estimation scheme are as follows.

Step 1: (Construct empirical block measures) For $k \in \{1, ..., n\}$, define the probability measure $\hat{\mu}_{k,n} := \hat{\mu}_k[X_1^n] \in \mathcal{M}(\mathcal{X}^k)$ by

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

for every $x_1^k \in \mathcal{X}^k$. 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 analogous manner. Note that $\hat{\mu}_{k,n}$ and $\hat{\nu}_{k,n}$ are probability measures on $\mathcal{X}^k$ and $\mathcal{Y}^k$, respectively, and may be thought of as estimates of the k-dimensional distributions of μ and ν, written as μ_k and ν_k.

Step 2: (Find optimal coupling) After constructing $\hat{\mu}_{k,n}$ and $\hat{\nu}_{k,n}$, we find an optimal coupling of the two with respect to the k-step total cost c_k. Formally, let $\hat{\pi}_{k,n} := \hat{\pi}_k[X_1^n, Y_1^n] \in \mathcal{M}(\mathcal{X}^k \times \mathcal{Y}^k)$ be any probability measure satisfying

$$\hat{\pi}_{k,n} \in \operatorname*{argmin}_{\pi \in \Pi(\hat{\mu}_{k,n}, \hat{\nu}_{k,n})} \int c_k \, d\pi_k.$$

Thus, $\hat{\pi}_{k,n}$ has expected k-step cost equal to $\mathcal{T}(c_k; \hat{\mu}_{k,n}, \hat{\nu}_{k,n})$. To simplify notation, we define $\hat{\rho}_k(X_1^n, Y_1^n) := \frac{1}{k}\, \mathcal{T}(c_k; \hat{\mu}_{k,n}, \hat{\nu}_{k,n})$.

Step 3: (Construct stationary process measure) Given $\hat{\pi}_{k,n}$, we let $\hat{\lambda}^k[X_1^n, Y_1^n] \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$ be the stationary process measure such that $\hat{\lambda}^k[X_1^n, Y_1^n] = \Lambda^k[\hat{\pi}_{k,n}]$.

We propose $\hat{\lambda}^k[X_1^n, Y_1^n]$ as the estimated optimal joining and $\hat{\rho}_k(X_1^n, Y_1^n)$ as the estimated optimal joining cost, respectively. To simplify notation, we will occasionally write $\hat{\lambda}^{k,n}$ for $\hat{\lambda}^k[X_1^n, Y_1^n]$ and $\hat{\rho}_{k,n}$ for $\hat{\rho}_k(X_1^n, Y_1^n)$. We establish in Appendix C.1 that $\hat{\lambda}^{k,n}$ is in fact a joining of empirical estimates of μ and ν with expected cost equal to $\hat{\rho}_{k,n}$. This ensures that, after establishing that $\hat{\rho}_{k,n}$ converges to the optimal joining cost of μ and ν, $\hat{\lambda}^{k,n}$ will converge to the set of optimal joinings.

Remark 5.8. We note that other constructions are possible in Step 3. For example, one may construct a finite-order, stationary Markov process from $\hat{\pi}_{k,n}$. However, the expected cost of a stationary Markov chain constructed from $\hat{\pi}_{k,n}$ will generally not be equal to $\hat{\rho}_{k,n}$ 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 $\hat{\rho}_{k,n}$, which we show converges to the optimal joining cost.

Remark 5.9. Note that the optimal transport distance between k-block measures was proposed as an extension of optimal transport to stationary time series in (Muskulus and Verduyn-Lunel, 2011). However, that work did not consider the relationship of this approach to the optimal joining problem or the consistency of the proposed distance.

5.3.1 Consistency

Having detailed the proposed estimators $\hat{\lambda}^{k,n}$ and $\hat{\rho}_{k,n}$, 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, $\hat{\rho}_{k,n}$ will be close to $\frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k)$. However, as Proposition 5.4 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 $\{\hat{\rho}_{k(n),n}\}_{n\geq 1}$ 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 5.10 that under the stated assumptions, such a sequence $\{k(n)\}$ necessarily exists.

We will say that a sequence of Borel probability measures $\{\gamma^n\}\subset\mathcal{M}(\mathcal{U})$ converges weakly to a set $\Gamma\subset\mathcal{M}(\mathcal{U})$, written as $\gamma^n\Rightarrow\Gamma$, if every subsequence of $\{\gamma^n\}$ 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 5.10. *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. Let $\mathbb{P}$ be any coupling of μ and ν. Then there exists a sequence $\{k(n)\}$ with $k(n)\to\infty$ such that with $\mathbb{P}$-probability one, $\hat{\rho}_{k,n}\to\mathcal{S}(c;\mu,\nu)$ and $\hat{\lambda}^{k,n}\Rightarrow\mathcal{J}_{min}(\mu,\nu)$ as $n\to\infty$.*

In the special case that $\mathcal{X}=\mathcal{Y}$, c is 0-1 cost, and μ is a stationary coding of an iid process, the first conclusion of Theorem 5.10 has been established for $k(n)=\lceil\log_{|\mathcal{X}|}n\rceil$ (Ornstein and Weiss, 1990; Ornstein and Shields, 1994). In relation to that result, Theorem 5.10 addresses more general pairs of measures (requiring only ergodicity) but does not provide specific information about the sequence $\{k(n)\}$.

Remark 5.11. The arguments underlying the proof of Theorem 5.10 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. We omit the proof of the result at this level of generality in order to keep the stated assumptions consistent throughout the paper.

5.3.2 Choice of $k(n)$

The conclusion of Theorem 5.10 begs the question of how the choice of the sequence $\{k(n)\}$ depends on the marginal processes μ and ν. In the proof of Theorem 5.10, we find that the sequence $\{k(n)\}$ is related to a notion of *c-admissibility.* Before defining c-admissibility and detailing its relationship to the choice of $\{k(n)\}$, we require the following definition.

Definition 5.12. *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.

Adapted cost functions capture the variability of the cost in each of its arguments. In particular, the adapted cost arises naturally when studying the Lipschitz properties of the optimal transport and optimal joining costs (see Lemmas 5.25 and 5.29). We will use $c_k^{\mathcal{X}} : \mathcal{X}^k \times \mathcal{X}^k \to \mathbb{R}_+$ and $c_k^{\mathcal{Y}} : \mathcal{Y}^k \times \mathcal{Y}^k \to \mathbb{R}_+$ to denote the $\mathcal{X}$- and $\mathcal{Y}$-adapted costs summed over k coordinates. Formally, for any $x_1^k, \tilde{x}_1^k \in \mathcal{X}^k$, we let $c_k^{\mathcal{X}}(x_1^k, \tilde{x}_1^k) = \sum_{\ell=1}^k c_k^{\mathcal{X}}(x_\ell, \tilde{x}_\ell)$ and define $c_k^{\mathcal{Y}}$ analogously. Note that $(\mathcal{X}^k, \frac{1}{k} c_k^{\mathcal{X}})$ and $(\mathcal{Y}^k, \frac{1}{k} c_k^{\mathcal{Y}})$ are well-defined pseudometric spaces for every $k \geq 1$.

Definition 5.13. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a cost function. We will say that a nondecreasing sequence $\{k(n)\}$ with $k(n) \to \infty$ is* c-admissible *for $\mu \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ if*

$$\mu\left(\left\{\mathbf{x} \in \mathcal{X}^{\mathbb{N}} : \lim_{n \to \infty} \frac{1}{k} \mathcal{T}\left(c_k^{\mathcal{X}}; \hat{\mu}_k[x_1^n], \mu_k\right) = 0\right\}\right) = 1.$$

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

The c-admissibility property quantifies the rate at which the finite dimensional distributions of a process may be coupled to the empirical distributions of the process. To the best of our knowledge, c-admissibility has not appeared previously in the literature, although it may be regarding as a weakening of the notions of admissibility and $\overline{d}$-admissibility discussed in (Shields, 1996). We note that certain sequences growing like $\mathcal{O}(\log n)$ are known to be admissible for aperiodic Markov chains (Marton and Shields, 1994).

In the proof of the Theorem 5.10, we find that the sequence $\{k(n)\}$ for which the $\hat{\lambda}^{k,n}$ and $\hat{\rho}_{k,n}$ are consistent is related to c-admissibility as follows.

Proposition 5.14. *Under the hypotheses of Theorem 5.10, if a sequence $\{k(n)\}$ is c-admissible for both μ and ν, then with $\mathbb{P}$-probability one, $\hat{\rho}_{k,n} \to \mathcal{S}(c; \mu, \nu)$ and $\hat{\lambda}^{k,n} \Rightarrow \mathcal{J}_{min}(\mu, \nu)$ as $n \to \infty$.*

In particular, we find that the c-admissibility of a sequence $\{k(n)\}$ for both μ and ν is sufficient for the consistency of $\hat{\lambda}^{k,n}$ and $\hat{\rho}_{k,n}$. For example, as discussed above, if μ and ν are aperiodic Markov chains, the conclusion of Theorem 5.10 will hold with $k(n) = C \log n$ for an appropriate choice of $C > 0$.

5.4 Finite-Sample Error Bound

In this section, we detail an upper bound on the expected error of the proposed estimate of the optimal joining cost. As we show in Section 5.7, this task is related to the problem of obtaining bounds on the optimal transport cost between a measure and an empirical measure constructed from a finite number of samples. A substantial body of work has considered this problem in the iid case from both asymptotic and finite-sample perspectives (Dudley, 1969; Boissard and Le Gouic, 2014; Fournier and Guillin, 2015; Weed and Bach, 2019; Mena and Niles-Weed, 2019; Genevay et al., 2019; Klatt et al., 2020). Other work has focused on rates of convergence and central limit theorems for the 1-Wasserstein distance on $\mathcal{M}(\mathbb{R})$ when samples are drawn from a stationary process satisfying a certain mixing condition (Dede, 2009; Boissard and Le Gouic, 2014; Dedecker and Merlevède, 2017; Berthet et al., 2020).

Building upon the intuition laid out in Section 5.3, we expect that the magnitude of error in the estimated optimal joining cost will depend on the rate at which the k-step optimal transport cost $\frac{1}{k}\,\mathcal{T}(c_k; \mu_k, \nu_k)$ converges to the optimal joining cost $\mathcal{S}(c; \mu, \nu)$ and the rate at which the empirical k-step optimal transport cost $\hat{\rho}_{k,n}$ converges to the true k-step optimal transport cost in expectation. Both of these quantities will depend on the extent to which the k-dimensional distributions of μ and ν capture the behavior of the full processes. This will be quantified via their respective ϕ-mixing coefficients.

Definition 5.15. *Let $\mathcal{U}$ be finite. We will say that $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ has a ϕ-mixing coefficient $\phi_\gamma : \mathbb{N}_0 \to \mathbb{R}_+$ if $\phi_\gamma(0) = 1$ and for any $g \geq 1$,*

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

*where $\gamma(\mathcal{U}^{g-1} \times B|A) = \gamma(A \times \mathcal{U}^{g-1} \times B)/\gamma(A)$. The process measure γ will be 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. For more details on ϕ-mixing and its relationship to other strong mixing conditions, we refer the reader to (Bradley, 2005).

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

$$\begin{aligned}\mathbb{E}\left[|\hat{\rho}_k(X_1^n, Y_1^n) - \mathcal{S}(c; \mu, \nu)|\right] &\leq \|c\|_\infty \left(\frac{k(\phi_\mu(g+1) + \phi_\nu(g+1))}{k+g} + \frac{3g}{k}\right) \\ &+ C\left(t + \left(\frac{1}{n^2}\sum_{\ell=0}^{n}(n-\ell+1)\phi_\mu^{1/2}(\ell)\right)^{1/2} \int_t^{\|c\|_\infty/4} \mathcal{N}(\mathcal{X}^k, \frac{1}{k}c_k^{\mathcal{X}}, \varepsilon)^{1/2}\, d\varepsilon\right. \\ &+ \left.\left(\frac{1}{n^2}\sum_{\ell=0}^{n}(n-\ell+1)\phi_\nu^{1/2}(\ell)\right)^{1/2} \int_t^{\|c\|_\infty/4} \mathcal{N}(\mathcal{Y}^k, \frac{1}{k}c_k^{\mathcal{Y}}, \varepsilon)^{1/2}\, d\varepsilon\right).\end{aligned}$$

Theorem 5.16 gives an abstract upper bound on the expected error of the proposed 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 $\frac{1}{k}c_k^{\mathcal{X}}$ and $\frac{1}{k}c_k^{\mathcal{Y}}$. The latter terms may be viewed as assessing the regularity of the adapted costs $c^{\mathcal{X}}$ and $c^{\mathcal{Y}}$ in some sense. In particular, when the cost c does not vary much, $c^{\mathcal{X}}$ and $c^{\mathcal{Y}}$ will yield smaller covering numbers $\mathcal{N}(\mathcal{X}^k, \frac{1}{k}c_k^{\mathcal{X}}, \varepsilon)$ and $\mathcal{N}(\mathcal{Y}^k, \frac{1}{k}c_k^{\mathcal{Y}}, \varepsilon)$. This coincides with the intuition that estimation should be easier when c is mostly constant.

Next, we use Theorem 5.16 to obtain an explicit upper bound.

Corollary 5.17. *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 $n \geq 1$, $k \in \{1,...,n\}$ and $g \geq 0$,

$$\mathbb{E}\left[|\hat{\rho}_k(X_1^n, Y_1^n) - \mathcal{S}(c;\mu,\nu)|\right] \leq \|c\|_\infty \left(\frac{k(\phi_\mu(g+1)+\phi_\nu(g+1))}{k+g} + \frac{3g}{k} + \frac{C(|\mathcal{X}|^{k/2}+|\mathcal{Y}|^{k/2})}{n^{1-p/2}}\right).$$

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

Corollary 5.17 gives explicit, finite-sample control on the mean error in the estimated entropic 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 (Csiszár and Talata, 2010; Talata, 2013, 2010; Gallo et al., 2013; Bressaud et al., 1999) 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 5.16 and Corollary 5.17 include the special case of Ornstein's $\overline{d}$-distance and thus provides some additional insight into the estimation scheme for this distance proposed in (Ornstein and Weiss, 1990). To provide further context for Corollary 5.17, we consider two examples below.

Example 5.18 ((IID processes)). If μ and ν are iid processes, then $\mathcal{S}(c;\mu,\nu) = \mathcal{T}(c;\mu_1,\nu_1)$ and $\phi_\mu(g) = \phi_\nu(g) = 0$ for every $g \geq 1$, and so we may let $k = 1$ and $g = 0$. Then by Corollary 5.17, we see that

$$\mathbb{E}\left[|\hat{\rho}_1(X_1^n, Y_1^n) - \mathcal{S}(c;\mu,\nu)|\right] = \mathcal{O}\left(n^{-1/2}\right).$$

This rate is consistent with known rates for the estimation of the 1-Wasserstein distance on finite spaces (Boissard and Le Gouic, 2014).

For iid processes, the optimal joining problem reduces to the optimal coupling problem of their 1-dimensional marginal measures, which yields the error bound above. However, when at least one of the measures is not iid, the optimal joining need not be Markov of any order (Ellis, 1976), 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.

Example 5.19 ((Markov processes)). If μ and ν are aperiodic and irreducible Markov chains, then there exist constants C_μ, $C_\nu < \infty$ and $\rho_\mu, \rho_\nu \in [0,1)$ such that $\phi_\mu(g) \leq C_\mu \rho_\mu^g$ and $\phi_\nu(g) \leq C_\nu \rho_\nu^g$ (Davydov, 1974; Bradley, 2005). Thus the summability conditions in Corollary 5.17 are satisfied with $p = 1$. Applying Corollary 5.17 and letting $k(n) = \left\lfloor \frac{\alpha \log(n)}{\log(|\mathcal{X}| \vee |\mathcal{Y}|) \vee 1} \right\rfloor$ and $g(n) = \left\lfloor -\frac{\log(\alpha \log(n))}{\log(\rho_\mu \wedge \rho_\nu)} \right\rfloor$ for some $\alpha \in (0,1)$ and n large enough, we find

$$\mathbb{E}\left[|\hat{\rho}_k(X_1^n, Y_1^n) - \mathcal{S}(c; \mu, \nu)|\right] = \mathcal{O}\left(\frac{\log(\log(n))}{\log(n)}\right).$$

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. Recall from Chapter 2 that 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:

$$\mathcal{T}_\eta(c; \mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int c \, d\pi - \eta H(\pi) \right\}.$$

In this section, we extend entropic regularization techniques from optimal transport to the optimal joining problem. In particular, we aim to regularize the optimal joining problem in a manner that leads to efficient computation. We propose to find the "natural" penalty term for the optimal joining problem by viewing the regularized problem as a limit of entropic optimal transport problems. By solving a series of entropic optimal transport problems with increasing dimension, we observe that they converge to a regularized optimal joining cost with the *entropy rate* as the penalty term. Entropy rate is the process analogue of entropy and has been an object of interest in stochastic processes and information theory for many years, dating back to Shannon (Shannon, 1948). For $k \geq 1$ and $\pi \in \mathcal{M}(\mathcal{X}^k \times \mathcal{Y}^k)$, let $H_k(\pi) := -\sum_{x_1^k, y_1^k} \pi(x_1^k, y_1^k) \log \pi(x_1^k, y_1^k)$ where we use $H_k(\cdot)$ instead of $H(\cdot)$ to emphasize the dependence of this quantity on the dimension k. For a process $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$, we will abuse notation and let $H_k(\gamma) = H_k(\gamma_k)$.

Definition 5.20. *Let $\mathcal{U}$ be finite and $\gamma \in \mathcal{M}_s(\mathcal{U}^{\mathbb{N}})$ be a stationary process. Then the* entropy rate *of γ is defined by the limit $h(\gamma) := \lim_{k\to\infty} \frac{1}{k} H_k(\gamma)$, which is known to exist by subadditivity.*

In other words, the entropy rate is the limiting joint entropy per symbol of the finite dimensional distributions of the process. As an example, an iid process with one dimensional distribution p has entropy rate equal to $H(p)$. Moreover, a stationary, aperiodic and irreducible Markov chain with stationary distribution p and transition matrix P has entropy rate given by $-\sum_{ij} p_i P_{ij} \log P_{ij}$. We that, again by subadditivity, the limit in the definition of the entropy rate is in fact an infimum: $h(\gamma) = \inf_{k\geq 1} \frac{1}{k} H_k(\gamma)$.

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

$$\inf_{\lambda \in \mathcal{J}(\mu,\nu)} \left\{ \int c \, d\lambda_1 - \eta h(\lambda) \right\}. \tag{5.4}$$

As in the unregularized problem, one can show under the stated assumptions that the infimum in (5.4) is attained. We include a proof of this fact in Appendix C.2. From now on, we will denote the set of joinings achieving the infimum in (5.4) by $\mathcal{J}^{\eta}_{\min}(\mu,\nu)$ and the value of the infimum by $\mathcal{S}_\eta(c;\mu,\nu)$. Note that (5.4) 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.

At this point, we may formalize the motivation for entropy rate as the regularizer in the optimal joining problem. Recall from Section 5.3 that Proposition 5.4 describes how the optimal joining cost may be obtained as a limit of k-step optimal transport costs. In Proposition 5.21, we extend Proposition 5.4 to the entropic optimal joining problem.

Proposition 5.21. *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 $\eta \geq 0$,*

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

This result demonstrates how the entropic optimal joining cost can be obtained as a limit of entropic optimal transport costs. In this way, the entropy rate is a natural regularizer for the optimal joining problem. Furthermore, it suggests that the proposed approach to estimating an optimal joining and the optimal joining cost may extend to the regularized setting. Intuitively, a good estimate of $\frac{1}{k}\mathcal{T}_\eta(c_k;\mu_k,\nu_k)$ for large k should be a good estimate of $\mathcal{S}_\eta(c;\mu,\nu)$. From this perspective, one may also see how the choice of entropy rate as the regularization yields improved computational efficiency. In particular, existing algorithms for computing $\frac{1}{k}\mathcal{T}_\eta(c_k;\mu_k,\nu_k)$ efficiently will translate to faster estimates of $\mathcal{S}_\eta(c;\mu,\nu)$.

Before moving on to extend the proposed estimation scheme to this problem, we consider the stability of the entropic optimal joining cost in η. We wish to assert that for small η, $\mathcal{S}_\eta(c;\mu,\nu)$ will be a reasonable approximation of $\mathcal{S}(c;\mu,\nu)$ in some sense. Considering the limit as the regularization coefficient η converges to zero, we find in the next proposition that the entropic optimal joining cost converges to the unregularized optimal joining cost.

Proposition 5.22. *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*

$$\lim_{\eta\to 0}\mathcal{S}_\eta(c;\mu,\nu) = \mathcal{S}(c;\mu,\nu).$$

Proposition 5.22 is consistent with analogous results in computational optimal transport (Peyré and Cuturi, 2019).

5.5.1 Extension of the Estimation Procedure

The proposed estimation scheme (described in Section 5.3) 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': (Find entropic optimal coupling) Let

$$\hat{\pi}^\eta_{k,n} \in \operatorname*{argmin}_{\pi\in\Pi(\hat{\mu}_{k,n},\hat{\nu}_{k,n})}\left\{\int c_k\, d\pi - \eta H_k(\pi)\right\}.$$

Thus, $\hat{\pi}^\eta_{k,n}$ has expected entropic k-step cost equal to $\mathcal{T}_\eta(c_k;\hat{\mu}_{k,n},\hat{\nu}_{k,n})$. To simplify notation, we define $\hat{\rho}^\eta_k(X_1^n,Y_1^n) = \frac{1}{k}\mathcal{T}_\eta(c_k;\hat{\mu}_{k,n},\hat{\nu}_{k,n})$.

One may then construct a stationary process $\hat{\lambda}^{\eta,k}[X_1^n,Y_1^n] \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}}\times\mathcal{Y}^{\mathbb{N}})$ from $\hat{\pi}^\eta_{k,n}$ following Step 3 of the estimation scheme. We propose $\hat{\lambda}^{\eta,k}[X_1^n,Y_1^n]$ as an estimate of an entropic optimal joining of μ and ν and $\hat{\rho}_k^\eta(X_1^n, Y_1^n)$ as an estimate of the entropic optimal joining cost. In order to reduce notation, we will often write $\hat{\lambda}^{\eta,k,n}$ for $\hat{\lambda}^{\eta,k}[X_1^n, Y_1^n]$ and $\hat{\rho}_{k,n}^\eta$ for $\hat{\rho}_k^\eta(X_1^n, Y_1^n)$.

As in the unregularized case, we establish in Appendix C.1 that the estimate $\hat{\lambda}^{\eta,k,n}$ is a joining of empirical estimates of μ and ν with the desired expected entropic cost $\hat{\rho}_{k,n}^\eta$. For large n, one expects that $\hat{\rho}_{k,n}^\eta$ will be close to $\frac{1}{k}\,\mathcal{T}_\eta(c_k; \mu_k, \nu_k)$ and by Proposition 5.21, $\frac{1}{k}\,\mathcal{T}_\eta(c_k; \mu_k, \nu_k)$ will be close to $\mathcal{S}(c; \mu, \nu)$ when k is large.

5.5.2 Consistency

If one is interested in the behavior of the proposed estimates as n goes to infinity, similar reasoning as in Section 5.3 suggests that it is necessary to consider sequences $\{\hat{\lambda}^{\eta,k,n}\}_{n\geq 1}$ and $\{\hat{\rho}_{k,n}^\eta\}_{n\geq 1}$ for some sequence $\{k(n)\}$ with $k(n) \to \infty$. The following result extends Theorem 5.10 to the regularized setting.

Theorem 5.23. *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 $\{k(n)\}$ with $k(n) \to \infty$ such that with $\mathbb{P}$-probability one, $\hat{\rho}_{k,n}^\eta \to \mathcal{S}_\eta(c; \mu, \nu)$ and $\hat{\lambda}^{\eta,k,n} \Rightarrow \mathcal{J}_{min}^\eta(\mu, \nu)$ as $n \to \infty$.*

The proof of Theorem 5.23 is similar to that of Theorem 5.10; 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. Consequently, a conclusion analogous to Proposition 5.14 holds for the regularized estimation scheme presented above. We omit a formal statement of such a result to simplify presentation.

5.6 Discussion

The extension of optimal transport techniques to stochastic processes is an important problem in statistics and machine learning. In this chapter, 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 and consistency result 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 5.3 that the arguments in the proof of Theorem 5.10 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 5.16 do not extend easily to continuous spaces, and so further consideration is necessary.

5.7 Proofs

5.7.1 Lipschitz Continuity of Optimal Transport

Before proving our main results, we prove a lemma regarding the Lipschitz continuity of the (entropic) optimal transport cost. In particular, we show in Lemma 5.25 that the map $(\alpha, \beta) \mapsto \mathcal{T}_\eta(c; \alpha, \beta)$ is 1-Lipschitz with respect to a certain "adapted" optimal transport cost with an additional term to account for the entropic penalty. First we will require some basic results regarding the entropic optimal transport problem and the notion of a (c, η)-*transform*. Let $\mathcal{U}$ and $\mathcal{V}$ be finite spaces, $\alpha \in \mathcal{M}(\mathcal{U})$ and $\beta \in \mathcal{M}(\mathcal{V})$, $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$, and $\eta > 0$. It was established in (Cuturi and Peyré, 2018, Proposition 2.4) that the entropic optimal transport problem satisfies

$$\mathcal{T}_\eta(c; \alpha, \beta) = \max_{f:\mathcal{U}\to\mathbb{R}} \left\{ \int f \, d\alpha + \int f^{(c,\eta,\beta)} \, d\beta \right\} = \max_{g:\mathcal{V}\to\mathbb{R}} \left\{ \int g^{(c,\eta,\alpha)} \, d\alpha + \int g \, d\beta \right\}, \tag{5.5}$$

where

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

and

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

The formulation (5.5) is referred to as the semidual of the entropic optimal transport problem while the quantities $f^{(c,\eta,\beta)}$ and $g^{(c,\eta,\alpha)}$ 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

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

to simplify notation. Note that $g^{(c,\eta,\alpha)}(u) = \eta \log \alpha(u) + g^{(c,\eta)}(u)$. Our proof of Lemma 5.25 will leverage the duality (5.5) as well as the following basic facts about $f^{(c,\eta)}$ and $g^{(c,\eta)}$.

Proposition 5.24. *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 $|c(u,v) - c(u',v')| \leq L(d_{\mathcal{U}}(u,u') + d_{\mathcal{V}}(v,v'))$ for all $u, u' \in \mathcal{U}$ and $v, v' \in \mathcal{V}$ for some $L \in \mathbb{R}$. Then for any $\eta > 0$, $f^{(c,\eta)}$ and $g^{(c,\eta)}$ satisfy $|f^{(c,\eta)}(v) - f^{(c,\eta)}(v')| \leq L d_{\mathcal{V}}(v,v')$ and $|g^{(c,\eta)}(u) - g^{(c,\eta)}(u')| \leq L d_{\mathcal{U}}(u,u')$ for all $u, u' \in \mathcal{U}$ and $v, v' \in \mathcal{V}$.*

A proof of Proposition 5.24 is provided in Appendix C.3. A more detailed discussion of the (c, η)-transform and its use in optimal transport can be found in (Peyré and Cuturi, 2019). Now we may proceed to the result of interest.

Lemma 5.25. *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 $\eta \geq 0$ and any $\alpha, \alpha' \in \mathcal{M}(\mathcal{U})$ and $\beta, \beta' \in \mathcal{M}(\mathcal{V})$,*

$$\begin{aligned} \left| \mathcal{T}_\eta(c; \alpha, \beta) - \mathcal{T}_\eta(c; \alpha', \beta') \right| \leq \mathcal{T}(c^{\mathcal{U}}; \alpha, \alpha') + \eta |H(\alpha) - H(\alpha')| \\ + \mathcal{T}(c^{\mathcal{V}}; \beta, \beta') + \eta |H(\beta) - H(\beta')|. \end{aligned} \tag{5.6}$$

Proof. We begin by considering the case $\eta > 0$. By the triangle inequality,

$$\left| \mathcal{T}_\eta(c; \alpha, \beta) - \mathcal{T}_\eta(c; \alpha', \beta') \right| \leq \left| \mathcal{T}_\eta(c; \alpha, \beta) - \mathcal{T}_\eta(c; \alpha', \beta) \right| + \left| \mathcal{T}_\eta(c; \alpha', \beta) - \mathcal{T}_\eta(c; \alpha', \beta') \right|. \tag{5.7}$$

By (5.5), there exists $g : \mathcal{V} \to \mathbb{R}$ be such that

$$\mathcal{T}_\eta(c; \alpha, \beta) = \int g^{(c,\eta,\alpha)}\, d\alpha + \int g\, d\beta. \tag{5.8}$$

Rewriting $g^{(c,\eta,\alpha)}$, we have

$$\begin{aligned}\mathcal{T}_\eta(c; \alpha, \beta) &= \int \left(g^{(c,\eta)} + \eta \log \alpha\right)\, d\alpha + \int g\, d\beta \\ &= \int g^{(c,\eta)}\, d\alpha + \int g\, d\beta - \eta H(\alpha).\end{aligned}$$

Since g is also feasible for the semidual problem of $\mathcal{T}_\eta(c; \alpha', \beta)$, (5.5) implies that

$$\begin{aligned}\mathcal{T}_\eta(c; \alpha', \beta) &\geq \int g^{(c,\eta,\alpha')}\, d\alpha' + \int g\, d\beta \\ &= \int \left(g^{(c,\eta)} + \eta \log \alpha'\right)\, d\alpha' + \int g\, d\beta \\ &= \int g^{(c,\eta)}\, d\alpha' + \int g\, d\beta - \eta H(\alpha'). \qquad (5.9)\end{aligned}$$

Combining (5.8) and (5.9),

$$\mathcal{T}_\eta(c; \alpha, \beta) - \mathcal{T}_\eta(c; \alpha', \beta) \leq \int g^{(c,\eta)}\, d\alpha - \int g^{(c,\eta)}\, d\alpha' + \eta(H(\alpha') - H(\alpha)).$$

Note that c necessarily satisfies $|c(u, v) - c(\tilde{u}, \tilde{v})| \leq c^{\mathcal{U}}(u, \tilde{u}) + c^{\mathcal{V}}(v, \tilde{v})$. Thus by Proposition 5.24, $g^{(c,\eta)}$ satisfies $g^{(c,\eta)}(u) - g^{(c,\eta)}(\tilde{u}) \leq c^{\mathcal{U}}(u, \tilde{u})$. Thus the pair $(g^{(c,\eta)}, -g^{(c,\eta)})$ is feasible for the dual of $\mathcal{T}(c^{\mathcal{U}}; \alpha, \alpha')$ and it follows that

$$\mathcal{T}_\eta(c; \alpha, \beta) - \mathcal{T}_\eta(c; \alpha', \beta) \leq \mathcal{T}(c^{\mathcal{U}}; \alpha, \alpha') + \eta(H(\alpha') - H(\alpha)).$$

A similar argument can be used to upper bound $\mathcal{T}_\eta(c; \alpha', \beta) - \mathcal{T}_\eta(c; \alpha, \beta)$, yielding

$$|\mathcal{T}_\eta(c; \alpha, \beta) - \mathcal{T}_\eta(c; \alpha', \beta)| \leq \mathcal{T}(c^{\mathcal{U}}; \alpha, \alpha') + \eta|H(\alpha) - H(\alpha')|.$$

The same line of reasoning will show that

$$|\mathcal{T}_\eta(c;\alpha',\beta) - \mathcal{T}_\eta(c;\alpha',\beta')| \leq \mathcal{T}(c^{\mathcal{V}};\beta,\beta') + \eta|H(\beta) - H(\beta')|,$$

which combines with (5.7) to give the result. Taking the limit as $\eta \to 0$ of (5.6) (see (Peyré and Cuturi, 2019, Remark 4.3)), we obtain the result for $\eta = 0$. □

We conclude the subsection with a simple proposition that details the implication of Lemma 5.25 for the k-step entropic optimal transport cost.

Proposition 5.26. *For any $\eta \geq 0$, $n \geq 1$, and $k \in \{1, ..., n\}$,*

$$\begin{aligned}\left|\hat{\rho}_k^\eta(X_1^n, Y_1^n) - \frac{1}{k}\mathcal{T}_\eta(c_k;\mu_k,\nu_k)\right| &\leq \frac{1}{k}\mathcal{T}(c_k^{\mathcal{X}};\hat{\mu}_{k,n},\mu_k) + \eta\left|\frac{1}{k}H_k(\hat{\mu}_{k,n}) - \frac{1}{k}H_k(\mu_k)\right| \\ &\quad + \frac{1}{k}\mathcal{T}(c_k^{\mathcal{Y}};\hat{\nu}_{k,n},\nu_k) + \eta\left|\frac{1}{k}H_k(\hat{\nu}_{k,n}) - \frac{1}{k}H_k(\nu_k)\right|.\end{aligned}$$

Proof. The result follows from an application of Lemma 5.25 and the pointwise inequalities $(\frac{1}{k}c_k)^{\mathcal{X}^k} \leq \frac{1}{k}c_k^{\mathcal{X}}$ and $(\frac{1}{k}c_k)^{\mathcal{Y}^k} \leq \frac{1}{k}c_k^{\mathcal{Y}}$. □

5.7.2 Proofs from Section 5.2

Next, we prove the second equality in Proposition 5.4, which states that the optimal joining cost is equal to the optimal transport cost with respect to the averaged cost $\bar{c}$. We will do this by showing that solutions to the optimal joining problem are characterized by a local property of their support known as *cyclical monotonicity.*

Definition 5.27. *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 $(u^1, v^1), ..., (u^N, v^N) \in C$,*

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

with the convention that $v^{N+1} = v^1$. A probability measure on $\mathcal{U} \times \mathcal{V}$ is called c-cyclically monotone *if it is concentrated on a c-cyclically monotone set.*

Cyclical monotonicity has been studied in the optimal transport literature as a means of characterizing optimality of couplings. In particular, we make reference to the following result.

Theorem B ((Beiglböck, 2015))**.** *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 $\pi \in \Pi(\mu, \nu)$ with $\int c \, d\pi < \infty$ is a solution to $\mathcal{T}(c; \mu, \nu)$.*

Under stronger assumptions on c (i.e. lower semicontinuity and integrability conditions), one may also establish the reverse implication, namely that any optimal coupling is c-cyclically monotone (see (Villani, 2008)). In this way, cyclical monotonicity characterizes solutions to the optimal transport problem and provides another perspective from which to study the problem. In the following lemma, we show that an analogous result holds for the optimal joining problem.

Lemma 5.28. *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 $\mathcal{S}(c; \mu, \nu)$ if and only if it is $\overline{c}$-cyclically monotone.*

Proof. We begin by showing that an ergodic, $\overline{c}$-cyclically monotone joining is optimal in the optimal joining problem. By construction, the averaged cost $\overline{c}$ is invariant under the joint left-shift map $\sigma \times \tau$. Thus by the pointwise ergodic theorem and the integrability of c, $\int c \, d\lambda_1 = \int \overline{c} \, d\lambda$ for every $\lambda \in \mathcal{J}(\mu, \nu)$. It follows by taking infima that $\mathcal{S}(c; \mu, \nu) = \mathcal{S}(\overline{c}; \mu, \nu)$. Fixing an ergodic and $\overline{c}$-cyclically monotone joining $\lambda \in \mathcal{J}(\mu, \nu)$, Theorem B implies that λ is a solution to $\mathcal{T}(\overline{c}; \mu, \nu)$. Thus

$$\int c \, d\lambda_1 = \int \overline{c} \, d\lambda = \mathcal{T}(\overline{c}; \mu, \nu) \leq \mathcal{S}(\overline{c}; \mu, \nu) = \mathcal{S}(c; \mu, \nu)$$

and it follows that λ is necessarily a solution to $\mathcal{S}(c; \mu, \nu)$.

Now it suffices to prove that any ergodic optimal joining is $\overline{c}$-cyclically monotone. Let $\lambda \in \mathcal{J}_{\min}(\mu, \nu)$ be ergodic. We will construct a $\overline{c}$-cyclically monotone set $C \subset \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ such that $\lambda(C) = 1$. Note that as a consequence of the ergodicity of λ, μ, and ν, and the pointwise ergodic theorem, the following two conditions hold.

1. There exists a set $D \subset \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ such that $\lambda(D) = 1$ on which $\overline{c}$ is constant and equal to $\int c \, d\lambda_1$.

2. There exist $E \subset \mathcal{X}^{\mathbb{N}}$ 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 $\mu_{\mathbf{x}}^n := \frac{1}{n} \sum_{\ell=0}^{n-1} \delta_{\sigma^\ell \mathbf{x}}$ and $\nu_{\mathbf{y}}^n := \frac{1}{n} \sum_{\ell=0}^{n-1} \delta_{\tau^\ell \mathbf{y}}$ satisfy $\mu_{\mathbf{x}}^n \Rightarrow \mu$ and $\nu_{\mathbf{y}}^n \Rightarrow \nu$.

Letting $C := D \cap (E \times F)$, one may easily establish that $\lambda(C) = 1$. So we need only show that the set C is $\overline{c}$-cyclically monotone. Let $N \geq 1$, $(\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^N, \mathbf{y}^N) \in C$ and suppose for the sake of contradiction that

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

Define the sequence of probability measures $\{\lambda^n\}$ by

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

where we use the convention $\mathbf{y}^{N+1} = \mathbf{y}^1$. Note that $\lambda^n \in \Pi(\overline{\mu}^n, \overline{\nu}^n)$ for every $n \geq 1$, where $\overline{\mu}^n := \frac{1}{N} \sum_{\ell=1}^{N} \mu^n_{\mathbf{x}^\ell}$ 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 as defined in the second condition above. By the choice of C, one may establish that $\overline{\mu}^n \Rightarrow \mu$ and $\overline{\nu}^n \Rightarrow \nu$ as $n \to \infty$. Thus by Lemma C.3 there is a subsequence of $\{\lambda^n\}$ converging weakly to some $\tilde{\lambda} \in \mathcal{J}(\mu, \nu)$. To simplify notation, we refer to this subsequence again as $\{\lambda^n\}$. Using the fact that $\overline{c}$ is constant on C and the continuity and boundedness of c, we have

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

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

Proposition 5.4. *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*

$$\mathcal{S}(c;\mu,\nu) = \lim_{k\to\infty} k^{-1}\,\mathcal{T}(c_k;\mu_k,\nu_k) = \mathcal{T}(\overline{c};\mu,\nu).$$

Proof. Under the stated conditions, it is well-known (Shields, 1996) that there exists $\lambda \in \mathcal{J}_{\min}(\mu,\nu)$ that is ergodic. By Lemma 5.28, λ is $\overline{c}$-cyclically monotone. Finally, by Theorem B, λ is a solution to $\mathcal{T}(\overline{c};\mu,\nu)$ and it follows that $\mathcal{S}(c;\mu,\nu) = \int c\,d\lambda_1 = \mathcal{T}(\overline{c};\mu,\nu)$. □

5.7.3 Proofs from Section 5.3

In this section, we prove Theorem 5.10 regarding the consistency of the proposed estimates without entropic regularization.

Theorem 5.10. *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. Let $\mathbb{P}$ be any coupling of μ and ν. Then there exists a sequence $\{k(n)\}$ with $k(n) \to \infty$ such that with $\mathbb{P}$-probability one, $\hat{\rho}_{k,n} \to \mathcal{S}(c;\mu,\nu)$ and $\hat{\lambda}^{k,n} \Rightarrow \mathcal{J}_{min}(\mu,\nu)$ 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 (Marton and Shields, 1994), 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

$$\begin{aligned}|\hat{\rho}_k(X_1^n,Y_1^n) - \mathcal{S}(c;\mu,\nu)| \leq{}& \left|\hat{\rho}_k(X_1^n,Y_1^n) - \frac{1}{k}\,\mathcal{T}\left(c_k;\mu_k,\nu_k\right)\right| \\ &+ \left|\frac{1}{k}\,\mathcal{T}\left(c_k;\mu_k,\nu_k\right) - \mathcal{S}(c;\mu,\nu)\right|.\end{aligned} \tag{5.10}$$

Applying Proposition 5.26, we have

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

which by the c-admissibility of k for μ and ν implies that the first term on the right hand side in (5.10) goes to zero, $\mathbb{P}$-almost surely as $n \to \infty$. An application of Proposition 5.4 with the fact that $k(n) \to \infty$ shows that the second term on the right hand side in (5.10) goes to zero. It follows that

$$|\hat{\rho}_k(X_1^n, Y_1^n) - \mathcal{S}(c; \mu, \nu)| \to 0, \quad \mathbb{P}\text{-almost surely.} \tag{5.11}$$

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 an element $\omega \in \Omega$ of the sample space in the set of $\mathbb{P}$-measure one on which (5.11) holds. Let $\{\hat{\lambda}^{k,n}\}_{n \geq 1}$ be the corresponding sequence of estimated optimal joinings where the dependence on $X_1^n(\omega)$ and $Y_1^n(\omega)$ has been suppressed. By Lemma C.3, 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

$$\int c \, d\lambda_1 = \lim_{\ell \to \infty} \int c \, d\hat{\lambda}_1^{k,n_\ell} = \lim_{\ell \to \infty} \frac{1}{k} \mathcal{T}\left(c_k; \mu_k^{n_\ell}, \nu_k^{n_\ell}\right) = \mathcal{S}(c; \mu, \nu),$$

where the first equality follows from the continuity and boundedness of c, the second equality follows from Proposition C.1, and the third equality follows from (5.11). 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 ω, this convergence occurs $\mathbb{P}$-almost surely. □

5.7.4 Proofs from Section 5.4

In this section, we prove Theorem 5.16 regarding the expected error of the estimated optimal joining cost $\hat{\rho}_{k,n}$. Our argument may be broken down into three steps: First, we prove a Lipschitz result for the optimal joining cost akin to Lemma 5.25 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.

Lipschitz Bound. To begin, we establish a Lipschitz property for the optimal joining cost akin to the result stated in Lemma 5.25. In particular, we show that $(\mu, \nu) \mapsto \mathcal{S}(c; \mu, \nu)$ is $\|c\|_\infty$-Lipschitz with respect to the $\overline{d}$-distance with an additional term to account for the difference in entropy rates of the marginal processes. Since we will require the Lipschitz bound for the proof of Theorem 5.23 as well, we prove Lemma 5.25 more generally for the regularized optimal joining cost $\mathcal{S}_\eta(c;\mu,\nu)$. Briefly, we remind the reader that the $\overline{d}$-distance between two processes, introduced in (Ornstein, 1973), may be defined as the optimal joining cost with respect to the single-letter Hamming metric $(\mathbf{u},\mathbf{u}') \mapsto \delta(u_1 \neq u_1')$. The distance $\overline{d}$ may be thought of as the process analogue to the total variation distance.

Lemma 5.29. *Let $\alpha,\alpha' \in \mathcal{M}_s(\mathcal{X}^{\mathbb{N}})$ and $\beta,\beta' \in \mathcal{M}_s(\mathcal{Y}^{\mathbb{N}})$ be stationary process measures. Then for any $\eta \geq 0$,*

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

Proof. Fix $k \geq 1$ and recall that by Proposition 5.26,

$$\begin{aligned}\left|\frac{1}{k}\mathcal{T}_\eta\left(c_k;\alpha_k,\beta_k\right) - \frac{1}{k}\mathcal{T}_\eta\left(c_k;\alpha_k',\beta_k'\right)\right| \leq \frac{1}{k}\mathcal{T}(c_k^{\mathcal{X}};\alpha_k,\alpha_k') &+ \frac{1}{k}\mathcal{T}(c_k^{\mathcal{Y}};\beta_k,\beta_k') \\ &+ \eta\left|\frac{1}{k}H_k(\alpha_k) - \frac{1}{k}H_k(\alpha_k')\right| \\ &+ \eta\left|\frac{1}{k}H_k(\beta_k) - \frac{1}{k}H_k(\beta_k')\right|.\end{aligned}$$

Now note that $c_k^{\mathcal{X}}$ and $c_k^{\mathcal{Y}}$ satisfy the pointwise upper bounds $c_k^{\mathcal{X}} \leq \|c\|_\infty \delta_k$ and $c_k^{\mathcal{Y}} \leq \|c\|_\infty \delta_k$ where $\delta_k(x_1^k,\tilde{x}_1^k) = \sum_{\ell=1}^k \delta(x_\ell \neq \tilde{x}_\ell)$ is the k-step Hamming distance. Thus,

$$\begin{aligned}\left|\frac{1}{k}\mathcal{T}_\eta\left(c_k;\alpha_k,\beta_k\right) - \frac{1}{k}\mathcal{T}_\eta\left(c_k;\alpha_k',\beta_k'\right)\right| \leq \frac{\|c\|_\infty}{k}\mathcal{T}(\delta_k;\alpha_k,\alpha_k') &+ \frac{\|c\|_\infty}{k}\mathcal{T}(\delta_k;\beta_k,\beta_k') \\ &+ \eta\left|\frac{1}{k}H_k(\alpha_k) - \frac{1}{k}H_k(\alpha_k')\right| \\ &+ \eta\left|\frac{1}{k}H_k(\beta_k) - \frac{1}{k}H_k(\beta_k')\right|.\end{aligned}$$

Letting $k \to \infty$ and applying Proposition 5.21, we find

$$\left|\mathcal{S}_\eta(c;\alpha,\beta) - \mathcal{S}_\eta(c;\alpha',\beta')\right| \leq \|c\|_\infty\left(\mathcal{S}(\delta;\alpha,\alpha') + \mathcal{S}(\delta;\beta,\beta')\right) + \eta|h(\alpha) - h(\alpha')| + \eta|h(\beta) - h(\beta')|.$$

Recognizing that $\overline{d}(\alpha,\alpha') = \mathcal{S}(\delta;\alpha,\alpha')$ and $\overline{d}(\beta,\beta') = \mathcal{S}(\delta;\beta,\beta')$, 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 5.30 ((Block approximation with gaps))**.** *Let $\mathcal{U}$ be a finite space and $k, g \geq 1$. We define $\tilde{\Lambda}^{k,g} : \mathcal{M}_s(\mathcal{U}^{\mathbb{N}}) \times \mathcal{M}(\mathcal{U}^g) \to \mathcal{M}(\mathcal{U}^{\mathbb{N}})$ 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 $C \subset \mathcal{U}^{\mathbb{N}}$,*

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

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

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

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

Note 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, $\Lambda^{k,g}[\mu, \alpha]$ and $\Lambda^{k,g}[\nu, \beta]$ are well-defined. We will show later that $\Lambda^{k,g}[\mu, \alpha]$ and $\Lambda^{k,g}[\nu, \beta]$ arise naturally in the proof of Theorem 5.16. 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 5.31. *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 $\alpha \in \mathcal{M}(\mathcal{U}^g)$,*

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

Proof. Fix $k \geq 1$, $g \geq 0$, $\alpha \in \mathcal{M}(\mathcal{U}^g)$. To simplify notation, let $\tilde{\xi} = \tilde{\Lambda}^{k,g}[\gamma, \alpha]$ and $\xi = \Lambda^{k,g}[\gamma, \alpha]$. 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 $\mathbf{U}$ generated as follows:

1. Draw a sequence $\mathbf{U} \in \mathcal{U}^{\mathbb{N}}$ according to γ.
2. For every $\ell \geq 0$, replace $U_{\ell(k+g)+k+1}^{(\ell+1)(k+g)}$ with a random draw from $\alpha \in \mathcal{M}(\mathcal{U}^g)$. In other words, replace the g-blocks of $\mathbf{U}$ with independent draws from α.

Thus $\tilde{\zeta}$ is comprised of alternating blocks of size k and g where letters in the k-blocks are distributed according to the corresponding letters of γ and the g-blocks are drawn independently according to α. 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 choosing a good coupling of γ and $\tilde{\zeta}$ and then randomizing the start to obtain a good 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 $(\mathbf{U}, \tilde{\mathbf{U}})$ generated as follows:

1. Sample $\mathbf{U} \in \mathcal{U}^{\mathbb{N}}$ according to γ.
2. Set $\tilde{\mathbf{U}} = \mathbf{U}$.
3. For every $\ell \geq 0$, replace $\tilde{U}_{\ell(k+g)+k+1}^{(\ell+1)(k+g)}$ with a random draw from $\alpha \in \mathcal{M}(\mathcal{U}^g)$. In other words, replace g-blocks of $\tilde{\mathbf{U}}$ with independent draws from α.

In particular, $U_\ell = \tilde{U}_\ell$ 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 and abusing notation somewhat, we obtain

$$\begin{aligned}\overline{d}(\gamma,\zeta) &\leq \int \delta(u_1 \neq \tilde{u}_1)\, d\lambda(\mathbf{u},\tilde{\mathbf{u}}) \\ &= \int \frac{1}{k+g} \sum_{\ell=1}^{k+g} \delta(u_\ell \neq \tilde{u}_\ell)\, d\pi(\mathbf{u},\tilde{\mathbf{u}}) \\ &= \frac{1}{k+g} \sum_{\ell=k+1}^{k+g} \int \delta(u_\ell \neq \tilde{u}_\ell)\, d\pi(\mathbf{u},\tilde{\mathbf{u}}) \\ &\leq \frac{g}{k+g}.\end{aligned}$$

Next we bound $\overline{d}(\zeta,\xi)$. By Proposition 5.4,

$$\overline{d}(\zeta,\xi) = \lim_{m\to\infty} \frac{1}{m} \mathcal{T}\left(\delta_m; \zeta_m, \xi_m\right)$$

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

$$\overline{d}(\zeta,\xi) = \lim_{L\to\infty} \frac{1}{m(L)} \mathcal{T}\left(\delta_{m(L)}; \zeta_{m(L)}, \xi_{m(L)}\right). \tag{5.12}$$

It suffices to obtain a bound on

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

for fixed $L \in \mathbb{N}_0$ and take a limit as $L \to \infty$. Similar to the first bound, we will achieve this by constructing a good coupling of $\tilde{\zeta}_{m(L)}$ and $\tilde{\xi}_{m(L)}$ and randomizing the start to obtain a good coupling of $\zeta_{m(L)}$ and $\xi_{m(L)}$. 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 $\tilde{\xi}_{m(L)}$).

Fix an index $L \in \mathbb{N}$. Define the process measures $\rho^0, ..., \rho^L \in \mathcal{M}(\mathcal{U}^{\mathbb{N}})$ such that for every $\ell \in \{0, ..., L\}$ and $U_{m(\ell)} \subset \mathcal{U}^{m(\ell)}$, it holds that $\rho^\ell(U_{m(\ell)} \times \mathcal{U}^{\mathbb{N}}) = \tilde{\zeta}(U_{m(\ell)} \times \mathcal{U}^{\mathbb{N}})$ and for every measurable $U \subset \mathcal{U}^{\mathbb{N}}$, it holds that $\rho^\ell(\mathcal{U}^{m(\ell)} \times U) = \tilde{\xi}(\mathcal{U}^{m(\ell)} \times U)$. 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 again,

$$\mathcal{T}\left(\delta_{m(L)}; \tilde{\zeta}_{m(L)}, \tilde{\xi}_{m(L)}\right) \leq \sum_{\ell=0}^{L-1} \mathcal{T}\left(\delta_{m(L)}; \rho^\ell_{m(L)}, \rho^{\ell+1}_{m(L)}\right). \tag{5.13}$$

In order to bound the terms on the right hand side of (5.13), we will construct a good coupling of $\rho^\ell_{m(L)}$ and $\rho^{\ell+1}_{m(L)}$ for each $\ell \in \{0, ..., L-1\}$. In particular, we will couple $\rho^\ell_{m(L)}$ 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 let $\pi^\ell \in \Pi(\rho^\ell, \rho^{\ell+1})$ be the coupling corresponding to the distribution of the process $(\mathbf{U}, \tilde{\mathbf{U}})$ generated as follows:

1. Draw $U_1^{m(\ell)+g} \in \mathcal{U}^{m(\ell)+g}$ according to $\tilde{\zeta}_{m(\ell)+g}$.
2. Set $\tilde{U}_1^{m(\ell)+g} = U_1^{m(\ell)+g}$.
3. Draw $(U_{m(\ell)+g+1}^{m(\ell+1)}, \tilde{U}_{m(\ell)+g+1}^{m(\ell+1)})$ according to an optimal coupling π^* of the conditional distributions of the next k-block $\rho^\ell_k(\cdot | U_1^{m(\ell)+g})$ and $\rho^{\ell+1}_k(\cdot | \tilde{U}_1^{m(\ell)+g})$ with respect to the cost δ_k.
4. Draw $U_{m(\ell+1)+1}^{m(\ell+1)+g}$ according to α and $U_{m(\ell+1)+g+1}^{m(L)}$ according to $\tilde{\xi}_{m(L)-m(\ell+1)-g}$.
5. Set $\tilde{U}_{m(\ell+1)+1}^{m(L)} = U_{m(\ell+1)+1}^{m(L)}$.

We note that for any $(u_1^{m(L)}, \tilde{u}_1^{m(L)}) \in \mathcal{U}^{m(L)} \times \mathcal{U}^{m(L)}$, one may write,

$$\begin{aligned}
\pi^\ell(u_1^{m(L)}, \tilde{u}_1^{m(L)}) = &\underbrace{\tilde{\zeta}_{m(\ell)+g}(u_1^{m(\ell)+g})}_{\text{Step 1}} \underbrace{\delta(u_1^{m(\ell)+g} = \tilde{u}_1^{m(\ell)+g})}_{\text{Step 2}} \\
&\times \underbrace{\pi^*\left((u_{m(\ell)+g+1}^{m(\ell+1)}, \tilde{u}_{m(\ell)+g+1}^{m(\ell+1)}) | (u_1^{m(\ell)+g}, \tilde{u}_1^{m(\ell)+g})\right)}_{\text{Step 3}} \\
&\times \underbrace{\alpha(u_{m(\ell+1)+1}^{m(\ell+1)+g}) \tilde{\xi}_{m(L)-m(\ell+1)-g}(u_{m(\ell+1)+g+1}^{m(L)})}_{\text{Step 4}} \underbrace{\delta(u_{m(\ell+1)+1}^{m(L)} = \tilde{u}_{m(\ell+1)+1}^{m(L)})}_{\text{Step 5}}.
\end{aligned}$$

In particular, $u_1^{m(\ell)+g} = \tilde{u}_1^{m(\ell)+g}$ and $u_{m(\ell+1)+1}^{m(L)} = \tilde{u}_{m(\ell+1)+1}^{m(L)}$ with π^ℓ-probability one. It follows that

$$\mathcal{T}\left(\delta_{m(L)}; \rho^\ell_{m(L)}, \rho^{\ell+1}_{m(L)}\right) \leq \int_{\mathcal{U}^{m(L)} \times \mathcal{U}^{m(L)}} \delta_{m(L)}(u_1^{m(L)}, \tilde{u}_1^{m(L)}) \, d\pi^\ell(u_1^{m(L)}, \tilde{u}_1^{m(L)})$$

$$= \int_{\mathcal{U}^{m(L)} \times \mathcal{U}^{m(L)}} \delta_k(u_{m(\ell)+g+1}^{m(\ell+1)}, \tilde{u}_{m(\ell)+g+1}^{m(\ell+1)})\, d\pi^\ell(u_1^{m(L)}, \tilde{u}_1^{m(L)}).$$

By the construction of π^ℓ, this implies

$$\mathcal{T}\left(\delta_{m(L)}; \rho_{m(L)}^\ell, \rho_{m(L)}^{\ell+1}\right) \leq \int_{\mathcal{U}^{m(\ell)+g}} \mathcal{T}\left(\delta_k; \rho_k^\ell(\cdot|u_1^{m(\ell)+g}), \rho_k^{\ell+1}(\cdot|u_1^{m(\ell)+g})\right)\, d\tilde{\zeta}_{m(\ell)+g}(u_1^{m(\ell)+g}).$$

Finally, using the fact that the optimal transport cost with respect to δ_k is bounded by k times the total variation distance, we have

$$\mathcal{T}\left(\delta_{m(L)}; \rho_{m(L)}^\ell, \rho_{m(L)}^{\ell+1}\right) \leq k \sup_{u_1^{m(\ell)+g}} \sup_{A \subset \mathcal{U}^k} \left|\rho_k^{\ell+1}(A|u_1^{m(\ell)+g}) - \rho_k^\ell(A|u_1^{m(\ell)+g})\right|.$$

Now since k-block $\ell+2$ of ρ^ℓ is independent of all previous blocks, it follows that $\rho_k^\ell(\cdot|u_1^{m(\ell)+g}) = \gamma_k$ for every $u_1^{m(\ell)+g} \in \mathcal{U}^{m(\ell)+g}$. Moreover, $\rho_k^{\ell+1}(\cdot|u_1^{\ell(k+g)}) = \gamma_k(\cdot|C(u_1^{\ell(k+g)}))$ where $C(u_1^{\ell(k+g)})$ is the set obtained by taking the union of $u_1^{\ell(k+g)}$ over all possible g-blocks, i.e.,

$$C(u_1^{\ell(k+g)}) = \bigcup_{i=0}^{\ell-1} \bigcup_{u_{m(i)+1}^{m(i+1)}} \{u_1^{\ell(k+g)}\}.$$

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

$$\mathcal{T}\left(\delta_{m(L)}; \rho_{m(L)}^\ell, \rho_{m(L)}^{\ell+1}\right) \leq k \sup_{u_1^{m(\ell)+g}} \sup_{A \in \mathcal{U}^k} \left|\gamma_k(A|C(u_1^{\ell(k+g)})) - \gamma_k(A)\right| \leq k\phi_\gamma(g+1).$$

Plugging this result into (5.13),

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

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

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

Plugging this into (5.12) 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 Mean k-Step Optimal Transport Cost. In the final step before proving Theorem 5.16, 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 (Boissard and Le Gouic, 2014, Proposition 1.7), stated below as Theorem C, 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 $\mathbf{U} = (U_1, U_2, ...) : \Omega \to \mathcal{U}^{\mathbb{N}}$ distributed according to γ,

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

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 C ((Boissard and Le Gouic, 2014))**.** *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 $\chi_n = \frac{1}{n^2}\sum_{m=0}^{n}\sum_{g=0}^{m}\rho_\gamma(g)$ and let $\Delta = \mathrm{Diam}(\mathcal{U})$. 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 $t \in (0, \Delta/4]$,*

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

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 5.32. *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 $t \in (0, \|c\|_\infty/4]$,*

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

Proof. The result follows from two applications of Theorem C for μ and ν. Considering first the case of μ, let $\mathcal{U} = \mathcal{X}^k$ with pseudo-metric $d = \frac{1}{k}c_k^{\mathcal{X}}$. Moreover, let $\tilde{\mu}^k \in \mathcal{M}_s((\mathcal{X}^k)^{\mathbb{N}})$ be the distribution of the stationary process $(X_1, ..., X_k), (X_2, ..., X_{k+1}), ...$ where $(X_1, X_2, ...)$ is drawn according to μ. Note that the ρ- and ϕ-mixing coefficients of $\tilde{\mu}^k$ satisfy $\rho_{\tilde{\mu}^k}(g) \leq 2\phi_{\tilde{\mu}^k}^{1/2}(g)$ for every $g \geq 0$ (Bradley, 2005). One may also easily establish that $\phi_{\tilde{\mu}^k}(g) \leq \phi_\mu(g)$ for every $g \geq 0$. Then a direct application of Theorem C to $\tilde{\mu}^k$ yields

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

for some constant $C < \infty$ and any $t \in (0, \frac{\Delta}{4}]$, where the $n - k + 1$ term comes from the fact that one has $n - k + 1$ 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 5.16 and Corollary 5.17.

Theorem 5.16. *Let μ and ν have ϕ-mixing coefficients ϕ_μ and ϕ_ν, respectively. Then there exists a universal constant $C < \infty$ such that for every $n \geq 1$, $k \in \{1, ..., n\}$, $g \geq 0$ and $t \in (0, \|c\|_\infty/4]$,*

$$\mathbb{E}\left[|\hat{\rho}_k(X_1^n, Y_1^n) - \mathcal{S}(c; \mu, \nu)|\right] \leq \|c\|_\infty\left(\frac{k(\phi_\mu(g+1) + \phi_\nu(g+1))}{k+g} + \frac{3g}{k}\right)$$

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

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

$$\begin{aligned}
&|\hat\rho_k(X_1^n, Y_1^n) - \mathcal{S}(c;\mu,\nu)| \\
&\quad = \left|\frac{1}{k}\mathcal{T}(c_k;\hat\mu_{k,n},\hat\nu_{k,n}) - \mathcal{S}(c;\mu,\nu)\right| \\
&\quad \leq \left|\frac{1}{k}\mathcal{T}(c_k;\hat\mu_{k,n},\hat\nu_{k,n}) - \frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k)\right| + \left|\frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k) - \mathcal{S}(c;\mu,\nu)\right|.
\end{aligned}$$

We begin by establishing an upper bound on $\left|\frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k) - \mathcal{S}(c;\mu,\nu)\right|$. Note by Proposition 5.4, $\frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k) \leq \mathcal{S}(c;\mu,\nu)$, so it suffices to upper bound $\mathcal{S}(c;\mu,\nu) - \frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k)$. Let $\pi_k \in \Pi(\mu_k,\nu_k)$ achieve the minimum in the problem $\mathcal{T}(c_k;\mu_k,\nu_k)$ and let $\gamma \in \mathcal{M}((\mathcal{X}\times\mathcal{Y})^g)$ be any probability measure on $(\mathcal{X}\times\mathcal{Y})^g$. Denote the $\mathcal{X}^g$ and $\mathcal{Y}^g$ marginals of γ by $\alpha \in \mathcal{M}(\mathcal{X}^g)$ 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 $\lambda^{k,g} = \Lambda^{k+g}[\pi_k \otimes \gamma]$. Note that $\lambda^{k,g} \in \mathcal{J}(\Lambda^{k,g}[\mu,\alpha], \Lambda^{k,g}[\nu,\beta])$. Then by the construction of $\lambda^{k,g}$,

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

Rearranging terms and adding $\mathcal{S}(c;\mu,\nu)$ to both sides, we obtain that

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

Now using the fact that $\frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k) \leq \mathcal{S}(c;\mu,\nu)$ as established in Lemma 5.33, we have

$$\begin{aligned}
\left|\frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k) - \mathcal{S}(c;\mu,\nu)\right| &= \mathcal{S}(c;\mu,\nu) - \frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k) \\
&\leq \mathcal{S}(c;\mu,\nu) - \frac{k+g}{k}\mathcal{S}\left(c;\Lambda^k[\mu,\alpha],\Lambda^k[\nu,\beta]\right) + \frac{g}{k}\|c\|_\infty
\end{aligned}$$

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

By Lemmas 5.29 and 5.31, we see that

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

Considering the other term, Lemma 5.25 gives us

$$\left|\frac{1}{k}\mathcal{T}(c_k;\hat{\mu}_{k,n},\hat{\nu}_{k,n}) - \frac{1}{k}\mathcal{T}(c_k;\mu_k,\nu_k)\right| \leq \frac{1}{k}\mathcal{T}(c_k^{\mathcal{X}};\hat{\mu}_{k,n},\mu_k) + \frac{1}{k}\mathcal{T}(c_k^{\mathcal{Y}};\hat{\nu}_{k,n},\nu_k).$$

Combining the bounds proven above, taking an expectation, and applying Proposition 5.32, we obtain the result. □

Corollary 5.17. *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 $n \geq 1$, $k \in \{1,...,n\}$ and $g \geq 0$,

$$\mathbb{E}\left[|\hat{\rho}_k(X_1^n,Y_1^n) - \mathcal{S}(c;\mu,\nu)|\right] \leq \|c\|_\infty\left(\frac{k(\phi_\mu(g+1)+\phi_\nu(g+1))}{k+g} + \frac{3g}{k} + \frac{C(|\mathcal{X}|^{k/2}+|\mathcal{Y}|^{k/2})}{n^{1-p/2}}\right).$$

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

Proof. For every $\varepsilon \in (0,\|c\|_\infty/4]$, we have $\mathcal{N}(\mathcal{X}^k,\frac{1}{k}c_k^{\mathcal{X}},\varepsilon) \leq |\mathcal{X}|^k$ and $\mathcal{N}(\mathcal{Y}^k,\frac{1}{k}c_k^{\mathcal{Y}},\varepsilon) \leq |\mathcal{Y}|^k$. After applying these inequalities and the summability conditions for ϕ_μ and ϕ_ν in Theorem 5.16, we obtain the result by letting $t \to 0$. □

5.7.5 Proofs from Section 5.5

In this section, we prove the results stated in Section 5.5. We begin with Lemma 5.33, which states that the limit in Proposition 5.21 exists and is equal to a supremum.

Lemma 5.33. *For any $\eta \geq 0$,*

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

Proof. By Fekete's lemma, it suffices to show that the sequence $\{\mathcal{T}_\eta(c_k; \mu_k, \nu_k)\}_{k\geq 1}$ is superadditive. Fix $k, \ell \geq 1$ and let $\pi \in \Pi(\mu_{k+\ell}, \nu_{k+\ell})$ be a solution to $\mathcal{T}_\eta(c_{k+\ell}; \mu_{k+\ell}, \nu_{k+\ell})$. 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 $H_{k+\ell}(\cdot)$,

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

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

Proposition 5.21. *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 $\eta \geq 0$,*

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

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

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

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

$$\mathcal{T}_\eta(c_k; \mu_k, \nu_k) \leq \int c_k\, d\lambda_k - \eta H_k(\lambda_k).$$

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

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

By Lemma 5.33 we may take a limit in k and let $\varepsilon \to 0$ to establish that

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

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

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

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, written as $m_\mathcal{X}(\lambda^k)$ and $m_\mathcal{Y}(\lambda^k)$, 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 $m_{\mathcal{X}}(\lambda^k) \Rightarrow \mu$ and one may use a similar argument to show that $m_{\mathcal{Y}}(\lambda^k) \Rightarrow \nu$. So by Lemma C.3, $\lambda^{k_\ell} \Rightarrow \lambda \in \mathcal{J}(\mu,\nu)$ 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 $h(\lambda^{k_\ell}) = \frac{1}{k_\ell} H_{k_\ell}(\pi^{k_\ell})$. Thus, by the upper semicontinuity of $h(\cdot)$ and the continuity and boundedness of c,

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

giving the result. □

Proposition 5.22. *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*

$$\lim_{\eta\to 0} \mathcal{S}_\eta(c;\mu,\nu) = \mathcal{S}(c;\mu,\nu).$$

Proof. Let $\{\eta_n\}$ be a sequence of non-negative integers such that $\eta_n \to 0$ and for every $n \geq 1$ let $\lambda^n \in \mathcal{J}_{\min}^{\eta_n}(\mu,\nu)$. 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 $\mathcal{S}(c;\mu,\nu)$ and λ^* for $\mathcal{S}_{\eta_n}(c;\mu,\nu)$, 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

$$0 \leq \int c \, d\lambda_1^n - \int c \, d\lambda_1^* \leq \eta_n (h(\lambda^n) - h(\lambda^*)). \tag{5.14}$$

As $h(\cdot)$ is bounded, we have $\lim_{n\to\infty} \eta_n(h(\lambda^n) - h(\lambda^*)) = 0$. Taking limits in (5.14) and using the continuity and boundedness of c,

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

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

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

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

Theorem 5.23. *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 $\{k(n)\}$ with $k(n) \to \infty$ such that with $\mathbb{P}$-probability one, $\hat{\rho}^{\eta}_{k,n} \to \mathcal{S}_\eta(c; \mu, \nu)$ and $\hat{\lambda}^{\eta,k,n} \Rightarrow \mathcal{J}^{\eta}_{min}(\mu, \nu)$ as $n \to \infty$.*

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

$$\lim_{n\to\infty} \left| \hat{\rho}^{\eta}_{k(n)}(X_1^n, Y_1^n) - \mathcal{S}_\eta\left(c; \mu, \nu\right) \right| = 0, \quad \mathbb{P} - a.s. \tag{5.15}$$

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 5.26,

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

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 $H_i(\cdot)$ is weakly continuous, there exists an $n(i) \in \mathbb{N}$ such that

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

Then by Borel-Cantelli,

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

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

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

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 an element $\omega \in \Omega$ of the sample space in the set of $\mathbb{P}$-measure one on which (5.15) holds. Let $\{\hat{\lambda}^{\eta,k,n}\}_{n\geq 1}$ be the corresponding sequence of estimated entropic optimal joinings where the dependence on the observations $X_1^n(\omega)$ and $Y_1^n(\omega)$ has been suppressed. By Lemma C.3, for any subsequence $\{\hat{\lambda}^{\eta,k,n_\ell}\}_{\ell\geq 1}$, 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 $\{\hat{\lambda}^{\eta,k,n_\ell}\}_{\ell\geq 1}$. Using the weak lower semicontinuity of $-h(\cdot)$ and the continuity and boundedness of c, one may establish

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

Thus, $\lambda \in \mathcal{J}^{\eta}_{\min}(\mu,\nu)$ and since the subsequence was arbitrary, we conclude that $\hat{\lambda}^{\eta,k,n} \Rightarrow \mathcal{J}^{\eta}_{\min}(\mu,\nu)$. By the choice of ω, we conclude that $\hat{\lambda}^{\eta,k,n} \Rightarrow \mathcal{J}^{\eta}_{\min}(\mu,\nu)$, $\mathbb{P}$-almost surely. The proof of the upper bound on $\mathbb{E}[|\hat{\rho}^{\eta}_{k}(X_1^n, Y_1^n) - \mathcal{S}_{\eta}(c;\mu,\nu)|]$ may be found in the proof of Theorem 5.16. □

APPENDIX A

APPENDIX TO CHAPTER 3

A.1 Properties of the OTC Problems

In this appendix, we prove that solutions to the OTC and constrained OTC problems exist via continuity and compactness arguments and establish the triangle inequality for the unconstrained problem. For a metric space $\mathcal{U}$ and a sequence of Borel probability measures $\{\mu^n\} \subset \mathcal{M}(\mathcal{U})$, we say that μ^n *converges weakly to* $\mu \in \mathcal{M}(\mathcal{U})$, denoted by $\mu^n \Rightarrow \mu$, if for every continuous and bounded function $f : \mathcal{U} \to \mathbb{R}$, $\int f \, d\mu^n \to \int f \, d\mu$. A set $\Pi \subset \mathcal{M}(\mathcal{U})$ is said to be *weakly compact* if every sequence in Π contains a subsequence converging weakly to an element of Π. Π is said to be *tight* if for every $\varepsilon > 0$, there exists a compact set $K \subset \mathcal{U}$ such that $\mu(K) > 1 - \varepsilon$ for every $\mu \in \Pi$. Tightness and relative compactness are related by Prohorov's theorem which states that if $\mathcal{U}$ is a separable metric space, $\Pi \subset \mathcal{M}(\mathcal{U})$ is tight if and only if its closure is relatively compact. Note that $\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ is complete and separable when equipped with the metric

$$d((\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^2, \mathbf{y}^2)) = \sum_{k=0}^{\infty} 2^{-k} \delta((x_k^1, y_k^1) \neq (x_k^2, y_k^2)).$$

Finally, we remark that since $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ is continuous and bounded, $\tilde{c}(\mathbf{x}, \mathbf{y}) = c(x_0, y_0)$ is as well.

A.1.1 Existence for the OTC Problem

We begin by proving that $\Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q})$ is weakly compact.

Lemma A.1. $\Pi_{TC}(\mathbb{P}, \mathbb{Q})$ *is weakly compact.*

Proof. By (Villani, 2008, Lemma 4.4), $\Pi(\mathbb{P}, \mathbb{Q})$ is tight. Since $\Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q}) \subset \Pi(\mathbb{P}, \mathbb{Q})$, $\Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q})$ is tight as well. Thus by Prohorov's theorem, the closure of $\Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q})$ is weakly compact. So we need only prove that $\Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q})$ is closed. Take a sequence $\{\pi^n\} \subset \Pi_{\text{TC}}(\mathbb{P}, \mathbb{Q})$ such that $\pi^n \Rightarrow \pi \in \mathcal{M}(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$. Since $\Pi(\mathbb{P}, \mathbb{Q})$ is weakly compact (Villani, 2008), $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$. Then it suffices to prove that π is stationary, Markov, and has a transition matrix that satisfies the transition coupling property.

We begin by proving that π is stationary. Let $\sigma : \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ be the left-shift map defined for every $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ by $\sigma(\mathbf{x}, \mathbf{y}) = (x_1^\infty, y_1^\infty)$. Then stationarity of any $\mu \in \mathcal{M}(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$ is defined by $\mu = \mu \circ \sigma^{-1}$. Since each π^n is stationary, $\pi^n = \pi^n \circ \sigma^{-1}$. Noting that σ is continuous, the continuous mapping theorem implies that $\pi^n \circ \sigma^{-1} \Rightarrow \pi \circ \sigma^{-1}$, so $\pi^n \Rightarrow \pi \circ \sigma^{-1}$. Since weak limits are unique, we conclude that $\pi = \pi \circ \sigma^{-1}$ and π is stationary.

Next we prove that π is Markov. Since $\mathcal{X} \times \mathcal{Y}$ is finite, for any cylinder set $[s_0^k] = \{(\mathbf{x}, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^{\mathbb{N}} : (x_j, y_j) = s_j, 0 \leq j \leq k\}$, $\pi^n([s_0^k]) \to \pi([s_0^k])$. Then

$$\frac{\pi^n([s_0 \cdots s_k])}{\pi^n([s_0 \cdots s_{k-1}])} \to \frac{\pi([s_0 \cdots s_k])}{\pi([s_0 \cdots s_{k-1}])} \tag{A.1}$$

and

$$\frac{\pi^n([s_{k-1} s_k])}{\pi^n([s_{k-1}])} \to \frac{\pi([s_{k-1} s_k])}{\pi([s_{k-1}])}, \tag{A.2}$$

where we let $0/0 = 0$. But since π^n is Markov for each $n \geq 1$,

$$\frac{\pi^n([s_0 \cdots s_k])}{\pi^n([s_0 \cdots s_{k-1}])} = \frac{\pi^n([s_{k-1} s_k])}{\pi^n([s_{k-1}])}.$$

As a result, $\pi([s_0 \cdots s_k])/\pi([s_0 \cdots s_{k-1}]) = \pi([s_{k-1} s_k])/\pi([s_{k-1}])$. Thus, π is Markov.

Now, we need only show that π satisfies the transition coupling property. Letting R_n and R denote the transition matrices of π^n and π, respectively, (A.1) and (A.2) imply that $R_n(s, s') \to R(s, s')$ for every $s, s' \in \mathcal{X} \times \mathcal{Y}$. Then for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $y' \in \mathcal{Y}$,

$$\sum_{x'} R_n((x, y), (x', y')) \to \sum_{x'} R((x, y), (x', y')). \tag{A.3}$$

But as $R_n \in \Pi_{\mathrm{TC}}(P, Q)$, $\sum_{x'} R_n((x, y), (x', y')) = Q(y, y')$ and it follows that $\sum_{x'} R((x, y), (x', y')) = Q(y, y')$. Employing a similar argument to the other marginal of R, one may show that in fact $R \in \Pi_{\mathrm{TC}}(P, Q)$. Therefore, $\pi \in \Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$ and we conclude that $\Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$ is weakly compact. □

Proposition A.2. *The OTC problem* (3.2) *has a solution.*

Proof. Let $\{\pi^n\} \subset \Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$ be a sequence such that

$$\int \tilde{c}\, d\pi^n \to \inf_{\pi \in \Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})} \int \tilde{c}\, d\pi.$$

By Lemma A.1, $\Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$ is weakly compact. Thus, there exists a subsequence $\{\pi^{n_k}\}$ such that $\pi^{n_k} \Rightarrow \pi^*$ for some $\pi^* \in \Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$. Since $\tilde{c}$ is continuous and bounded,

$$\int \tilde{c}\, d\pi^* = \lim_{k\to\infty} \int \tilde{c}\, d\pi^{n_k} = \inf_{\pi \in \Pi_{\mathrm{TC}}(\mathbb{P},\mathbb{Q})} \int \tilde{c}\, d\pi.$$

Thus π^* is an optimal solution for Problem (3.2). □

A.1.2 Existence for the Constrained OTC Problem

We begin by proving that $\Pi^{\eta}_{\mathrm{TC}}(P, Q)$ is convex and compact as a subset of $\mathbb{R}^{d^2 \times d^2}$.

Lemma A.3. *For any $\eta > 0$, the constrained set of transition coupling matrices $\Pi^{\eta}_{TC}(P, Q)$ is convex and compact.*

Proof. Fixing $\eta > 0$, we begin by showing that $\Pi^{\eta}_{\mathrm{TC}}(P, Q)$ is convex. Let $R, R' \in \Pi^{\eta}_{\mathrm{TC}}(P, Q)$, $\lambda \in (0, 1)$, and define $R_\lambda := \lambda R + (1-\lambda)R'$. Since $\Pi_{\mathrm{TC}}(P, Q)$ is convex, $R_\lambda \in \Pi_{\mathrm{TC}}(P, Q)$. Moreover, using the convexity of the KL-divergence, for any $s \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned}
\mathcal{K}(R_\lambda(s, \cdot) \| P \otimes Q(s, \cdot)) &= \mathcal{K}(\lambda R(s, \cdot) + (1-\lambda)R'(s, \cdot) \| P \otimes Q(s, \cdot)) \\
&\leq \lambda \mathcal{K}(R(s, \cdot) \| P \otimes Q(s, \cdot)) + (1-\lambda)\mathcal{K}(R'(s, \cdot) \| P \otimes Q(s, \cdot)) \\
&\leq \lambda\eta + (1-\lambda)\eta \\
&= \eta.
\end{aligned}$$

Thus $R_\lambda \in \Pi^{\eta}_{\mathrm{TC}}(P, Q)$ and we conclude that $\Pi^{\eta}_{\mathrm{TC}}(P, Q)$ is convex.

Next we prove compactness. Note that as a subset of the compact set $\Pi_{\mathrm{TC}}(P, Q)$ we need only show that $\Pi^{\eta}_{\mathrm{TC}}(P, Q)$ is closed. Let $\{R_n\} \subset \Pi^{\eta}_{\mathrm{TC}}(P, Q)$ be a sequence converging to $R \in \mathbb{R}^{d^2 \times d^2}$. By the compactness of $\Pi_{\mathrm{TC}}(P, Q)$, $R \in \Pi_{\mathrm{TC}}(P, Q)$. Now for any $s \in \mathcal{X} \times \mathcal{Y}$, note that $R(s, \cdot)$ is absolutely continuous with respect to $P \otimes Q(s, \cdot)$. This implies that, for every $s' \in \mathcal{X} \times \mathcal{Y}$,

$$R(s, s') \log \frac{R(s, s')}{P \otimes Q(s, s')} < \infty,$$

where we let $0 \log(0/0) = 0$. Then $\mathcal{K}(\cdot \| P \otimes Q(s, \cdot))$ is continuous at $R(s, \cdot)$ and we have that

$$\mathcal{K}(R(s, \cdot) \| P \otimes Q(s, \cdot)) = \lim_{n \to \infty} \mathcal{K}(R_n(s, \cdot) \| P \otimes Q(s, \cdot)) \leq \eta.$$

Thus $R \in \Pi_{\mathrm{TC}}^{\eta}(P, Q)$ and we conclude that $\Pi_{\mathrm{TC}}^{\eta}(P, Q)$ is compact. □

Next, we show that $\Pi_{\mathrm{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})$ is weakly compact.

Lemma A.4. *For any $\eta \geq 0$, $\Pi_{TC}^{\eta}(\mathbb{P}, \mathbb{Q})$ is weakly compact.*

Proof. Let $\{\pi_n\} \subset \Pi_{\mathrm{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})$ be a sequence such that $\pi_n \Rightarrow \pi \in \mathcal{M}(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}})$. By Lemma A.3, $\Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$ is weakly compact so $\pi \in \Pi_{\mathrm{TC}}(\mathbb{P}, \mathbb{Q})$. Letting R be the transition matrix of π, we need only show that $R \in \Pi_{\mathrm{TC}}^{\eta}(P, Q)$. Letting R_n be the transition matrix of π_n, it follows from (A.3) that $R_n \to R$. Using the weak lower semicontinuity of the KL-divergence, for every $s \in \mathcal{X} \times \mathcal{Y}$,

$$\mathcal{K}(R(s, \cdot) \| P \otimes Q(s, \cdot)) \leq \liminf_{n \to \infty} \mathcal{K}(R_n(s, \cdot) \| P \otimes Q(s, \cdot)) \leq \eta.$$

Therefore, $R \in \Pi_{\eta}(P, Q)$ and we find that $\pi \in \Pi_{\mathrm{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})$. Thus, we conclude that $\Pi_{\mathrm{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})$ is weakly compact. □

Proposition A.5. *For any $\eta > 0$, the constrained OTC problem* (3.4) *has a solution.*

Proof. Let $\{\pi^n\} \subset \Pi_{\mathrm{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})$ be a sequence such that

$$\int \tilde{c} \, d\pi^n \to \inf_{\pi \in \Pi_{\mathrm{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})} \int \tilde{c} \, d\pi.$$

By Lemma A.4, $\Pi_{\mathrm{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})$ is weakly compact. So there exists a subsequence $\{\pi^{n_k}\}$ such that $\pi^{n_k} \Rightarrow \pi^*$ for some $\pi^* \in \Pi_{\mathrm{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})$. Since $\tilde{c}$ is continuous and bounded,

$$\int \tilde{c} \, d\pi^* = \lim_{k \to \infty} \int \tilde{c} \, d\pi^{n_k} = \inf_{\pi \in \Pi_{\mathrm{TC}}^{\eta}(\mathbb{P}, \mathbb{Q})} \int \tilde{c} \, d\pi.$$

Thus π^* is an optimal solution for Problem (3.4). □

A.1.3 Triangle Inequality

Next we prove that the optimal transition coupling cost satisfies the triangle inequality when the cost does. For probability measures p_1, p_2, $p_3 \in \mathcal{M}(\mathcal{X})$, we let $\Pi(p_1, p_2, p_3)$ denote the set of three-way couplings of p_1, p_2, and p_3 defined in the obvious way. For stationary Markov process measures $\mathbb{P}_1$, $\mathbb{P}_2$, $\mathbb{P}_3 \in \mathcal{M}(\mathcal{X}^{\mathbb{N}})$ we let $\Pi_{\text{TC}}(\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3)$ denote the set of three-way transition couplings of $\mathbb{P}_1$, $\mathbb{P}_2$, and $\mathbb{P}_3$, again defined in the obvious way. If the three process measures have transition matrices P_1, P_2, and $P_3 \in \mathbb{R}^{d \times d}$, we let $\Pi(P_1, P_2, P_3)$ denote the set of three-way transition coupling matrices of P_1, P_2, and P_3.

Lemma A.6 ((Gluing Lemma))**.** *Let $\mathbb{P}_1$, $\mathbb{P}_2$, $\mathbb{P}_3 \in \mathcal{M}(\mathcal{X}^{\mathbb{N}})$ be stationary and irreducible Markov chains with stationary distributions p_1, p_2, $p_3 \in \mathcal{M}(\mathcal{X})$, and let $\pi_{12} \in \Pi_{TC}(\mathbb{P}_1, \mathbb{P}_2)$ and $\pi_{23} \in \Pi_{TC}(\mathbb{P}_2, \mathbb{P}_3)$. Then there exists $\pi_{123} \in \Pi_{TC}(\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3)$ such that $\pi_{123}(A_1 \times A_2 \times \mathcal{X}^{\mathbb{N}}) = \pi_{12}(A_1 \times A_2)$ and $\pi_{123}(\mathcal{X}^{\mathbb{N}} \times A_2 \times A_3) = \pi_{23}(A_2 \times A_3)$ for any A_1, A_2, $A_3 \subset \mathcal{X}^{\mathbb{N}}$. Furthermore, any stationary distribution $\lambda_{123} \in \mathcal{M}(\mathcal{X} \times \mathcal{X} \times \mathcal{X})$ of R_{123} necessarily satisfies $\lambda_{123} \in \Pi(p_1, p_2, p_3)$.*

Proof. Let $\mathbb{P}_1$, $\mathbb{P}_2$, $\mathbb{P}_3$, π_{12} and π_{23} have transition matrices P_1, P_2, P_3, R_{12} and R_{23}, respectively. By the gluing lemma for optimal couplings (Villani, 2008), for every x_1, x_2, $x_3 \in \mathcal{X}$, there exists a coupling $r_{(x_1,x_2,x_3)} \in \Pi(P_1(x_1, \cdot), P_2(x_2, \cdot), P_3(x_3, \cdot))$ such that

$$\sum_{\tilde{x}_3} r_{(x_1,x_2,x_3)}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = R_{12}((x_1, x_2), (\tilde{x}_1, \tilde{x}_2))$$

and

$$\sum_{\tilde{x}_1} r_{(x_1,x_2,x_3)}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = R_{23}((x_2, x_3), (\tilde{x}_2, \tilde{x}_3)).$$

Let $R_{123} \in \mathbb{R}^{d^3 \times d^3}$ be the transition matrix such that for every $(x_1, x_2, x_3), (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) \in \mathcal{X} \times \mathcal{X} \times \mathcal{X}$, $R_{123}((x_1, x_2, x_3), (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)) = r_{(x_1,x_2,x_3)}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$. By construction, $R_{123} \in \Pi(P_1, P_2, P_3)$ and we may let π_{123} be the stationary Markov process measure constructed from R_{123} and some stationary distribution $\lambda_{123} \in \mathcal{M}(\mathcal{X} \times \mathcal{X} \times \mathcal{X})$ of R_{123}. To see that $\lambda_{123} \in \Pi(p_1, p_2, p_3)$, let the first $\mathcal{X}$-marginal

$$R = \begin{array}{cc} & \begin{array}{ccccccccc} (0,0) & (0,1) & (0,2) & (1,0) & (1,1) & (1,2) & (2,0) & (2,1) & (2,2) \end{array} \\ \begin{array}{c} (0,0) \\ (0,1) \\ (0,2) \\ (1,0) \\ (1,1) \\ (1,2) \\ (2,0) \\ (2,1) \\ (2,2) \end{array} & \left[\begin{array}{ccccccccc} 0 & 0.25 & 0 & 0.25 & 0 & 0 & 0 & 0 & 0.50 \\ 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0 \\ 0.25 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0.25 & 0.25 & 0 & 0 & 0 & 0.25 & 0.25 \\ 0 & 0.25 & 0 & 0 & 0 & 0.25 & 0.50 & 0 & 0 \\ 0 & 0.25 & 0 & 0.25 & 0 & 0 & 0 & 0 & 0.50 \\ 0.25 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0.25 & 0.25 \\ 0 & 0.25 & 0 & 0 & 0 & 0.25 & 0.50 & 0 & 0 \end{array}\right] \end{array}.$$

Figure A.1: A reducible transition coupling of irreducible transition matrices P and Q defined in (A.5) and (A.6), respectively.

of λ_{123} be $\tilde{p}_1 \in \mathcal{M}(\mathcal{X})$. Then for every $x_1 \in \mathcal{X}$,

$$\begin{aligned} \tilde{p}_1(x_1) &= \sum_{x_2,x_3} \lambda_{123}(x_1,x_2,x_3) \\ &= \sum_{x_2,x_3} \sum_{\tilde{x}_1,\tilde{x}_2,\tilde{x}_3} \lambda_{123}(\tilde{x}_1,\tilde{x}_2,\tilde{x}_3) \times R_{123}((\tilde{x}_1,\tilde{x}_2,\tilde{x}_3),(x_1,x_2,x_3)) \\ &= \sum_{\tilde{x}_1,\tilde{x}_2,\tilde{x}_3} \lambda_{123}(\tilde{x}_1,\tilde{x}_2,\tilde{x}_3) P_1(\tilde{x}_1,x_1) \\ &= \sum_{\tilde{x}_1} \tilde{p}_1(\tilde{x}_1) P_1(\tilde{x}_1,x_1), \end{aligned}$$

so $\tilde{p}_1$ is stationary with respect to P_1. Since $\mathbb{P}_1$ is irreducible, the stationary distribution of P_1 is unique and it follows that $\tilde{p}_1 = p_1$. Repeating the argument for the second and third marginals, it follows that $\lambda_{123} \in \Pi(p_1,p_2,p_3)$ and thus $\pi_{123} \in \Pi_{\text{TC}}(\mathbb{P}_1,\mathbb{P}_2,\mathbb{P}_3)$. □

Proposition A.7 ((Triangle Inequality))**.** *Let $\mathbb{P}_1$, $\mathbb{P}_2$, $\mathbb{P}_3 \in \mathcal{M}(\mathcal{X}^{\mathbb{N}})$ be stationary and irreducible Markov chains and let $\tilde{c}(\mathbf{x},\tilde{\mathbf{x}}) = c(x_0,\tilde{x}_0)$ for every $\mathbf{x}$, $\tilde{\mathbf{x}} \in \mathcal{X}^{\mathbb{N}}$. If c satisfies the triangle inequality, then the OTC problem satisfies*

$$\min_{\pi \in \Pi_{TC}(\mathbb{P}_1,\mathbb{P}_3)} \int \tilde{c}\, d\pi \leq \min_{\pi \in \Pi_{TC}(\mathbb{P}_1,\mathbb{P}_2)} \int \tilde{c}\, d\pi + \min_{\pi \in \Pi_{TC}(\mathbb{P}_2,\mathbb{P}_3)} \int \tilde{c}\, d\pi. \tag{A.4}$$

Proof. By Proposition A.2, there exist $\pi_{12} \in \Pi_{\text{TC}}(\mathbb{P}_1,\mathbb{P}_2)$ and $\pi_{23} \in \Pi_{\text{TC}}(\mathbb{P}_2,\mathbb{P}_3)$ that are optimal in the two problems on the right hand side of (A.4). Then by Lemma A.6, there exists $\pi_{123} \in \Pi_{\text{TC}}(\mathbb{P}_1,\mathbb{P}_2,\mathbb{P}_3)$ that admits π_{12} and π_{23} as $(\mathcal{X} \times \mathcal{X})^{\mathbb{N}}$-marginals. Define the measure $\pi_{13} \in \mathcal{M}((\mathcal{X} \times$

$\mathcal{X})^{\mathbb{N}}$) by $\pi_{13}(A_1 \times A_3) = \pi_{123}(A_1 \times \mathcal{X}^{\mathbb{N}} \times A_3)$ for every $A_1,\ A_3 \subset \mathcal{X}^{\mathbb{N}}$. Clearly, $\pi_{13} \in \Pi_{\mathrm{TC}}(\mathbb{P}_1, \mathbb{P}_3)$. Moreover, $\tilde{c}$ satisfies the triangle inequality on $(\mathcal{X} \times \mathcal{X})^{\mathbb{N}}$ since c satisfies it on $\mathcal{X} \times \mathcal{X}$. Thus,

$$\begin{aligned}
\min_{\pi \in \Pi_{\mathrm{TC}}(\mathbb{P}_1,\mathbb{P}_3)} \int \tilde{c}\, d\pi &\le \int_{(\mathcal{X}\times\mathcal{X})^{\mathbb{N}}} \tilde{c}(\mathbf{x}_1, \mathbf{x}_3)\, d\pi_{13}(\mathbf{x}_1, \mathbf{x}_3) \\
&= \int_{(\mathcal{X}\times\mathcal{X}\times\mathcal{X})^{\mathbb{N}}} \tilde{c}(\mathbf{x}_1, \mathbf{x}_3)\, d\pi_{123}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \\
&\le \int_{(\mathcal{X}\times\mathcal{X}\times\mathcal{X})^{\mathbb{N}}} (\tilde{c}(\mathbf{x}_1, \mathbf{x}_2) + \tilde{c}(\mathbf{x}_2, \mathbf{x}_3))\, d\pi_{123}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \\
&= \int_{(\mathcal{X}\times\mathcal{X})^{\mathbb{N}}} \tilde{c}(\mathbf{x}_1, \mathbf{x}_2)\, d\pi_{12}(\mathbf{x}_1, \mathbf{x}_2) + \int_{(\mathcal{X}\times\mathcal{X})^{\mathbb{N}}} \tilde{c}(\mathbf{x}_2, \mathbf{x}_3)\, d\pi_{23}(\mathbf{x}_2, \mathbf{x}_3) \\
&= \min_{\pi \in \Pi_{\mathrm{TC}}(\mathbb{P}_1,\mathbb{P}_2)} \int \tilde{c}\, d\pi + \min_{\pi \in \Pi_{\mathrm{TC}}(\mathbb{P}_2,\mathbb{P}_3)} \int \tilde{c}\, d\pi.
\end{aligned}$$

□

A.2 Reducible Transition Coupling of Irreducible Chains

In this appendix, we provide an example showing that a transition coupling of two irreducible transition matrices is not necessarily irreducible. Let

$$P = \begin{array}{c|ccc|} & 0 & 1 & 2 \\ 0 & 0.25 & 0.25 & 0.50 \\ 1 & 0.25 & 0.25 & 0.50 \\ 2 & 0.25 & 0.25 & 0.50 \end{array} \tag{A.5}$$

and

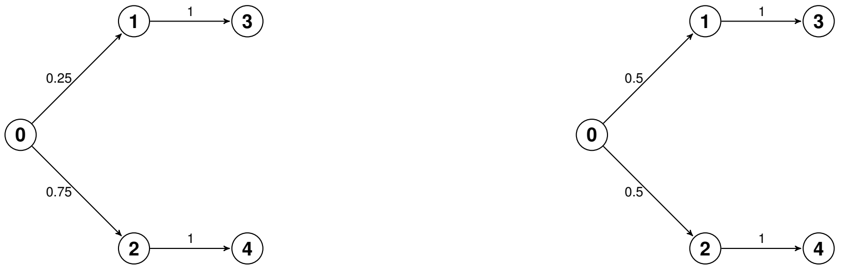
$$Q = \begin{array}{c|ccc|} & 0 & 1 & 2 \\ 0 & 0.25 & 0.25 & 0.50 \\ 1 & 0.25 & 0.25 & 0.50 \\ 2 & 0.50 & 0.25 & 0.25 \end{array}. \tag{A.6}$$

Both P and Q are clearly irreducible, but the transition coupling R, given in Figure A.1, is reducible. While we do not provide an example here, we remark that transition coupling matrices of aperiodic and irreducible transition matrices may also have multiple recurrent classes.

A.3 Comparison to 1-step Optimal Transition Coupling

In this appendix, we demonstrate how the 1-step transition coupling problem described in Section 3.2 prioritizes expected cost in the next step over long-term average cost as the OTC problem does.

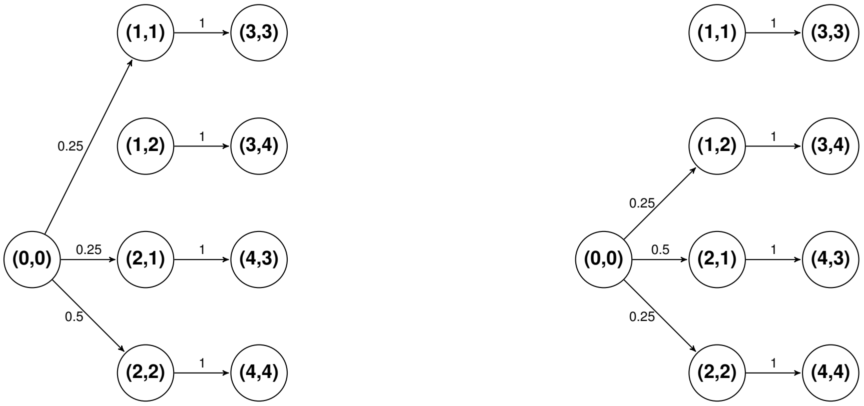
Example A.8. Consider stationary Markov chains X and Y with transition distributions defined by the graphs in Figure A.2. In order to find an OTC of X and Y, we must specify a cost for every pair of states $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Let states $(0, 0)$, $(1, 2)$, $(2, 1)$, $(2, 2)$, and $(3, 3)$ have cost 0, states $(1, 1)$, $(3, 4)$ and $(4, 3)$ have cost 1, state $(4, 4)$ have cost 9, and let all other states have a cost sufficiently large 1-step OTC and OTC do not assign them positive probability.



(a) X transition probabilities **(b)** Y transition probabilities

Figure A.2: Marginal stationary Markov chains. Both chains return to state 0 from states 3 and 4 with probability one.

The transition distributions of the OTC and 1-step OTC are largely the same except for the transitions from $(0, 0)$ to $(1, 1)$, $(1, 2)$, $(2, 1)$ and $(2, 2)$ (see Figure A.3 for an illustration). In particular, since the OTC chooses the transitions to minimize expected cost over the complete trajectory of the chain, it assigns lower probability to the transition $(0, 0) \to (2, 2)$ in order to avoid the costly state $(4, 4)$. On the other hand, the 1-step OTC does not utilize this information in deciding how to transition from $(0, 0)$ and assigns a higher probability to the transition $(0, 0) \to (2, 2)$. As a result, the expected cost of the 1-step OTC is $5/3$ compared to an expected cost of 1 for the OTC. In fact, by increasing the cost of the state $(4, 4)$, one can make the difference between the 1-step OTC and OTC costs arbitrarily large. The lower expected cost indicates that the OTC constitutes a better alignment of X and Y as compared to the 1-step OTC.



(a) 1-step OTC (expected cost of 5/3)

(b) OTC (expected cost of 1)

Figure A.3: An example where the 1-step OTC has sub-optimal expected cost. Both chains return to state $(0, 0)$ from states $(3, 3)$, $(3, 4)$, $(4, 3)$, and $(4, 4)$ with probability one. Note that Figures A.3a and A.3b omit the edges that are the same between the two transition couplings.

APPENDIX B

APPENDIX TO CHAPTER 4

B.1 Experimental Details

In this appendix, we provide further details for the experiments discussed in Section 4.5.

B.1.1 Point Cloud Alignment

In the experiment described in Section 4.5.1, point clouds in $\mathbb{R}^3$ were generated from a 4-component Gaussian mixture model as follows: For a given $\sigma_\mu > 0$, we generate $\mu_1, ..., \mu_4 \sim \mathcal{N}_3(0, \sigma_\mu^2 \mathbf{I})$ independently. The vectors $\mu_1, ..., \mu_4$ will be the means of the four mixture components under consideration. Note that as σ_μ is increased, these means will tend to be further separated and the task of distinguishing between points from each component becomes easier. Then given $n_i \in \mathbb{N}$, we generate n_i points IID from $\mathcal{N}_3(\mu_i, \mathbf{I})$ for each $i = 1, ..., 4$. Gathering these points, we obtain a point cloud $D \subset \mathbb{R}^3$ of size $N := \sum_{i=1}^4 n_i$.

The overlap regimes referred to in Figure 4.3 of Section 4.5.1 correspond to $\sigma_\mu = 1$ (high overlap), $\sigma_\mu = 2$ (moderate overlap), and $\sigma_\mu = 3$ (low overlap). In each iteration, we sampled 10 points from each mixture component to form the first graph and 5 points from each mixture component to form the second graph. For each choice of σ_μ, cross validation was performed for FGW (to select $\alpha \in \{0, 0.1, ..., 1\}$) by randomly generating 5 pairs of graphs and computing alignments of vertices and edges. Parameters that yielded the highest average alignment accuracy were selected. Separate parameters were chosen for optimizing vertex and edge alignment. The `ExactOTC` algorithm was used to compute solutions to the GraphOTC problem. The experiment was developed and run in Matlab on a 6-core, personal machine.

B.1.2 Graph Classification

In order to compute approximate solutions to the GraphOTC problem, we used the `EntropicOTC` algorithm with $L = 10$, $T = 50$, $\xi = 100$, and 50 Sinkhorn iterations. The FGW cost was computed with a default parameter choice of $\alpha = 0.5$. The experiment was developed in Matlab and run on a 24-core node in a university-owned computing cluster.

APPENDIX C

APPENDIX TO CHAPTER 5

C.1 Properties of the Proposed Estimates

Proposition C.1. *For any $\eta \geq 0$, the proposed estimates satisfy $\hat{\lambda}^{\eta,k,n} \in \mathcal{J}(\Lambda^k[\hat{\mu}_{k,n}], \Lambda^k[\hat{\nu}_{k,n}])$ and*

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

Proof. Let $\hat{\pi}_{k,n}^{\eta} \in \mathcal{M}(\mathcal{X}^k \times \mathcal{Y}^k)$ be as defined in Section 5.5. We start by proving that $\hat{\lambda}^{\eta,k,n}$ is invariant under the left-shift $\sigma \times \tau$ on $\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$. Note first that by construction, $\tilde{\Lambda}^k[\hat{\pi}_{k,n}^{\eta}] \circ (\sigma \times \tau)^{-k} = \tilde{\Lambda}^k[\hat{\pi}_{k,n}^{\eta}]$. Then,

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

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

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

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

$$\int c\, d\hat{\lambda}_1^{\eta,k,n} - \eta h(\hat{\lambda}^{\eta,k,n}) = \frac{1}{k}\int c_k\, d\hat{\pi}_k^{\eta,n} - \frac{\eta}{k} H_k(\hat{\pi}_k^{\eta,n}) = \frac{1}{k}\mathcal{T}_\eta\left(c_k; \hat{\mu}_{k,n}, \hat{\nu}_{k,n}\right) = \hat{\rho}_{k,n}^{\eta},$$

where the first equality follows from the well-known fact that randomizing the start of a block-IID process preserves entropy rate. □

C.2 Existence of an Entropic Optimal Joining

Proposition C.2. *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 $\int c\, d\lambda_1^n - \eta h(\lambda^n) \to \mathcal{S}_\eta(c; \mu, \nu)$. 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,

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

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

C.3 Properties of the (c, η)-Transform

Proposition 5.24. *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 $|c(u, v) - c(u', v')| \leq L(d_{\mathcal{U}}(u, u') + d_{\mathcal{V}}(v, v'))$ for all $u, u' \in \mathcal{U}$ and $v, v' \in \mathcal{V}$ for some $L \in \mathbb{R}$. Then for any $\eta > 0$, $f^{(c,\eta)}$ and $g^{(c,\eta)}$ satisfy $|f^{(c,\eta)}(v) - f^{(c,\eta)}(v')| \leq L d_{\mathcal{V}}(v, v')$ and $|g^{(c,\eta)}(u) - g^{(c,\eta)}(u')| \leq L d_{\mathcal{U}}(u, u')$ for all $u, u' \in \mathcal{U}$ and $v, v' \in \mathcal{V}$.*

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

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

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

C.4 Weak Convergence of Couplings and Joinings

Lemma C.3. *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 (Villani, 2008)). 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. $\square$

BIBLIOGRAPHY

E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018. URL `http://jmlr.org/papers/v18/16-480.html`.

E. Abbe and C. Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688, 2015.

D. Aldous and J. Fill. Reversible Markov chains and random walks on graphs, 2002.

D. J. Aldous and P. Diaconis. https://www.stat.berkeley.edu/ aldous/unpub/persi.pdf. 2009.

M. Allan and C. K. Williams. Harmonising chorales by probabilistic inference. *Advances in Neural Information Processing Systems*, 17:25–32, 2005.

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.

D. Alvarez-Melis and T. Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

C. Ames. The Markov process as a compositional model: A survey and tutorial. *Leonardo*, 22(2): 175–187, 1989.

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

J. Backhoff, D. Bartl, M. Beiglböck, and J. Wiesel. Estimating processes in adapted Wasserstein distance. *arXiv preprint arXiv:2002.07261*, 2020.

L. Bahl, P. Brown, P. De Souza, and R. Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 49–52. IEEE, 1986.

S. Banerjee and W. Kendall. Coupling polynomial Stratonovich integrals: the two-dimensional Brownian case. *Electronic Journal of Probability*, 23, 2018.

S. Banerjee and W. S. Kendall. Coupling the Kolmogorov diffusion: maximality and efficiency considerations. *Advances in Applied Probability*, 48(A):15–35, 2016.

S. Banerjee and W. S. Kendall. Rigidity for Markovian maximal couplings of elliptic diffusions. *Probability Theory and Related Fields*, 168(1-2):55–112, 2017.

A. Barbe, M. Sebban, P. Gonçalves, P. Borgnat, and R. Gribonval. Graph diffusion Wasserstein distances. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2020.

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

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

M. Beiglböck and W. Schachermayer. Duality for Borel measurable cost functions. *Transactions of the American Mathematical Society*, 363(8):4203–4224, 2011.

M. Beiglböck, P. Henry-Labordère, and F. Penkner. Model-independent bounds for option prices—a mass transport approach. *Finance and Stochastics*, 17(3):477–501, 2013.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

C. Bell. Algorithmic music composition using dynamic Markov chains and genetic algorithms. *Journal of Computing Sciences in Colleges*, 27(2):99–107, 2011.

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.

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

D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2): 5120–5150, 2019.

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.

J. F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems.* Springer Science & Business Media, 2013.

O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schölkopf. From optimal transport to generative modeling: the VEGAN cookbook. *stat*, 1050:22, 2017.

M. Boyle and K. Petersen. Hidden Markov processes in the context of symbolic dynamics. *arXiv preprint arXiv:0907.1858*, 2009.

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

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.

E. Cazelles, A. Robert, and F. Tobar. The Wasserstein-Fourier distance for stationary time series. *IEEE Transactions on Signal Processing*, 2020.

L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, and J. Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020.

T. Chen and S. Kiefer. On the total variation distance of labelled Markov chains. In *Proceedings of the Joint Meeting of the Twenty-Third EACSL Annual Conference on Computer Science Logic (CSL) and the Twenty-Ninth Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–10, 2014.

Y. Chen, J. Ye, and J. Li. Aggregated Wasserstein distance and state registration for hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2133–2147, 2019.

Y. Chen, Z. Lin, and H.-G. Müller. Wasserstein regression. *Journal of the American Statistical Association*, (just-accepted):1–40, 2021.

S. Cohen, G. Luise, A. Terenin, B. Amos, and M. Deisenroth. Aligning time series on incomparable spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 1036–1044. PMLR, 2021.

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.

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

M. Cuturi and M. Blondel. Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*, pages 894–903. PMLR, 2017.

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

P. Daca, T. A. Henzinger, J. Kretínskỳ, and T. Petrov. Linear distances between Markov chains. In *27th International Conference on Concurrency Theory: CONCUR 2016*, 2016.

O. Das, B. Kaneshiro, and T. Collins. Analyzing and classifying guitarists from rock guitar solo tablature. In *Proceedings of the Sound and Music Computing Conference, Limassol, Chypre*, 2018.

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

T. de la Rue. An introduction to joinings in ergodic theory. *Discrete & Continuous Dynamical Systems*, 15(1):121, 2006.

T. de la Rue. Joinings in ergodic theory, 2020.

A. K. Debnath, R. L. Lopez de Compadre, G. Debnath, A. J. Shusterman, and C. Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34 (2):786–797, 1991.

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

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.

R. Dekker. Counter examples for compact action Markov decision chains with average reward criteria. *Stochastic Models*, 3(3):357–368, 1987.

E. Del Barrio, J.-M. Loubes, et al. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019.

P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, and R. Singh. Gromov-Wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, 2020.

Y. Dong and W. Sawin. Copt: Coordinated optimal transport on graphs. *Advances in Neural Information Processing Systems*, 33, 2020.

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

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, 2018.

M. 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.

M. Ellis. Distances between two-state Markov processes attainable by Markov joinings. *Transactions of the American Mathematical Society*, 241:129–153, 1978.

M. Ellis. Conditions for attaining $\bar{d}$ by a Markovian joining. *The Annals of Probability*, 8(3): 431–440, 1980a.

M. Ellis. On Kamae's conjecture concerning the $\overline{d}$-distance between two-state Markov processes. *The Annals of Probability*, pages 372–376, 1980b.

A. Forrow, J.-C. Hütter, M. Nitzan, P. Rigollet, G. Schiebinger, and J. Weed. Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465. PMLR, 2019.

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.

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

H. Furstenberg. Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Mathematical systems theory*, 1(1):1–49, 1967.

C. Furusawa and K. Kaneko. A dynamical-systems view of stem cell biology. *Science*, 338(6104): 215–217, 2012.

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.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

E. Glasner. *Ergodic theory via joinings*. Number 101. American Mathematical Soc., 2003.

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

D. S. Griffeath. *Coupling methods for Markov processes*. Cornell University, January, 1976.

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.

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.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

R. Howard. Dynamic programming and Markov processes. 1960.

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

H. Janati, M. Cuturi, and A. Gramfort. Wasserstein regularization for sparse multi-task regression. In *AISTATS 2019-22nd International Conference on Artificial Intelligence and Statistics*, volume 89, 2019.

H. Janati, M. Cuturi, and A. Gramfort. Spatio-temporal alignments: Optimal transport through space and time. In *International Conference on Artificial Intelligence and Statistics*, pages 1695–1704. PMLR, 2020.

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

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

L. V. Kantorovich. On a problem of Monge. *Journal of Mathematical Sciences*, 133(4):1383–1383, 2006a.

L. V. Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4): 1381–1382, 2006b.

K. Kersting, N. M. Kriege, C. Morris, P. Mutzel, and M. Neumann. Benchmark data sets for graph kernels, 2016. http://graphkernels.cs.tu-dortmund.de.

S. Kiefer. On computing the total variation distance of hidden Markov models. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

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.

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.

N. M. Kriege, M. Fey, D. Fisseler, P. Mutzel, and F. Weichert. Recognizing Cuneiform signs using graph based methods. In *International Workshop on Cost-Sensitive Learning*, pages 31–44. PMLR, 2018.

C. Laclau, I. Redko, B. Matei, Y. Bennani, and V. Brault. Co-clustering through optimal transport. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1955–1964, 2017.

C. Lee and D. Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4:1–50, 2019.

D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

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.

T. Lindvall. *Lectures on the coupling method.* Courier Corporation, 2002.

Y.-W. Liu and E. Selfridge-Field. Modeling music as Markov chains: Composer identification, 2002.

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.

A. O. Lopes and J. K. Mengue. Duality theorems in ergodic transport. *Journal of Statistical Physics*, 149(5):921–942, 2012.

L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993.

J. Malka, R. Flamary, and N. Courty. Gromov-Wasserstein optimal transport for heterogeneous domain adaptation.

H. P. Maretic, M. E. Gheche, G. Chierchia, and P. Frossard. Got: An optimal transport framework for graph comparison. *Advances in Neural Information Processing Systems 32*, 32(CONF), 2019.

H. P. Maretic, M. E. Gheche, M. Minder, G. Chierchia, and P. Frossard. Wasserstein-based graph alignment. *arXiv preprint arXiv:2003.06048*, 2020.

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

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

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

K. McGoff, S. Mukherjee, and A. Nobel. Gibbs posterior convergence and the thermodynamic formalism. *Annals of Applied Probability*, 2021.

A. Medio and G. Gallo. *Chaotic dynamics: Theory and applications to economics.* Cambridge University Press, 1995.

F. Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 4541–4551, 2019.

L. Mi, W. Zhang, X. Gu, and Y. Wang. Variational Wasserstein clustering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 322–337, 2018.

A. Moameni. Invariance properties of the Monge-Kantorovich mass transport problem. *Discrete & Continuous Dynamical Systems*, 36(5):2653, 2016.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie royale des sciences de Paris*, 1781.

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

N. Moriel, E. Senel, N. Friedman, N. Rajewsky, N. Karaiskos, and M. Nitzan. NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*, pages 1–24, 2021.

V. Moulos. Bicausal optimal transport for Markov chains via dynamic programming. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1688–1693. IEEE, 2021.

C. Mufa. Optimal Markovian couplings and applications. *Acta Mathematica Sinica*, 10(3):260–275, 1994.

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.

K. O'Connor, K. McGoff, and A. B. Nobel. Estimation of stationary optimal transport plans. *arXiv preprint arXiv:2107.11858*, 2021a.

K. O'Connor, K. McGoff, and A. B. Nobel. Optimal transport for stationary Markov chains via policy iteration. *To appear in the Journal of Machine Learning Research*, 2021b.

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*, 2021c.

N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96(2):163–180, 2004.

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

D. S. Ornstein and P. C. Shields. The d-recognition of processes. *Advances in Mathematics*, 104 (2):182–224, 1994.

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

N. Papadakis. *Optimal transport for image processing.* PhD thesis, Université de Bordeaux; Habilitation thesis, 2015.

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

A. Pikrakis, S. Theodoridis, and D. Kamarotos. Classification of musical patterns using variable duration hidden Markov models. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1795–1807, 2006.

M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons Inc., 2005.

J. Rabin and N. Papadakis. Convex color image segmentation with optimal transport distances. In *International conference on scale space and variational methods in computer vision*, pages 256–269. Springer, 2015.

J. Rabin, S. Ferradans, and N. Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4852–4856. IEEE, 2014.

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems: Applications.* Springer Science & Business Media, 2006.

L. Ren, D. Dunson, S. Lindroth, and L. Carin. Dynamic nonparametric Bayesian models for analysis of music. *Journal of the American Statistical Association*, 105(490):458–472, 2010.

K. Riesen and H. Bunke. Iam graph database repository for graph based pattern recognition and machine learning. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 287–297. Springer, 2008.

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

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

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

T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63, 2015.

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.

P. Schweitzer. On undiscounted Markovian decision processes with compact action spaces. *RAIRO-Operations Research*, 19(1):71–86, 1985.

C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3): 379–423, 1948.

P. C. Shields. *The ergodic theory of discrete sample paths*, volume 13. American Mathematical Soc., 1996.

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

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.

J. Song, Y. Gao, H. Wang, and B. An. Measuring the distance between finite Markov decision processes. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pages 468–476. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

B. Su and G. Hua. Order-preserving Wasserstein distance for sequence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1057, 2017.

B. Su and G. Hua. Order-preserving optimal transport for distances between sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2961–2974, 2018.

J. J. Sutherland, L. A. O'brien, and D. F. Weaver. Spline-fitting with a genetic algorithm: A method for developing classification structure- activity relationships. *Journal of chemical information and computer sciences*, 43(6):1906–1915, 2003.

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

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

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.

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.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR 2018)*. OpenReview. net, 2018.

A. Tong, J. Huang, G. Wolf, D. Van Dijk, and S. Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International Conference on Machine Learning*, pages 9526–9536. PMLR, 2020.

V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, 1958.

A. Varga and R. Moore. Hidden Markov model decomposition of speech and noise. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 845–848. IEEE, 1990.

T. Vayer, R. Flamary, R. Tavenard, L. Chapel, and N. Courty. Sliced Gromov-Wasserstein. In *NeurIPS 2019-Thirty-third Conference on Neural Information Processing Systems*, volume 32, 2019.

T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. Fused Gromov-Wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

P. Walters. *An introduction to ergodic theory*, volume 79. Springer Science & Business Media, 2000.

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.

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.

M. Weiland, A. Smaill, and P. Nelson. Learning musical pitch structures with hierarchical hidden Markov models. *Journees d'Informatique Musical*, 2005.

J. P. Williams, C. B. Storlie, T. M. Therneau, C. R. J. Jr, and J. Hannig. A Bayesian approach to multistate hidden Markov models: application to dementia progression. *Journal of the American Statistical Association*, 115(529):16–31, 2020.

H. Xu, W. Wang, W. Liu, and L. Carin. Distilled Wasserstein learning for word embedding and topic modeling. In *Advances in Neural Information Processing Systems*, pages 1716–1725, 2018.

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, 2019a.

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, 2019b.

J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A hidden Markov model approach to text segmentation and event tracking. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 1, pages 333–336. IEEE, 1998.

X. Yan. `https://sites.cs.ucsb.edu/~xyan/dataset.htm`.

Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, volume 7, pages 2969–2975, 2018.

A. K. Yanchenko and S. Mukherjee. Classical music composition using state space models. *arXiv preprint arXiv:1708.03822*, 2017.

K. D. Yang, K. Damodaran, S. Venkatachalapathy, A. C. Soylemezoglu, G. Shivashankar, and C. Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828, 2020.

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

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

S. Zhang. Existence of the optimal measurable coupling and ergodicity for Markov processes. *Science in China Series A: Mathematics*, 42(1):58–67, 1999.

S. Zhang. Existence and application of optimal Markovian coupling with respect to non-negative lower semi-continuous functions. *Acta Mathematica Sinica*, 16(2):261–270, 2000.

W.-B. Zhang. *Discrete dynamical systems, bifurcations and chaos in economics*. elsevier, 2006.

X.-Q. Zhao. *Dynamical systems in population biology*, volume 16. Springer, 2003.

W. Zucchini, I. L. MacDonald, and R. Langrock. *Hidden Markov models for time series: an introduction using R*. CRC press, 2017.