Metrics on the Space of Probability Distributions on Countable Products of Polish Spaces

Jacob Armstrong Goodall

Supervisor: Professor Robert MacKay

Contents

1 Context 5

1.1 Interacting Particle Systems (IPS) 5

1.1.1 Origins, Motivations and Informal Definitions. 5

1.1.2 Probabilistic Cellular Automata (PCA) 6

1.1.3 Emergence and Complexity 8

2 Mathematical Setting & Key Result 9

2.1 Mathematical Description of PCA 10

2.2 Two Metrics Between Distributions of PCA. 12

3 Example of Explicit Calculation 16

3.1 Setup 16

3.2 Dobrushin's Metric (MacKay and Diakonova,2011) 16

4 Kantorovich Duality 17

5 Proof of duality. 22

5.1 Definitions and Technical Requirements. 22

5.2 Developing Duality with an Informal Analogy 26

5.3 Statement of the Duality 32

5.4 Idea for Proof: 34

5.5 Rigorous Proof. 36

5.6 Some Equivalent Statements About Optimal Transmission Methods. . . 45

6 Conclusion 50

7 Bibliographic Notes 50

8 Appendix **52**

8.1 $\bar{d}(\mu, \nu)$ is a complete metric on $\mathcal{P}(\mathcal{X})$. 52

Preface

The goal of this dissertation was to prove the equality of two metrics introduced to quantify distances between probability distributions defined on the countable product of compact Polish spaces. All together there are eight sections.

In the first section I provide context for the rest of dissertation, this includes background information describing the motivation for undertaking this project.

In the second section I introduce preliminary mathematical definitions, including what we call Steif's and Dobrushin's metrics, and state informally the main result. The third section is an explicit example of the calculation of distance between two stationary distributions using Dobrushin's metric.

The fourth section states the Kantorovich Duality, which is a generalization of the Kantorovich Rubinstein theorem. The proof of this duality is the basis for the main result in this dissertation.

Section five develops the "Dobrushin-Steif" duality with an informal analogy from information theory. Then the result is proven, following closely (but extending upon where necessary) the proof of the Kantorovich Rubinstein Theorem.

The bibliography details my use of the references, explaining which reference was used for each result.

Finally, there is an appendix where it is shown that Steif's metric is a complete metric on $\mathcal{P}(\mathcal{X})$.

I would like to thank my advisor Professor Robert MacKay for giving me the chance to attempt this problem as well as his encouragement and feedback. I should also thank Dr Marie-Therese Wolfram for giving me the opportunity to explain some of my ideas.

1 Context

In this section I will briefly and informally introduce some ideas that provide context for the main result.

1.1 Interacting Particle Systems (IPS)

1.1.1 Origins, Motivations and Informal Definitions.

The precursors to the study of IPS can be traced back to the works of Russian cyberneticist M.L. Tsetlin in the 1960's when he studied cellular automata[1] in random media. Along with Tsetlin, I.M. Gelfand wrote about the importance of new mathematical techniques to model collective behavior. This in turn inspired R.L. Dobrushin who was a statistical mechanic whith an interest in information theory and signal propagation in uncertain media. At around the same time (Late 1960's) in the USA, probablist F. Spitzer also began studying interacting Markov processes which eventually became the theory of IPS. Dobrushin and Spitzer can well be regarded as the founding fathers of this branch of probability.

Informally, Interacting particle systems are Markov processes (stochastic process with independent increments), in continuous or discrete time, which describe 'particles' moving in some underlying discrete space, subject to some random noise and interactions. Foundational results such as existence and uniqueness were quickly laid down by the mid-seventies, but it turned out to be extraordinarily hard to analyse many properties of these systems and as a result, many problems that are simple to state resisted attempts at solution. It seems that early motivations for studying IPS were the modeling of biological phenomena (e.g neurons) and statistical mechanics (e.g crystalline substances). However since then, similar problems have arisen naturally throughout

[1] Historically cellular automata came first but mathematically they fit into the framework of IPS so are defined later in the text.

the sciences such as the behavior of financial markets, computer networks and traffic flows - the abstract framework seems to have promoted it's broadening application and the discrete time version (probabilistic cellular automata or PCA) turned out to be important computational tools. IPS, despite their seeming simplicity, are incredibly good at capturing subtle aspects of phenomena and often making the model more realistic does little to change the overall behavior of the model. But how do we investigate the behavior of such systems?

1. Limiting Behavior: Existence and uniqueness of invariant measures (limit points as time extends indefinitely);

2. Attractor Basins: second is whether and how fast these invariant distributions attract other initial probability measures - we can classify initial distributions according to their attractors;

3. Dependence of answers to the above quetions on specific parameters: Qualitative changes in large scale macroscopic behavior depending on the system parameters are often known as phase transitions, and are of particular interest;

4. Classification: Universality classes (collections of systems that share asymptotic behavior and critical exponent[2]) are part of a pervasive philosophy[3] in the study of IPS.

1.1.2 Probabilistic Cellular Automata (PCA)

Probabilistic cellular automata are the discrete time version of interacting particle systems. PCA are particularly important to the present discussion as the study of such systems is the precise context that gives rise to the importance of the main result.

[2] Critical exponents describe the behavior of physical quantities near phase transitions.

[3] It is generally believed but not rigorously proven that such universality classes exist.

Cellular Automata (CA) are lattices of interconnected countable-state sites or cells which evolve synchronously in discrete time steps according to deterministic rules involving the states of adjacent automata. They are an interesting example of a dynamical system with simple rules that can display rich and complex long term behavior.

PCA are the stochastic extension of such entities where each site is updated according to the probability distribution defined by the values of its neighbors. Mathematically they are interacting Markov chains evolving in a parallel but locally coupled fashion[4]. PCA are defined on lattices, each vertex representing a site. The topology of the lattice determines the viable links between vertices which can be interpreted as a route of communication between sites. The influence of communication between sites decays as the distance (depending on the topology of the space) increases - hence we consider a neighborhood of dependence between sites.

One of the things of greatest interest in the study of PCA are phase transitions, these occur as you change the macroscopic variables of a system and manifest as abrupt changes in the properties of the system. Physical examples are changes from liquid to gas or solid. There are two classifications of phase transitions, first and second-order, of which we will discuss only the former. First order phase transitions have coexistence curves - i.e more than one possible limiting measures. For instance if the right combination of pressure and temperature is met then water can exist as either liquid or ice depending on how those conditions where reached.

Metastability can be thought of as a barrier delaying convergence, an apparent state of equilibrium in contrast to the expected limiting behavior of the system under the given parameters. A system exhibits metastability in the vicinity of first order phase transitions. A metastable state is one that a system may occupy for a large amount of time but is different from the true equilibrium of the system. For instance it was

[4] See definition of coupling in subsection 2.2.

mentioned before that if the correct combination of pressure and cooling is applied to water in a very smooth fashion then it may continue as a liquid even below zero, however it cannot do this indefinitely - its only true equilibrium state under these conditions is as a solid. So metastability can be described as the persistence of a system in a state other than the one described by an invariant measure with respect to the parameters of the system. Normally the presence of external noise or internal fluctuations cause the system to jump into the equilibrium state after some time has passed. Metastability relates to some attempts at quantifying emergence, for instance in (MacKay & Diakonova, 2011).

1.1.3 Emergence and Complexity

Emergence as a concept is the tendency of large interacting networks of participants, whose individual behaviour may be deterministic or non-deterministic, to produce macro-scale or approximately network wide synergy. That is, the system displays emergence if the interactions of its participants produce behaviour not reducible to the rules governing individual participants. In more mathematical terms one may say that this kind of emergence (which we call weak) is quantified as the distance of the probability distribution describing the entire system from the product of the probability distributions of the participants. Non-trivial emergence of collective behaviour can also manifest as phase transitions and produce metastable states. Existence of the metastable states we refer to as strong emergence and can be quantified as the diameter of the set of space-time phases. One may seek to find cases of metastable states that are the result of systematic behavior not reducible to the rules governing individual actions - i.e emergence that is both strong and weak.

Broadly speaking a complex system is one that displays self organization and emergence in the absence of global control. Swarms of insects, nervous systems of animals and financial markets are all examples and yet, despite their ubiquity in nature, there is so far not an adequate mathematical framework uniting these phenomena.

The main advantage of introducing the metrics described in this project (and previously in (MacKay, 2011) and (MacKay, 2018)) is that they act appropriately as a quantification of distance on large finite or even infinite IPS with large neighbourhoods of dependence. While other metrics on such distributions, for example absolute variation, fail to converge (see section 2 in (MacKay, 2011) these metrics capture the intuitive notions we have about the closeness of such distributions. This allows us to study the behaviour of complex systems, testing for properties like ergodicity, quantify the concept of emergence and hence aid in classifying complex dynamical systems.

2 Mathematical Setting & Key Result

In this section, after introducing the necessary definitions, I formulate PCA more precisely. After this formulation there is a description of the NEC voter model PCA, which is the subject of the calculation in section 3. Finally I state the definitions of Steif's and Dobrushin's Metrics and briefly explain what I plan to do in the main section 5.

Definition 1 (Probability Space)**.** A probability space is a triple $(\Omega, \mathcal{F}, P)$, where Ω is a sample space, $\mathcal{F}$ denotes a $\sigma-$field consisting of subsets of Ω and $P : \mathcal{F} \to [0, 1]$ is a probability measure. The sample space, Ω can be thought of as a set of results corresponding to all possible outcomes of a random experiment. The $\sigma-$field $\mathcal{F}$ is called the event space. $P : \mathcal{F} \to [0, 1]$ is a countably additive function, i.e. if $\{A_i\}_{i\in\mathbb{N}} \subset \mathcal{F}$ is a countable collection of pairwise disjoint sets, then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

P assigns to each event a probability between 0 and 1 such that $P(\Omega) = \int_\Omega dP = 1$.

Definition 2 (Markov Process)**.** The sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is called a Discrete time Markov chain if it has the *Markovian property*

$$P(X_{n+1} = j|X_0 = i_0, X_1 = i_1, ..., X_n = i) = P(X_{n+1} = j|X_n = i) = P_{ij}.$$

Where n represents the present step, and $n+1$ represents the step following step n. The probability P_{ij} can be interpreted as the probability that the Markov chain will be in state j at step $n+1$ given that is is in state i at step n.

Then next definition is required to understand how PCA evolve in discrete time, which is what will be explained in the next subsection.

Definition 3 (Markov Transition Probability Matrix)**.** For a discrete time Markov chain with statespace $S = \{1, ..., K\}$,

$$P = \begin{bmatrix} P_{00} & \cdots & P_{0K} \\ \vdots & P_{ij} & \vdots \\ P_{K0} & \cdots & P_{KK} \end{bmatrix}$$

is the one-step transition probability matrix. Here P_{ij} is the one-step transition probability for transition from state i to state j, and $\sum_{j\in S} P_{ij} = 1 \; \forall i \in S$.

2.1 Mathematical Description of PCA

A graph $G = (V, E)$ consists of two sets V and E. The elements of V are called the vertices and the elements of E the edges of G. Each edge corresponds to a pair of vertices which we say are adjacent. If two vertices s and t are adjacent we write

$s \sim t$. Consider the graph $G = (S, E)$ in which the set of vertices S represents the locations of the automaton (sites) and is a countable (but potentially infinite) set. The edge set E defines the topology of the space and can be thought of as representing communication channels between the sites. Denote by $\mathcal{A}$ the "universal state-space", this is the set of values from which we draw the state-space at each site. The local state-space (state-space of $s \in S$) X_s is a subset of $\mathcal{A}$, the state-space may be different at each site. Now define the configuration space as $\mathcal{X} = \prod_s X_s$, this represents the set of all possible configurations and has the product topology. We will denote configurations by $x = (x_s)_{s \in S}$ or $y = (y_s)_{s \in S}$. Each site has a neighbourhood of interactions defined by E and we denote this by T_s where $T_s = \{t \in S : t \sim s\}$.

The updating rule is defined by a *Markov* transition probability kernel in discrete time. If S and X_s are finite then the kernel can be written as $P(y|x)$, interpreted as the probability that the configuration at time $t+1$ is y given that the configuration at time t is x. A PCA then corresponds to the transition of the form

$$P(y|x) = \prod_{s \in S} p_s(y_s | x_{T_s})$$

where $\{p_s(\cdot|x_{T_s}),\ s \in S,\ x_{T_s} \in X_s^T\}$ is a family of probability distributions on X_s. This product corresponds to a family of Markov processes, one at each site.

Example 4 (The North-East-Center Majority Voter PCA)**.** The state space at each site $s \in S$ is $X_s = \{0, 1\}$. At each time step each site computes the majority state over its north and east neighbours and itself (this is called the "NEC neighbourhood" and corresponds to the set T_s in the above construction). Next with independent probability (corresponding to p_s above) the site is updated to the majority state with probability $1 - \lambda$ and to the opposite state with probability λ. We refer to λ as the error rate.

Many results have been proven rigorously about the PCA in the last example.

Firstly, if λ is near enough to $\frac{1}{2}$ there is a unique space time phase, in contrast, if $\lambda > 0$ is small enough there are at least two. The latter is a consequence of work by Toom that for λ small enough there is a function $c(\lambda) < \frac{1}{2}$ and at least two stationary probability distributions, for one of which the probability of any given site being 1 is $c(\lambda)$ and for the other it is $1 - c(\lambda)$. Each generates a space-time phase by time evolution. For a lengthier discussion see section 2 of (MacKay and Diakonova,2011).

2.2 Two Metrics Between Distributions of PCA.

Definition 5 (Complete Metric)**.** A complete metric is a metric in which every Cauchy sequence is convergent. A topological space with a complete metric compatible with the topology is called a complete metric space.

Definition 6 (Separable Metric Space)**.** A topological space is called separable if it contains a countable, dense subset; that is, there exists a sequence $\{x_n\}_{n\in\mathbb{N}}$ of elements of the space such that every non-empty open subset of the space contains at least one element of the sequence.

Definition 7 (Coupling)**.** Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces. Coupling μ and ν means constructing two *random variables* X and Y on some probability space $(\Omega, \mathbb{P})$, such that $\mathrm{Law}(X) = \mu$ and $\mathrm{Law}(Y) = \nu$. The couple (X, Y) is called a coupling of (μ, ν). By abuse of language, the law of (X, Y) is also called a coupling of (μ, ν).

Remark 8. Another way of rephrasing this definition is to define a coupling of μ and ν as a measure π on $\mathcal{X} \times \mathcal{Y}$ such that $(\mathrm{proj}_{\mathcal{X}})_{\#}\pi = \mu$, and $(\mathrm{proj}_{\mathcal{Y}})_{\#}\pi = \nu$, where $\mathrm{proj}_{\mathcal{X}}$ and $\mathrm{proj}_{\mathcal{Y}}$ respectively stand for the projection maps $(x, y) \mapsto x$ and $(x, y) \mapsto y$. It is also equivalent to say that for all integrable non-negative measurable functions ψ and ϕ on $\mathcal{X}$, $\mathcal{Y}$,

$$\int_{\mathcal{X}\times\mathcal{Y}} (\phi(x) + \psi(y)) d\pi(x, y) = \int_{\mathcal{X}} \phi d\mu + \int_{\mathcal{Y}} \psi d\nu.$$

We will refer to this as the *marginal condition.*

Example 9 (Optimal Transport)**.** Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be the cost function. The value $c(x, y)$ can be interpreted as the cost requird to transport a unit of mass from the point x to the point y. Then the Monge-Kantorovich minimization problem is

$$\inf \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y),$$

where the infimum runs over all joint probability measures π on $\mathcal{X}\times\mathcal{Y}$ with marginals μ and ν. The measures π are referred to a transport plans, and those achieving the minimum value are referred to as optimal transference plans.

This coupling leads to the definition of the Wasserstein distance .

Definition 10 (Wasserstein Distance)**.** Let $(\mathcal{X}, d)$ be a Polish metric space, and let $p \in [1, \infty)$. For any two probability measures μ, ν on $\mathcal{X}$, the Wasserstein distance of order p between μ and ν is defined by the formula

$$W_p(\mu, \nu) = \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}} d(x, y)^p d\pi(x, y) \right)^{\frac{1}{p}}$$

When $p = 1$ this is called the Kantorovich-Rubinstein distance and had the dual representation

$$W_1(\mu, \nu) = \sup_{||\psi||_{\text{Lip}}\leq 1} \left\{ \int_{\mathcal{X}} \psi d\mu - \int_{\mathcal{X}} \psi d\nu \right\}.$$

Definition 11 (Polish Space)**.** A Polish space is a complete, separable metric space, equipped with its Borel $\sigma-$algebra denoted by $\mathcal{M}$.

Remark 12. If $\{(X_s, \mathcal{M}_s)\}_{s\in S}$ is a family of measurable spaces and E_s is any element of $\mathcal{M}_s$, the product $\sigma-$algebra on $\mathcal{X} = \prod_{s\in S} X_s$ is the smallest $\sigma-$algebra on $\mathcal{X}$ that makes all the projection maps $\mu_s : \mathcal{X} \to X_s$ measurable, that is, the $\sigma-$algebra generated by the sets $\mu_s^{-1}(E_s)$ as E_s ranges over $\mathcal{M}_s$ and s ranges over S. It is denoted by $\otimes_{s\in S}\mathcal{M}_s$. This is analogous to the product topology on a product of topological spaces.

Say we want to construct a metric on a probabilistic cellular automata. Consider first the collection of state spaces $(X_s)_{s\in S}$ where each (X_s, d_s) is a Polish probability space[5] with bounded diameter. Next, define $\mathcal{X} = \prod_{s\in S} X_s$, introduce μ and ν as probability distributions on $\mathcal{X}$, and the set of couplings π of μ and ν. We will write $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and $\pi \in \Pi(\mu, \nu)$ respectively. Finally note that d_s extends trivially to a symmetric non-negative function $d_s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by defining for $(x, y) \in \mathcal{X} \times \mathcal{X}$, $d_s(x, y) = d_s(x_s, y_s)$ - here x_s represents the s^{th} component of the vector x. Hence we have the following definitions.

Definition 13 (Steif's Metric)**.**

$$\bar{\mathrm{d}}(\mu,\, \nu) = \inf_{\pi\in\Pi(\mu,\,\nu)} \sup_{s\in S} \int d_s(x_s, y_s)\mathrm{d}\pi(x, y)$$

Remark 14. This is an extension of a metric that was introduced by J. Steif in (Steif,1988). Steif considered the case where the statespace was $\{0, 1\}$ and d_s the indicator metric. In the appendix I prove that this is a complete metric on $\mathcal{P}(\mathcal{X})$.

Definition 15 (Dobrushin Metric)**.** Let F be the space of continuous functions $f : \mathcal{X} \to \mathbb{R}$ such that

$$||f||_F = \sum_{s\in S} \Delta_s(f) < \infty,$$

where

$$\Delta_s(f) = \sup\left\{\frac{f(x) - f(y)}{d_s(x_s, y_s)} : x_t = y_t\ \forall t \neq s, x_s \neq y_s\right\}.$$

[5] The term "Polish probability space" comes from Villani. The interpretation is that μ and ν are Borel measures on the Polish space which, in Villani's view, is implicitly equipped with it's $\sigma-$algebra.

Then we define Dobrushin's metric as,

$$D(\mu,\nu) = \sup_{f\in F\setminus C} \frac{(\mu-\nu)(f)}{\sum_{s\in S}\Delta_s(f)}$$

Here C denotes the constant functions, $\mu(f) = \int f \mathrm{d}\mu$.

The next proposition is a useful reformulation of Dobrushin's metric that will be used to prove that $D(\mu,\nu) = \bar{d}(\mu,\nu)$.

Proposition 16. *For* $\mathcal{E} = \big\{e = (e_s)_{s\in S} : e_s > 0,\ \sum_{s\in S} e_s \leq 1\big\}$,

$$\begin{aligned} D(\mu,\nu) &= \sup_{f\in F\setminus C} \frac{(\mu-\nu)(f)}{\sum_{s\in S}\Delta_s(f)} \\ &= \sup_{f\in F\setminus C,\, e\in\mathcal{E}} \int (d\mu - d\nu)(x) f(x) \textbf{\textit{ subject to }} f(x)-f(y) \leq e_s d_s(x_s,y_s), x_t = y_t \forall t\neq s \\ &= \sup_{f\in F\setminus C,\, e\in\mathcal{E}} \int (d\mu - d\nu)(x) f(x) \textbf{\textit{ subject to }} f(x)-f(y) \leq \sum_{s\in S} e_s d_s(x_s,y_s) \end{aligned}$$

Proof. Define

$$\sup\left\{\frac{f(x)-f(y)}{d_s(x,y)} : x_t = y_t \quad \forall t\neq s,\ x_s\neq y_s\right\} := e_s,$$

then $\frac{f(x)-f(y)}{d_s(x,y)} \leq e_s$ for all $(x,y)\in\mathcal{X}\times\mathcal{X}$ such that $x_t = y_t \quad \forall t\neq s,\ x_s\neq y_s$. Obviously we can rewrite $D(\mu,\nu)$ as

$$\begin{aligned} &\sup_{f\in F\setminus C} \{(\mu-\nu)(f) : ||f||_F \leq 1\} \\ &\qquad = \sup_{f\in F\setminus C}\sup_{e\in\mathcal{E}}\left\{(\mu-\nu)(f),\ \frac{f(x)-f(y)}{d_s(x,y)} \leq e_s,\ x_t = y_t\ \forall t\neq s,\ x_s\neq y_s\right\}. \quad (2.1) \end{aligned}$$

It can be shown (see Lemma 5.3 on page 28) that the conditions in the second line can be rewritten as

$$e_s \geq 0,\ \sum_{s\in S} \leq 1,\ f(x) - f(y) \leq \sum_{s\in S} e_s d_s(x,y).$$

□

The main goal of this project is to show that these two metrics are in fact identical by proving that the following duality exists.

$$\inf_{\pi\in\Pi(\mu,\nu)} \sup_{s\in S} \int d_s(x_s,y_s)\mathrm{d}\pi(x,y) \overset{\text{Duality}}{=} \sup_{f\in F\setminus C} \frac{(\mu-\nu)(f)}{\sum_{s\in S}\Delta_s(f)}$$

In the case of the Wasserstein distance this property is called the Kantorovich Duality. The dual interpretation has the benefit of allowing the user to pass back and forth between the two - a property that has proven of great technical convenience in the case of the Wasserstein distance.

3 Example of Explicit Calculation

3.1 Setup

The aim in this section is to calculate explicitly the distance between two stationary distributions of the NEC majority voter PCA (see example 4 on page 11) in terms of Dobrushin's metric. We consider the initial distributions $\delta_0\left(x_s=1\right)=0$ and $\mu_0\left(x_s=1\right)=c(\lambda)$.

3.2 Dobrushin's Metric (MacKay and Diakonova,2011)

We will prove that

$$D\left(\mu_0,\delta_0\right)=c(\lambda).$$

First by considering $f(x) = x_{00}$ the value of the state at site $(0,0)$ then $\mu_0(f) - \delta_0(f) = c(\lambda)$. We also have $||f|| = \sum_{s\in S} \Delta_s(f) = \sum_{s\in S} \sup_{x=y \text{ off } s} \frac{f(x)-f(y)}{d_s(x_s,y_s)} = 1$. So $D(\mu_0, \delta_0) \geq c(\lambda)$. Next we bound it above by the same. Consider all functions f with support on a finite subset of space I. I is countable so we can consider some arbitrary order imposed on the elements. Without loss of generallity we can let $f(x) = 0$ at the all zero state, then characterize $f(x)$ as the sum of changes from the all zero state. This is bounded above by $\sum_{s:x=1} \Delta_s(f)$. So

$$\begin{aligned}\mu_0(f) - \delta_0(f) = \sum_{s \text{ on } I} \mu_0(x)f(x) \leq \sum_{x \text{ on } I} \mu_0(x) \sum_{s:x_s=0} \Delta_s(f) \\ = \sum_{s\in I} \Delta_s(f) \sum_{x \text{ on } I:x_s=1} \mu_0(x) = \sum_s \Delta_s(f)c(\lambda) \leq c(\lambda)||f||\end{aligned}$$

4 Kantorovich Duality

In this section we will state the Kantorovich Duality. A proof can be found in (Villani,2009).

Definition 17 (Concentration of Measure)**.** If μ is a Borel measure on a topological space X, a set N is said to be $\mu-$negligible if N is included in a Borel set of zero $\mu-$measure. Then μ is said to be concentrated on a set C if $X\backslash C$ is negligible. (If C itself is Borel measurable, this is of course equivalent to $\mu[X\backslash C] = 0$.) By abuse of language, I may say that X has full $\mu-$measure if μ is concentrated on X .

Definition 18 (Upper Semi-continuity)**.** We say that a function f is upper semi-continuous at x_0 if for every $y > f(x_0)$ there exists a neighbourhood $U$of x_0 such that $f(x) < y$ for all $x \in U$. For the particular case of a metric space, this can be expressed as

$$\limsup_{x\to x_0} f(x) \leq f(x_0).$$

The function f is called upper semi-continuous if it is upper semi-continuous at every point of its domain.

Definition 19 (Lower Semi-continuity)**.** We say that a function f is lower semi continuous at x_0 if for every $y < f(x_0)$ there exists a neighbourhood $U$of x_0 such that $f(x) > y$ for all $x \in U$. For the particular case of a metric space, this can be expressed as

$$\liminf_{x \to x_0} f(x) \geq f(x_0).$$

The function f is called lower semi-continuous if it is lower semi-continuous at every point of its domain.

Definition 20 (Cyclical Monotonicity)**.** Let $\mathcal{X}, \mathcal{Y}$ be arbitrary sets, and $c : \mathcal{X} \times \mathcal{Y} \to (-\infty, +\infty]$ be a function. A subset $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ is said to be $c-$cyclically monotone if, for any $N \in \mathbb{N}$, and any family $(x_1, y_1), ..., (x_N, y_N)$ of points in Γ, holds the inequality

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

(with the convention that $y_{N+1} = y_1$). A transference plan is said to be $c-$cyclically monotone if it is concentrated on a $c-$cyclically monotone set.

Definition 21 (c-convexity)**.** Let $\mathcal{X}, \mathcal{Y}$ be sets, and $c : \mathcal{X} \times \mathcal{Y} \to (-\infty, +\infty]$. A function $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be $c-$convex if it is not identically $+\infty$, and there exists $\zeta : \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ such that

$$\forall x \in \mathcal{X} \qquad \psi(x) = \sup_{y \in \mathcal{Y}} \left(\zeta(y) - c(x, y)\right).$$

Then its $c-$transform is the function ψ^c defined by

$$\forall y \in \mathcal{Y} \qquad \psi^c(y) \inf_{x \in \mathcal{X}} \left(\psi(x) + c(x, y)\right),$$

and its $c-$sub-differential is the $c-$cyclically monotone set defined by

$$\partial_c \psi := \left\{ (x, y) \in \mathcal{X} \times \mathcal{Y}; \quad \psi^c(y) - \psi(x) = c(x, y) \right\}.$$

The functions ψ and ψ^c are said to be $c-$conjugate.

Moreover, the $c-$sub-differential of ψ at point x is

$$\partial_c \psi(x) = \left\{ y \in \mathcal{Y}; (x, y) \in \partial_c \psi \right\},$$

or equivalently

$$\forall z \in \mathcal{X}, \qquad \psi(x) + c(x, y) \leq \psi(z) + c(z, y).$$

Definition 22 (c-concavity)**.** With the same notation as in the previous definition, a function $\phi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ is said to be $c-$concave if it not identically $-\infty$, and there exists $\psi : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ such that $\phi = \psi^c$. then its $c-$transform is the function ϕ^c defined by

$$\forall x \in \mathcal{X} \qquad \phi^c(x) = \sup_{y \in \mathcal{Y}} \left(\phi(y) - c(x, y) \right);$$

and its $c-$sub-differential is the $c-$cyclically monotone set defined by

$$\partial^c \phi := \left\{ (x, y) \subset \mathcal{X} \times \mathcal{Y}; \quad \phi(y) - \phi^c(x) = c(x, y) \right\}.$$

Remark 23. If $c = d$ is a distance on some metric space $\mathcal{X}$,then a $c-$convex function is just a $1-$Lipschitz function and is its own $c-$transform. In general, if c satisfies the triangle inequality then ψ is $c-$convex if and only if $\psi(y) - \psi(x) \leq c(x, y)$ for all x, y; and then $\psi = \psi^c$.

Theorem 24 (**Kantorovich Duality**)**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$be a lower semicontinuous cost function such* *that* $\forall(x,y)$, $c(x,y)\geq a(x)+b(y)$ *for some real valued upper semi-continuous functions* $a\in L^1(\mu)$ *and* $b\in L^1(\nu)$. *Then,*

1. *There is duality*

$$\begin{aligned}
\min_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) &= \sup_{\substack{(\phi,\psi)\in C_0(\mathcal{X})\times C_0(\mathcal{Y})\\ \phi-\psi\leq C}}\left(\int_{\mathcal{Y}}\phi(y)\mathrm{d}\nu(y)-\int_{\mathcal{X}}\psi(x)\mathrm{d}\mu(x)\right)\\
&= \sup_{\substack{(\phi,\psi)\in L^1(\mathcal{X})\times L^1(\mathcal{Y})\\ \phi-\psi\leq C}}\left(\int_{\mathcal{Y}}\phi(y)\mathrm{d}\nu(y)-\int_{\mathcal{X}}\psi(x)d\mu(x)\right)\\
&= \sup_{\psi\in L^1(\mu)}\left(\int_{\mathcal{Y}}\psi^c(y)\mathrm{d}\nu(y)-\int_{\mathcal{X}}\psi(x)d\mu(x)\right) \qquad (4.1)\\
&= \sup_{\phi\in L^1(\nu)}\left(\int_{\mathcal{Y}}\phi(y)\mathrm{d}\nu(y)-\int_{\mathcal{X}}\phi^c(x)d\mu(x)\right)
\end{aligned}$$

   And in the above suprema we may as well impose that ψ *is* c-convex *and* ϕ *is* c-concave.

2. *If* c *is real valued and the optimal cost* $C(\mu,\nu)=\inf_{\pi\in\Pi(\mu,\nu)}\int c d\pi$ *is finite, then there is a* measurable c-cyclically monotone set $\Gamma\subset\mathcal{X}\times\mathcal{Y}$ *(closed if* a, b *and* c *are continuous.) such that for any* $\pi\in\Pi(\mu,\nu)$ *the following are equivalent.*

   (a) π *is optimal;*

   (b) π *is cyclically monotone;*

   (c) *There is a c-convex* ψ *such that,* π*-almost surely,* $\psi^c(y)-\psi(x)=c(x,y)$;

   (d) *There exist* $\psi:\mathcal{X}\to\mathbb{R}\cup\{+\infty\}$ *and* $\phi:\mathcal{Y}\to\mathbb{R}\cup\{-\infty\}$, *such that* $\phi(y)-\psi(x)\leq c(x,y)$ *for all* (x,y), *with equality* π-almost surely*;*

   (e) π *is* concentrated on Γ*.*

3. *If c is real valued, $C(\mu,\nu) < +\infty$, and one has the pointwise upper bound*

$$\begin{aligned} c(x,y) &\leq c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y), \\ (c_{\mathcal{X}}, c_{\mathcal{Y}}) &\in L^1(\mu) \times L^1(\nu) \end{aligned} \tag{4.2}$$

Then both the primal and the dual Kantorovich problem have solutions, so

$$\begin{aligned} \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) &= \max_{\substack{(\phi,\psi)\in L^1(\mathcal{X})\times L^1(\mathcal{Y}) \\ \phi-\psi\leq C}} \left(\int_{\mathcal{Y}} \phi(y)d\nu(y) - \int_{\mathcal{X}} \psi(x)d\mu(x) \right) \\ &= \max_{\psi\in L^1(\mu)} \left(\int_{\mathcal{Y}} \psi^c(y)d\nu(y) - \int_{\mathcal{X}} \psi(x)d\mu(x) \right) \end{aligned}$$

And in the latter expressions we may as well impose that ψ be convex and $\phi = \psi^c$. If in addition a, b and c are continuous, then there is a closed c-cyclically monotone set $\Gamma \subset \mathcal{X} \times \mathcal{Y}$, such that for any $\pi \in \Pi(\mu,\nu)$ and for any c-convex $\psi \in L^1(\mu)$

- *π is optimal in the Kantorovich problem if and only if $\pi(\Gamma) = 1$.*
- *ψ is optimal in the Kantorovich problem if and only if $\Gamma \subset \partial_c\psi$.*

Remark 25. As in case 23 there is the following particular variation of the above duality called the Kantorovich-Rubinstein theorem. When $c(x,y) = d(x,y)$ is a distance on a Polish space $\mathcal{X}$, and μ,ν belong to $\mathcal{P}(\mathcal{X})$,then

$$\inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}[d(x,y)] = \sup \mathbb{E}[\psi(x) - \psi(y)] = \sup \left\{ \int_{\mathcal{X}} \psi d\mu - \int_{\mathcal{Y}} \psi d\nu \right\},$$

where the supremum on the right is over all $1-$Lipschitz functions ψ.

5 Proof of duality.

This section contains the main result. The first subsection introduces the requisite definitions and Lemmas. The second introduces the duality, illustrating the ideas with a practical analogy taken from information theory. The third has a short informal description of the proof as well as a sketch. The fourth subsection contains the rigorous proof. The fifth is a theorem proving some equivalent statement about the domain of the transmission method - similar to part (ii) of the Kantorovich duality in the previous section.

5.1 Definitions and Technical Requirements.

Remark 26. *In the sequel, vectors in $\mathcal{X}$ are denoted x or y. To differentiate between two vectors I use superscripts. for example, a sequence of vectors in $\mathcal{X} \times \mathcal{X}$ is denoted $\{(x^i, y^i)\}_{i\in\mathbb{N}}$. Then the $s-$component of x is denoted x_s and so a sequence of real ordered pairs made up of the s-components of the above sequence of vectors is $\{(x^i_s, y^i_s)\}_{i\in\mathbb{N}}$.*

Definition 27 (Compact Set)**.** A subset C of a topological space X is compact if for every open cover of C there exists a finite sub-cover of C.

Definition 28 (Pre-compactness)**.** A set C in a normed space is pre-compact if every sequence of points in C has a subsequence converging in norm to an element of the space.

Definition 29 (Tight Set)**.** A set $\mathcal{X}$ is tight if for any $\epsilon > 0$ there is a compact set K_ϵ such that $\mu[\mathcal{X}\backslash K_\epsilon] \leq \epsilon$ for all $\mu \in P(\mathcal{X})$.

Remark 30. For the next result we need to recall a technique for constructing bounded metrics. Suppose that d is a metric on the set X. It is easy to check that the formula

$$d_0(x, y) = \min(1, d(x, y)) \tag{5.1}$$

defines a metric on X, that the metrics d and d_0 determine the same topology on X, and that X is complete under d_0 if and only if it is complete under d.

Theorem 31. *(Cohn,1980) A finite or infinite product of Polish spaces is Polish.*

Proof. Let I be a finite or infinite subset of $\mathbb{N}$. Let $(X_i, d_i)_{i \in I}$ be a sequence of Polish spaces, where each d_i is a complete metric which metrizes X_i. We may assume that no X_i is empty. Further, by (5.1) we may assume that $d_i(x_i, y_i) \leq 1$ holds for each i and each (x_i, y_i) in X_i. For points x, y in $\prod_i X_i$, with coordinates $x_1, x_2, ...$ and $y_1, y_2, ...$ respectively, let

$$d(x,y) = \sum_i \frac{1}{2^i} d_i(x_i, y_i).$$

It is easy to check that this defines a metric d on $\prod_i X_i$, that d metrizes the product topology on $\prod_i X_i$, and that $\prod_i X_i$ is complete under d.

We can prove the separability of $\prod_i X_i$by constructing a countable basis for $\prod_i X_i$. For each i choose a countable basis $\mathcal{B}_i$ for X_i. Then the collection of subsets of $\prod_i X_i$ that have the form

$$B_1 \times \cdots \times B_N \times X_{N+1} \times X_{N+2} \times \cdots$$

For some N and some choice of sets B_i in $\mathcal{B}_i$, $i = 1, ..., N$, is the required basis for $\prod_i X_i$. □

Theorem 32 (**Prokhorov's Theorem**). *If $\mathcal{X}$ is a Polish space, then a set $\mathcal{P} \subset P(\mathcal{X})$ is pre-compact for the weak topology if and only if it is tight.*

Theorem 33 (***Varadarajan's Theorem***). *Let $(\mathcal{X}$,$\mu)$ be a Polish probability space. Then the empirical measures $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x^i}$, ($x^i$being a sequence of random variables with values in $\mathcal{X}$ and law μ) converges to μ almost surely.*

Proof of Prokhorov's theorem and Varadarajan's theorem can be found in (Dudley,2002) on pages 404-405 and 399 respectively.

Theorem 34 (**Existence of an optimal coupling**). *Let $(\mathcal{X}, \mu)$ and $(\mathcal{X}, \nu)$ be two Polish probability spaces. Let $d_s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a metric on X_s. Then there is a coupling of (μ, ν) which minimizes maximal distance $\sup_s \mathbb{E} d_s(X, Y)$ over all sites, among all possible couplings (X, Y).*

For the proof we will need the following Lemmas.

Lemma 35. *Let (X_s, d_s) be a collection of Polish spaces with bounded diameter, each metrized respectively by d_s. Then $\mathcal{X} = \prod_s X_s$ is also a polish space by theorem 4. Let $d_s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a metric and $h_s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous function such that for all $s \in S$, $d_s \geq h_s$. Let $(\pi_k)_{k \in \mathbb{N}}$ be a sequence of probability measures on $\mathcal{X} \times \mathcal{X}$, converging weakly to some $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$, in such a way that $h_s \in L^1(\pi_k),\ h_s \in L^1(\pi)$, and*

$$\int_{\mathcal{X} \times \mathcal{X}} h_s d\pi_k \underset{k \to \infty}{\longrightarrow} \int_{\mathcal{X} \times \mathcal{X}} h_s d\pi.$$

Then for every $s \in S$

$$\int_{\mathcal{X} \times \mathcal{X}} d_s(x, y) d\pi \leq \liminf_{k \to \infty} \int_{\mathcal{X} \times \mathcal{X}} d_s(x, y) d\pi_k.$$

In particular, $F : \pi \to \int d_s(x, y) d\pi$ is lower semi-continuous on $\mathcal{P}(\mathcal{X} \times \mathcal{X})$, equipped with the topology of weak convergence.

Proof. Since d_s is non-negative it can be written as the pointwise limit of a non-decreasing family $(d_s^k)_{k \in \mathbb{N}}$ of continuous real-valued functions. By monotone convergence,

$$\int d_s d\pi = \lim_{l \to \infty} \int (d_s)_l d\pi = \lim_{l \to \infty} \lim_{k \to \infty} \int (d_s)_l d\pi_k \leq \liminf_{k \to \infty} \int d_s d\pi_k.$$

This is the desired result. □

Lemma 36 (***Tightness of transmission plans***). *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. Let $P \subset \mathcal{P}(\mathcal{X})$ and $Q \subset \mathcal{P}(\mathcal{Y})$ be tight subsets of $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ respectively. Then* *the sets* $\Pi(P, Q)$ *of all transference plans whose marginals lie in* P *and* Q *respectively, is itself tight in* $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

Proof. Let $\mu \in P$, $\nu \in Q$, and $\pi \in \Pi(\mu, \nu)$. By assumption, for any $\epsilon > 0$ there is a compact set $K_\epsilon \subset \mathcal{X}$, independent of the choice of μ in P, such that $\mu[\mathcal{X} \backslash K_\epsilon] \leq \epsilon$; and similarly there is a compact set $L_\epsilon \subset \mathcal{Y}$, independent of the choice of ν in Q, such that $\nu[\mathcal{X} \backslash L_\epsilon] \leq \epsilon$. Then for any coupling (X, Y) of (μ, ν),

$$\mathbb{P}\left[(X, Y) \notin K_\epsilon \times L_\epsilon\right] \leq \mathbb{P}\left[X \notin K_\epsilon\right] + \mathbb{P}\left[Y \notin L_\epsilon\right] \leq 2\epsilon$$

The desired result follows since this bound is independent of the coupling, and $K_\epsilon \times L_\epsilon$ is compact in $\mathcal{X} \times \mathcal{X}$. □

Proof of theorem 6. Since $\mathcal{X}$ is Polish, $\{\mu\}$ and $\{\nu\}$ are tight in $\mathcal{P}(\mathcal{X})$; By Lemma 15, $\Pi(\mu, \nu)$ is tight in $P(\mathcal{X} \times \mathcal{X})$, and by Prokhorov's theorem this set has a compact closure. By passing to the limit in the equation for the marginals, we see that $\Pi(\mu, \nu)$ is closed, so it is compact. Then let $(\pi_k)_{k \in \mathbb{N}}$ be a sequence of probability measures on $\mathcal{X} \times \mathcal{X}$, such that $\sup_s \int d_s d\pi_k$ converges to the infimum transmission error. Extracting a subsequence if necessary, we may assume π_k converges to some $\pi \in \Pi(\mu, \nu)$. The function $h_s : (x_s, y_s) \mapsto (0, 0)$ lies in $L^1(\pi_k)$ and in $L^1(\pi)$, and $d_s \geq h_s$ by assumption; moreover, $\int h_s d\pi_k = \int h_s d\pi = 0$; so Lemma 4.3 implies

$$\int d_s d\pi \leq \liminf_{k \to \infty} \int d_s d\pi_k$$

and

$$\sup_s \int d_s d\pi \leq \sup_s \liminf_{k \to \infty} \int d_s d\pi_k.$$

Thus π is minimizing. □

5.2 Developing Duality with an Informal Analogy

I start with the characterization of Steif's metric as an optimization problem.

Suppose we have an alphabet $\mathcal{A}$ consisting of symbols and S is a countable set with cardinality[6] N where $s \in S$ represents the s^{th} element in a set indexed by S. Let each X_s be a subset of $\mathcal{A}$ representing the set of allowable symbols at x_s (the s^{th} character of a word). So each vector $x \in \mathcal{X}$ represents a word and $\mathcal{X}$ can be referred to as the dictionary, while some sequence $(x^i)_{i\in\mathbb{N}} \subset \mathcal{X}$ is called a message.

We will consider the situation where messages are sent to a receiver via some transmission method. Let μ and ν be probability distributions representing the probability of a word before and after transmission respectively. Now π can be thought of as a transmission method. A transmission method consists of an encoding method, transmission medium and a decoding method. The encoded signal, while travelling through the medium may be influenced by external noise. We describe such noise with the vector $e = (e_s)_{s\in S}$ which we will from now on refer to as "intrinsic error". Let $e \in \mathcal{E}$ where $\mathcal{E} = \{e \in \mathbb{R}^N : \sum_s e_s \leq 1,\ e_s > 0\}$

The joint distribution $\pi((x, y))$ is the probability that a signal x transmitted via the method π is received as y.

Suppose now that you have cooked up some transmission process but have found that too often the output has errors. You know that when encoding a message the information is compressed and as such an error at any one site can completely change the signal (due to the decoding process). So if d_s is the error quantification (distance between two symbols), x the input signal and y the output signal then we want to find a plan π^* such that

[6] S may be infinite but for illustration purposes I will stick to finite sites. In the proof, step 2 extends the theory to infinite S.

$$\pi^* = \inf_{\pi \in \Pi(\mu,\nu)} \sup_{s \in S} \mathbb{E}^{\pi}[d_s(x,y)].$$

In words: π^* is the transmission method that provides the smallest maximum error that is to be expected between any symbol before transmission and it's corresponding symbol (occupying its position in the message) after transmission. Now consider $\{x^i, y^i\}_{i \in I \subset \mathbb{N}}$ as sequence of words in a message, then we have the following definition.

Definition 37 ($(S, e_s d_s)-$ Cyclical Monotonicity)**.** Let $\mathcal{X} = \prod X_s$, where (X_s, d_s) are polish space. A subset $\Gamma \subset \mathcal{X} \times \mathcal{X}$ is said to be $e_s d_s-$cyclically monotone if, for any collection $M \in \mathbb{N}$, and any family $(x_s^1, y_s^1), ..., (x_s^M, y_s^M)$ of points in Γ, holds the inequality

$$\sup_{s \in S} \sum_{i=1}^{M} e_s d_s(x_s^i, y_s^i) \leq \sup_{s \in S} \sum_{i=1}^{M} e_s d_s(x_s^i, y_s^{i+1}) \tag{5.2}$$

(with the conventions that $y_s^{N_s+1} = y_s^1$). A transmission method is said to be $e_s d_s-$cyclically monotone if it is concentrated on a $e_s d_s-$cyclically monotone set.

Cyclical monotonicity enables us to ignore methods that prioritize minimizing spurious errors. These errors are spurious because the surrounding symbols matter, two words are not encoded the same way and the s^{th} symbol of the $(i+1)^{th}$ word may not even be the same as that of the i^{th} word! That leads us to the following definition.

While it's fairly obvious that an optimal plan must be cyclically monotone the converse is less clear - but it never the less holds as soon the diameter of each X_s is finite and each d_s is real valued. Theorem 38 "equivalence statements about optimal plans" goes into the details of this.

Now we will consider the dual problem. In the primal problem we are trying to minimize the impact of errors, in the dual problem we will be looking for the most accurate coding strategy. Let $\psi(y)$ be the error rate when decoding y and $\phi(x)$ the error detection accuracy we can achieve when encoding x. Ideally we would detect all errors made in the encoding process and decode without making errors. We will refer to the pair (ψ, ϕ) as a coding strategy. The accuracy of our coding strategy is $\phi(x) - \psi(y)$ which is assumed to be non-negative. Of course this is for each word, if the word x occurs at rate $\mu(dx)$ then the total error rate will be $\psi(x)\mu(dx)$. Obviously the accuracy can not be greater than the distance between two words so

$$\phi(x) - \psi(y) \leq \sum_s d_s(x, y).$$

However, absolute accuracy may not be possible as there could be noise in the transmission media. Taking into account intrinsic error gives

$$\phi(x) - \psi(y) \leq \sum_{s \in S} e_s d_s(x_s, y_s). \tag{5.3}$$

In our analogy e shrinks the measurement of the error making it less likely to be picked up. We call a coding strategy satisfying this inequality plausible. The following lemma will come in very useful.

Lemma 38. *For any functions $\phi \in L^1(\mu)$ and $\psi \in L^1(\nu)$ and $e = (e_s)_{s \in S}$ a vector contained in $\mathcal{E}$. Then, we have*

$$\phi(x) - \psi(y) \leq \sum_{s \in S} e_s d_s(x_s, y_s).$$

If and only if, for each $s \in S$and all $(x, y) \in \mathcal{X} \times \mathcal{X}$such that $x_t = y_t$off s,we have

$$\phi(x) - \psi(y) \leq e_s d_s(x_s, y_s). \tag{5.4}$$

Proof. For sufficiency it is enough to simply specialize $\sum_s e_s d_s$ to $x = y$ off s. For necessity, fix $x, y \in \mathcal{X}$ and put $J = \{s \in S : x_s \neq y_s\}$. Without loss of generality we may assume that $J = (j_n)_{n\in\mathbb{N}}$. Let $(x_n)_{n\in\mathbb{N}}$be a sequence of configurations such that $x_0 = x$, for which $\lim_{n\to\infty} x_n = y$ and such that for all $n, m \in \mathbb{N}$:

$$x^n_{j_n} \neq x^{n+1}_{j_n}, \quad x^n_b = x^{n+1}_b \; (b = j_n), \quad x^{n+m+1}_{j_n} = y_{j_n}.$$

Which implies that

$$f(x) - f(y) \leq \sum_n f(z^n) - f(z^{n-1}) \leq \sum_n e_{j_n} d_{j_n}(x, y)$$

which is the desired conclusion. □

Where as before the task was to minimize the error, now it is to maximize the accuracy. This leads naturally to the dual problem:

Definition 39 (Dual problem)**.**

$$\sup\left\{\int_{\mathcal{Y}} \phi(y)d\nu(y) - \int_{\mathcal{X}} \psi(x)d\mu(x); \quad \phi(y) - \psi(x) \leq e_s d_s(x, y); \quad x = y \text{ off } s\right\} \tag{5.5}$$

Impose that ϕ and ψ be integrable: $\psi \in L^1(\mathcal{X}, \mu)$; $\phi \in L^1(\mathcal{Y}, \nu)$. By linearity of the integral and (5.4), for a given e we have

$$\sup_{\substack{\phi-\psi\leq e_s c_s \\ x=y \text{ off } s}} \left\{\int_{\mathcal{Y}} \phi(y)d\nu(y) - \int_{\mathcal{X}} \psi(x)d\mu(x)\right\} \leq \inf_{\pi\in\Pi}\left\{\sup_{s\in S}\int d_s(x_s, y_s)d\pi(x, y)\right\} \tag{5.6}$$

This is as expected since optimizing the coding strategy cannot increase the error rate of the transmission.

Denote by d_{s^*} the metric on the allowable symbols at site $s \in S$ with the largest expected error with respect to π. This exists because the state space at each site is of bounded diameter, hence d_s is bounded above and below for all s. As a result we may conclude that d_s is integrable and the supremum of $\int d_s d\pi$ can be identified. Of course in the dual problem we would like to choose the best possible plausible strategy. So for any given y we would choose the highest lower bound of $\psi(x) + e_{s^*} d_{s^*}(x_s, y_s)$ and consider the worst case scenario with regards to the intrinsic error. Likewise for any given x we would choose the supremum of $\psi(x) + e_{s^*} d_{s^*}(x_s, y_s)$ and take the infimum over all intrinsic error vectors.

Definition 40 (Tight Coding Strategy)**.** Let $|S| = N$. The coding strategy (ψ, ϕ) is tight if, for each s and $\forall (x, y) \in \mathcal{X} \times \mathcal{X} : x = y$ off s,

$$\phi(y) = \sup_{e \in \mathcal{E}} \inf_{x} \left(\psi(x) + e_{s^*} d_{s^*}(x_s, y_s) \right), \qquad \psi(x) = \inf_{e \in \mathcal{E}} \sup_{y} \left(\phi(y) - e_{s^*} d_{s^*}(x_s, y_s) \right) \quad (5.7)$$

and we denote by e^* any $e \in \mathcal{E}$ for which the above equality holds.

A coding strategy being tight implies that one cannot decrease the decoding error or increase the encoding accuracy while keeping the strategy plausible. Both formulae in the above definition will only hold simultaneously if ψ has the following convexity property.

Definition 41 ($e_s d_s-$convexity)**.** Let $\mathcal{X} = \prod_s X_s$ be a set and $d_s : \mathcal{X} \times \mathcal{X} \to [0, +\infty]$ such that $x = y$ off s. A function $\psi : \mathbb{R} \cup \{+\infty\}$ is said to be $e_s d_s-$convex if it is not identically $+\infty$, and there exists $\zeta : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ such that

$$\forall x \in \mathcal{X} \quad \psi(x) = \inf_{e \in \mathcal{E}} \sup_{y \in \mathcal{X}} \left(\zeta(y) - e_{s^*} d_{s^*}(x, y) \right) \quad (5.8)$$

Then its $e_s d_s-$transform is the function $\psi^{e_s d_s}(y)$ defined by

$$\forall y \in \mathcal{X} \quad \psi^{e_s d_s}(y) = \sup_{e \in \mathcal{E}} \inf_{x \in \mathcal{X}} \left(\psi(x) + e_{s^*} d_{s^*}(x, y) \right) \quad (5.9)$$

and its $e_s d_s-$sub-differential is the $e_s d_s-$cyclically monotone set defined by

$$\partial_{e_s d_s}\psi := \{(x,y) \in \mathcal{X} \times \mathcal{X}; \psi^{e_s d_s}(y) - \psi(x) = e^*_{s^*} d_{s^*}(x,y)\}$$

The functions ψ and $\psi^{e_s d_s}$ are said to be $e_s d_s-$conjugate. Moreover, the $e_s d_s-$sub-differential of ψ at point x is

$$\partial_{e_s d_s}\psi(x) = \{y \in \mathcal{Y}; (x,y) \in \partial_{e_s d_s}\psi\}$$

or equivalently

$$\forall z \in \mathcal{X}, \quad \psi(x) + e^*_{s^*} d_{s^*}(x,y) \leq \psi(z) + e^*_{s^*} d_{s^*}(z,y) \tag{5.10}$$

since,

$$\begin{aligned}
\psi^c(y) - \psi(x) &= e^*_s d_{s^*}(x,y) \\
&\Rightarrow \sup_{e\in\mathcal{E}} \inf_{x\in\mathcal{X}} \left(\psi(x) + e_{s^*} d_{s^*}(x,y)\right) = \psi(x) + e^*_s d_{s^*}(x,y) \\
&\Rightarrow \psi(x) + e^*_{s^*} d_{s^*}(x,y) \leq \psi(z) + e^*_{s^*} d_{s^*}(z,y)
\end{aligned}$$

Remark 42. By Lemma 8 we can generalize (0.8) to $\forall z \in \mathcal{X}, \quad \psi(x) + \sum_s e^*_s d_{s^*}(x,y) \leq \psi(z) + \sum_s e^*_s d_{s^*}(z,y)$

Remark 43. Since $\sum_s e^*_s d_{s^*}$ satisfies the triangle inequality $\sum_s e^*_s d_{s^*}(x,z) \leq \sum_s e^*_s d_{s^*}(x,y) + \sum_s e^*_s d_{s^*}(y,z)$, it follows that ψ is $c-$convex if and only if $\psi(y) - \psi(x) \leq \sum_s e^*_s d_{s^*}(x,y)$ for all x, y. Hence $\psi = \psi^{e_s d_s}$.

Proposition 44 (***Alternative Characteristic of*** $e_s d_s-$***convexity***)**.** *For any function $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, let its $e_s d_s-$ convexification be defined by $\psi^{dd} = (\psi^d)^d$. More* *explicitly,*

$$\psi^{dd}(x) = \sup_{y \in \mathcal{X}} \inf_{z \in \mathcal{X}} \left(\psi(z) + e^*_{s^*} d_{s^*}(z,y) - e^*_{s^*} d_{s^*}(x,y) \right).$$

Then ψ is $e_s d_s-$convex if and only if $\psi^{dd} = \psi$.

Proof. For any function $\phi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ (not necessarily $e_s d_s-$convex), it holds that $\phi^{ddd} = \phi^d$. More accurately we have

$$\phi^{ddd}(x) = \sup_y \inf_z \sup_w \left[\phi(w) - e^*_{s^*} d_{s^*}(z,w) + e^*_{s^*} d_{s^*}(z,y) - e^*_{s^*} d_{s^*}(x,y) \right].$$

Now, choose $z = x$ which gives us $\phi^{ddd}(x) \leq \phi^d(x)$; while the choice $w = y$ shows that $\phi^{ddd}(x) \geq \phi^d(x)$.

If ψ is $e_s d_s-$convex, then there is ζ such that $\psi = \zeta^d$,so $\psi^{dd} = \zeta^{ddd} = \zeta^d = \psi$. Conversely, if $\psi^{dd} = \psi$, then ψ is $e_s d_s-$convex, as the $e_s d_s-$transform of ψ^c. □

5.3 Statement of the Duality

Theorem 45. *Let S be any countable set and consider a collection $\{X_s\}_{s \in S}$ such that each X_s is a complete separable metric space of bounded diameter. Let $\mathcal{X} = \Pi_s X_s$ and d_s be a collection of metrics on $\mathcal{X}$ defined by $d_s(x,y) = d_s(x_s, y_s)$. Denote the set of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν by $\Pi(\mu, \nu)$. Then there is*

Duality;

$$\begin{aligned}
&\min_{\pi\in\Pi(\mu,\nu)}\sup_{s\in S}\int_{\mathcal{X}\times\mathcal{X}} d_s(x_s,y_s)d\pi(x,y)\\
&=\sup_{\substack{(\phi,\psi)\in C_0(\mathcal{X})\times C_0(\mathcal{X})\\ \phi-\psi\le e_sd_s}}\left(\int_{\mathcal{X}}\phi(y)d\nu(y)\text{-}\int_{\mathcal{X}}\psi(x)d\mu(x)\right)\\
&=\sup_{\substack{(\phi,\psi)\in L^1(\mathcal{X})\times L^1(\mathcal{X})\\ \phi-\psi\le e_sd_s}}\left(\int_{\mathcal{X}}\phi(y)d\nu(y)-\int_{\mathcal{X}}\psi(x)d\mu(x)\right) \qquad (5.11)\\
&=\sup_{\psi\in L^1(\mu)}\left(\int_{\mathcal{X}}\psi^{e_sd_s}(y)d\nu(y)-\int_{\mathcal{X}}\psi(x)d\mu(x);\psi^{e_sd_s}-\psi\le e_sd_s\right)\\
&=\sup_{\phi\in L^1(\nu)}\left(\int_{\mathcal{X}}\psi(y)d\nu(y)-\int_{\mathcal{X}}\psi(x)d\mu(x);\psi(y)-\psi(x)\le e_sd_s\right)
\end{aligned}$$

and in the above suprema one might as well impose that ψ be e_sd_s-convex and ϕ e_sd_s-concave.

Theorem. *Further, if for all $s\in S$ the distance metric is real valued, the optimal cost is finite, and one has the pointwise upper bound*

$$d_s(x,y)\le f_s(x)+g_s(y),\quad (f_s(x),g_s(y))\in L^1(\mu)\times L^1(\nu),$$

where f_s and g_s are functions from X_s to $\mathbb{R}$ then;

1. Both the primal and dual problems have solutions, so

$$\min_{\pi\in\Pi(\mu,\nu)}\max_{s\in S}\int_{\mathcal{X}\times\mathcal{X}} d_sd\pi=\max_{\psi\in L^1(\mu)}\left(\int_{\mathcal{X}}\psi(y)d\nu(y)-\int_{\mathcal{X}}\psi(x)d\mu(x)\right)$$

2. There is a closed e_sd_s-cyclically monotone set $\Gamma\subset\mathcal{X}\times\mathcal{X}$, such that for any $\pi\in\Pi(\mu,\nu)$ and for any e_sd_s-convex $\psi\in L^1(\mu)$,

   (a) π is optimal in the primal problem if and only if $\pi[\Gamma]=1$;

   (b) ψ is optimal in the dual problem if and only if $\Gamma\subset\partial_{e_sd_s}\psi$.

5.4 Idea for Proof:

The main steps are as follows

1. Prove that the optimal plan is cyclically monotone if the distributions μ and ν are delta functions at each point in statespace. This would be sufficient to prove the finite case.

2. Use the central limit theorem to extend this existence theorem to countable S and X_s.

3. Show that if π is optimal it must lie in the sub-differential of ψ.

4. Show that ψ and ϕ must be bounded and by step 3 are optimal.

5. Show that ψ and ϕ are integrable and hence a feasible point exists for both problems.

We have proved that an optimal plan π indeed exists. For each $s \in S$ the state-space X_s is bounded so there exists $s^* \in S$ such that $\int d_{s^*} d\pi = \sup_s \int d_s d\pi$. Now, let (ψ, ϕ) a coding strategy satisfying. Of course, if $x_t = y_t$ $\forall t : t \neq s$ then

$$\int e^*_{s^*} d_{s^*}(x,y) d\pi(x,y) \geq \int \phi(x) d\nu - \int \psi(y) d\mu = \int \left[\phi(y) - \psi(x)\right] d\pi(x,y)$$

So if the inequality is an equality we have $\int \left[e^*_{s^*} d_{s^*} - \phi + \psi\right] d\pi = 0$, and hence

$$\phi(y) - \psi(x) = e^*_{s^*} d_{s^*} \quad \pi(dxdy) - a.s.$$

Intuitively speaking, whenever y is an allowable input and x an allowable output, the a coding strategy is chosen so that the accuracy is equal to the maximal error at any site scaled according to the worst case intrinsic error. Now let $(x^i)_{0 \leq i \leq m}$ and $(y^i)_{0 \leq i \leq m}$

be such that $d_{s^*}(x^i_{s^*}, y^i_{s^*})$. Then we observe that

$$\begin{cases} \phi(y^0) - \psi(x^0) & = e^*_{s^*} d_{s^*}(x^0, y^0) \\ \phi(y^1) - \psi(x^1) & = e^*_{s^*} d_{s^*}(x^1, y^1) \\ \vdots & \vdots \\ \phi(y^m) - \psi(x^m) & = e^*_{s^*} d_{s^*}(x^m, y^m) \end{cases}$$

On the other hand, if x is an arbitrary point,

$$\begin{cases} \phi(y^0) - \psi(x^1) & \leq e^*_{s^*} d_{s^*}(x^1, y^0) \\ \phi(y^1) - \psi(x^2) & \leq e^*_{s^*} d_{s^*}(x^0, y^1) \\ \vdots & \vdots \\ \phi(y^m) - \psi(x) & \leq e^*_{s^*} d_{s^*}(x, y^m) \end{cases}$$

By subtracting these inequalities from the previous equalities and adding everything up, one obtains

$$\psi(x) \geq \psi(x^0) + \left[e^*_{s^*} d_{s^*}(x^0, y^0) - e^*_{s^*} d_{s^*}(x^1, y^0)\right] + \cdots + \left[e^*_{s^*} d_{s^*}(x^m, y^m) - e^*_{s^*} d_{s^*}(x, y^m)\right]$$

Of course, one can add an arbitrary constant to ψ, provided that one subtracts the same constant from ϕ; so it is possible to decide that $\psi(x_0) = 0$, where (x^0, y^0) is arbitrarily chosen in the support of π. Then

$$\psi(x) \geq \left[e^*_{s^*} d_{s^*}(x^0, y^0) - e^*_{s^*} d_{s^*}(x^1, y^0)\right] + \cdots + \left[e^*_{s^*} d_{s^*}(x^m, y^m) - e^*_{s^*} d_{s^*}(x, y^m)\right]$$

and this should be true for all choices of (x^i, y^i) $(1 \leq i \leq m)$ in the support of π, and for $m \geq 1$. So it becomes natural to define ψ as the supremum of all the function (in the variable x) appearing on the right hand side above. It will turn out that this ψ satisfies the equation

$$\psi^{e_s d_s}(y) - \psi(x) = e^*_{s^*} d_{s^*}(x,y) \quad \pi(dxdy) - a.s.$$

d_s

for all $(x,y) \in \mathcal{X} \times \mathcal{X} : x_t = y_t \iff t \neq s$. By lemma 8 we can generalize to

$$\psi^{e_s d_s}(y) - \psi(x) = \sum e^*_s d_{s^*}(x,y) \quad \pi(dxdy) - a.s.$$

for all $(x,y) \in \mathcal{X} \times \mathcal{X}$. Then, if ψ and $\psi^{e_s d_s}$ are integrable, one can write

$$\sup_s \int d_s d\pi = \frac{\int \psi^c d\pi - \int \psi d\pi}{\sum_s e^*_s} = \frac{\int \psi d\nu - \int \psi d\mu}{\sum_s e^*_s}.$$

The last equality following from remark 43. This shows at the same time that π is optimal in the Primal problem, and the function ψ is optimal in the dual problem.

5.5 Rigorous Proof.

Throughout the proof let $|S| = N$ and $|X_s| = n_s$.

Proof. **Step 1: If** $\mu = \frac{1}{\sum_s n_s} \sum_{s=1}^N \sum_{i=1}^{n_s} \delta_{x^i_s}$, $\nu = \frac{1}{\sum_s n_s} \sum_{s=1}^N \sum_{j=1}^{n_s} \delta_{y^j_s}$**, where the distances** $e_s d_s(x^i_s, y^j_s)$ **are finite, then there is at least one cyclically monotone transmission method.**

In this particular case, a transmission method between μ and ν can be identified with a bi-stochastic $N \times N$ array of real valued matrices a_s with components $a^{ij}_s \in [0,1]$: each a^{ij}_s tells us the rate at which the point x^i_s, occurring with probability $\frac{1}{\sum_s n_s}$ will be interpreted as y_s^j so the primal problem becomes

$$\inf_{\left(a_s^{ij}\right)} \sup_s \sum_{ij} a_s^{ij} e_s d_s(x_s^i, y_s^j)$$

Where the infimum is over all arrays (a_s^{ij}) satisfying

$$\sum_s \sum_i a_s^{ij} = 1, \quad \sum_s \sum_j a_s^{ij} = 1. \tag{5.12}$$

Here we are minimizing a linear function on the compact set $[0,1]^{(n_s \times n_s) \times (n_s \times n_s)}$, so obviously there exists a minimizer; the corresponding transmission method π can be written as

$$\pi = \frac{1}{\sum_s n_s} \sum_{s=1}^{N} \sum_{i,j=1}^{n_s} a_s^{ij} \delta_{(x_s^i, y_s^j)}$$

and its support Γ is the set of all couples (x^i, y^j) such that $a_s^{ij} > 0$. Assume that $\Gamma = \text{Supp}\pi$ is not cyclically monotone: Then there exist $M \in \mathbb{N}$ and $(x_s^{i_1}, y_s^{j_1}), ..., (x_s^{i_M}, y_s^{j_M})$ in Γ such that

$$\sup_s \left\{ e_s d_s \left(x_s^{i_1}, y_s^{j_2}\right) + ... + e_s d_s \left(x_s^{i_M}, y_s^{j_1}\right) \right\} < \sup_s \left\{ e_s d_s \left(x_s^{i_1}, y_s^{j_1}\right) + ... + e_s d_s \left(x_s^{i_M}, y_s^{j_M}\right) \right\}.$$

Let $a := \min\left(a_s^{i_1, j_1}, ..., a_s^{i_M, j_M}\right) > 0$. Define a new transmission method $\tilde{\pi}$ by the formula

$$\tilde{\pi} = \pi + \frac{a}{\sum_s n_s} \sum_{s=1}^{N} \sum_{l=1}^{n_s} \left(\delta_{(x_s^{i_l}, y_s^{j_{l+1}})} - \delta_{(x_s^{i_l}, y_s^{j_l})} \right)$$

One can check that this has the same marginals, and the error rate associated with $\tilde{\pi}$ is strictly less than the error rate associates with π. This is a contradiction, so Γ is indeed $(S, e_s d_s)-$cyclically monotone!

Step 2: If S and X_s are countably infinite, then there is a cyclically monotone transmission method.

To prove this, consider sequences of independent random variables $x_s^i \in \mathcal{X}$, $y_s^i \in \mathcal{X}$, with respective law μ, ν. According to Varadarajan's theorem, one has, with probability one,

$$\mu_{N,n} := \frac{1}{\sum_s n_s} \sum_{s=1}^{N} \sum_{i=1}^{n_s} \delta_{x_s^i} \to \mu, \quad \nu_{N,n} := \frac{1}{\sum_s n_s} \sum_{s=1}^{N} \sum_{j=1}^{n_s} \delta_{y_s^j} \to \nu \tag{5.13}$$

as $N, n_s \to \infty$, for all $s \in S$ in the sense of weak convergence of measures. In particular, by Prokhorovs theorem, μ_{N,n_s} and ν_{N,n_s} are tight sequences.

For each pair N, n_s, let π_{N,n_s} be a cyclically monotone transmission method between μ_{N,n_s} and ν_{N,n_s}. By Lemma 36 on page 24, $\{\pi_{N,n_s}\}_{N,n_s\in\mathbb{N}}$ is tight. By Prokhorov's theorem, there is a subsequence, still denoted (π_{N,n_s}) ,which converges weakly to some probability measure π, i.e.

$$\int h(x,y) d\pi_{N,n_s}(x,y) \to \int h(x,y) d\pi(x,y)$$

for all bounded continuous functions h on $\mathcal{X} \times \mathcal{X}$. By applying the previous identity with $h(x,y) = f(x)$ and $h(x,y) = g(y)$, we see that π has marginals μ and ν.

For each N and each n_s, the cyclical monotonicity of π_{N,n_s} implies that for all M and $\pi_{N,n_s}^{\otimes M}$−almost all $(x_s^1, y_s^1), ..., \left(x_s^M, y_s^M\right)$,the inequality (5.2) is satisfied; in other words, $\pi_{N,n_s}^{\otimes M}$ is concentrated on the set $\mathcal{C}(M)$ of all $\left((x_s^1, y_s^1), ..., \left(x_s^M, y_s^M\right)\right)_{s\in S} \in (\mathcal{X} \times \mathcal{X})^M$ satisfying ((5.2)). Since d_s is continuous $\forall s \in S$, $\mathcal{C}(M)$ is a closed set, so the weak limit $\pi^{\otimes M}$ of $\pi_{N,n_s}^{\otimes M}$ is also concentrated on $\mathcal{C}(M)$. Let $\Gamma = \mathrm{Supp}\pi$, then

$$\Gamma^M = (\mathrm{Supp}\pi)^M = \mathrm{Supp}\left(\pi^{\otimes M}\right) \subset \mathcal{C}(M),$$

and since this holds true for all M, Γ is cyclically monotone.

Step 3: If each d_s is real-valued and π is cyclically monotone, then there is an $e_s d_s -$convex ψ such that $\partial_{e_s d_s} \psi$ contains the support of π.

Since π is optimal and by steps one and two is cyclically monotone we have the following implication. If $s^* \in S$ is such that $\int e_{s^*} d_{s^*} d\pi = \sup_s \int e_s d_s d\pi$ then from cyclical monotonicity

$$\sum_{i=1} e_{s^*} d_{s^*}(x^i, y^i) \leq \sum e_{s^*} d_{s^*}(x^i, y^{i+1})$$

which we can rewrite as

$$\sum_{i=1} [e_{s^*} d_{s^*}(x^i, y^i) - e_{s^*} d_{s^*}(x^i, y^{i+1})] \leq 0.$$

Indeed, let Γ again denote the support of π (a closed set). Pick any $(x^0, y^0) \in \Gamma$,and define

$$\psi(x) := \inf_{e \in \mathcal{E}} \sup_{m \in \mathbb{N}} \sup \{ \left[e_{s^*} d_{s^*}(x^0, y^0) - e_{s^*} d_{s^*}(x^1, y^0) \right] + \left[e_{s^*} d_{s^*}(x^1, y^1) - e_{s^*} d_{s^*}(x^2, y^1) \right] +$$

$$... + [e_{s^*} d_{s^*}(x^m, y^m) - e_{s^*} d_{s^*}(x, y^m)]]; \;\; (x^1, y^1), ..., (x^m, y^m) \in \Gamma \} \quad (5.14)$$

By applying the definition with $m = 1$ and $(x^1, y^1) = (x^0, y^0)$, one immediately sees that $\psi(x^0) \geq 0$. On the other hand, $\psi(x^0)$ is the supremum of all the quantities $\inf_{e \in \mathcal{E}} \{[e_{s^*} d_{s^*}(x^0, y^0) - e_{s^*} d_{s^*}(x^1, y^0)] + ... + [e_{s^*} d_{s^*}(x^m, y^m) - e_{s^*} d_{s^*}(x^0, y^m)]\}$ which by cyclical monotonicity are all non-positive. So actually $\psi(x^0) = 0$.

By renaming y_m as y, obviously

$$\psi(x) = \sup_{y\in\mathcal{X}} \inf_{e\in\mathcal{E}} \sup_{m\in\mathbb{N}} \sup_{(x^1,y^1),...,(x^{m-1},y^{m-1}),x^m} \{ \left[e_{s^*}d_{s^*}(x^0,y^0) - e_{s^*}d_{s^*}(x^1,y^0)\right] + \left[e_{s^*}d_{s^*}(x^1,y^1) - e_{s^*}d_{s^*}(x^2,y^1)\right] + ... + [e_{s^*}d_{s^*}(x^m,y) - e_{s^*}d_{s^*}(x,y)]); \left(x^1,y^1\right), ..., (x^m,y) \in \Gamma\} \tag{5.15}$$

So $\psi(x) = \inf_{e\in\mathcal{E}} \sup_y \left[\zeta(y) - e_{s^*}d_{s^*}(x,y)\right]$, if ζ is defined by

$$\zeta(y) = \sup\{ \left[e^*_{s^*}d_{s^*}(x^0,y^0) - e^*_{s^*}d_{s^*}(x^1,y^0)\right] + \left[e^*_{s^*}d_{s^*}(x^1,y^1) - e^*_{s^*}d_{s^*}(x^2,y^1)\right] + ... + e^*_{s^*}d_{s^*}(x^m,y)); m\in\mathbb{N}, \left(x^1,y^1\right), ..., (x^m,y) \in \Gamma\} \tag{5.16}$$

(with the convention that $\zeta = -\infty$ out of $\mathrm{proj}_{\mathcal{X}}(\Gamma)$). Thus ψ is a e_sd_s-convex function.

Now let $(\bar{x},\bar{y}) \in \Gamma$. By choosing $x^m = \bar{x},\ y^m = \bar{y}$ in the definition of ψ,

$$\psi(x) \geq \inf_{e\in\mathcal{E}} \sup_{m} \sup_{(x^1,y^1),...,(x^{m-1},y^{m-1})} \{ \left[e_{s^*}d_{s^*}(x^0,y^0) - e_{s^*}d_{s^*}(x^1,y^0)\right] + \cdots + \left[e_{s^*}d_{s^*}\left(x^{m-1},y^{m-1}\right) - e_{s^*}d_{s^*}\left(\bar{x},y^{m-1}\right)\right] + \left[e_{s^*}d_{s^*}\left(\bar{x},\bar{y}\right) - e_{s^*}d_{s^*}\left(x,\bar{y}\right)\right]\}.$$

In the definition of ψ, it does not matter whether one takes the supremum over $m-1$ or over m variables, since one also takes the supremum over m. So the Previous inequality can be recast as

$$\psi(x) \geq \psi(\bar{x}) + e^*_{s^*}d_{s^*}(\bar{x},\bar{y}) - e^*_{s^*}d_{s^*}(x,\bar{y}).$$

In particular, $\psi(x) + e^*_{s^*}d_{s^*}(x,\bar{y}) \geq \psi(\bar{x}) + e^*_{s^*}d_{s^*}(\bar{x},\bar{y})$. As was proved in Lemma 4 this can be extended to the sum over $s \in S$;

$$\psi(x) + \sum_{s\in S} e^*_s d_{s^*}(x, \bar{y}) \geq \psi(\bar{x}) + \sum_{s\in S} e^*_s d_{s^*}(\bar{x}, \bar{y})$$

Taking the infimum over $x \in \mathcal{X}$ in the left-hand side, we deduce that

$$\psi^{e_s d_s}(\bar{y}) \geq \psi(\bar{x}) + \sum_{s\in S} e^*_s d_{s^*}(\bar{x}, \bar{y}).$$

Since the reverse inequality is always satisfied, actually

$$\psi^{e_s d_s}(\bar{y}) = \psi(\bar{x}) + \sum_{s\in S} e^*_s d_{s^*}(\bar{x}, \bar{y}),$$

and this means precisely that $(\bar{x}, \bar{y}) \in \partial_{e_s d_s}\psi$. So Γ does lie in the $e_s d_s-$sub-differential of ψ.

Step 4: There is duality.

Let $||d_{s^*}|| := \sup_{(x,y)\in\mathcal{X}\times\mathcal{X}} e^*_{s^*} d_{s^*}$. By steps two and three, there exists a transmission method π whose support is included in $\partial_{e_s d_s}\psi$ for some $e_s d_s-$convex ψ, and which was constructed "explicitly" in Step three. Let $\phi = \psi^{e_s d_s}$.

From equation (5.14), $\psi = \sup \psi_m$, where each ψ_m is a supremum of continuous functions, and therefore lower semi-continuous. In particular, ψ is measurable. The same is true of ϕ.

Next we check that ψ, ϕ are bounded. Let $(x^0, y^0) \in \partial_{e_s d_s}\psi$ be such that $\psi(x^0) < +\infty$; then consequently $\phi(y^0) > -\infty$. So, for any $x \in \mathcal{X}$,

$$\psi(x) = \inf_{e\in\mathcal{E}} \sup_{y} \left[\phi(y) - e_{s^*} d_{s^*}(x, y)\right] \geq \phi(y^0) - e^*_{s^*} d_{s^*}(x, y^0) \geq \phi(y^0) - ||d_{s^*}||;$$

$$\phi(y) = \inf_{e\in\mathcal{E}} \inf_{x} \left[\phi(x) + e_{s^*} d_{s^*}(x_{s^*}, y_{s^*})\right] \leq \psi(x^0) + e^*_{s^*} d_{s^*}(x^0_{s^*}, y_{s^*}) \leq \psi(x^0) + ||d_{s^*}||.$$

Re-injecting these bounds into the identities $\psi = \phi^{e_s d_s}, \phi = \psi^{e_s d_s}$ and applying Proposition 44, we get

$$\psi(x) \leq \sup_y \phi(y) \leq \psi(x^0) + ||d_s||;$$

$$\phi(y) \geq \inf_x \psi(x) \geq \phi(y^0) - ||d_s||.$$

So both ψ and ϕ are bounded from above and below.

Thus we can integrate $\phi,\ \psi$ against $\mu,\ \nu$ respectively, and, by the marginal condition,

$$\int \phi(y) d\nu(y) - \int \psi(x) d\mu(x) = \int \left[\phi(y) - \psi(x)\right] d\pi(x,y).$$

Since $\phi(y) - \psi(x) = \sum_s e_s^* d_{s^*}(x,y)$ on the support of π, we may write

$$\int \phi(y) d\nu(y) - \int \psi(x) d\mu(x) = \int \sum_s e_s^* d_{s^*}(x,y) d\pi(x,y).$$

It was shown in remark 43 on page 31 that $\phi = \psi^c = \psi$, so we can write

$$\int \psi(y) d\nu(y) - \int \psi(x) d\mu(x) = \int \sum_s e_s^* d_{s^*}(x,y) d\pi(x,y).$$

This proves the duality.

Step 5: If $d_s(x,y) \leq f_s(x) + g_s(y)$ then (ψ, ϕ) solves the dual problem.

The idea in this step is to prove that ψ and ϕ are integrable. The estimates in this step are similar to that of step 4, the difference being that we fix (x^0, y^0) such that

$\phi(y^0),\ \psi(x^0),\ f_s(x^0)$ and $g_s(y^0)$ are finite, and write

$$\begin{aligned}\psi(x) + f_{s^*}(x) &= \inf_{e \in \mathcal{E}} \sup_y [\phi(y) - e_{s^*} d_{s^*}(x, y) + f_{s^*}(x)] \\ &\geq \sup_y [\phi(y) - g_{s^*}(y)] \\ &\geq \phi(y^0) - g_{s^*}(y^0)\end{aligned}$$

and;

$$\begin{aligned}\psi(x) - g_{s^*}(y) &= \sup_{e \in \mathcal{E}} \inf_x [\psi(x) + e^*_{s^*} d_{s^*}(x, y) - g_{s^*}(y)] \\ &\geq \inf_x [\psi(x) + f_{s^*}(x)] \\ &\geq \phi(x^0) + f_{s^*}(x^0).\end{aligned}$$

So ψ is bounded below by the $\mu-$integrable function $\phi(y^0) - g_{s^*}(y^0) - f_{s^*}$ and ϕ is bounded above by the $\nu-$integrable function $\psi(x^0) + f_{s^*}(x^0) + g_{s^*}$; hence both $\int \psi d\mu$ and $\int \phi d\nu$ make sense in $\mathbb{R} \cup \{-\infty\}$. Further, both integrals are finite since $\int (\phi - \psi) d\pi = \int e^*_{s^*} d_{s^*} d\pi > -\infty$, and so

$$\int e^*_{s^*} d_{s^*} d\pi = \int \phi d\nu - \int \psi d\mu.$$

Hence, as a result of step 4 we can conclude that both π and (ψ, ϕ) are optimal in the primal and dual problems, respectively.

To prove the last part of the theorem, first note that d_s is continuous, so the subdifferential of any $e_s d_s-$convex function is a closed $(S, e_s d_s)-$cyclically monotone set.

Let π be an arbitrary optimal transmission method, and (ψ, ϕ) and an optimal coding strategy. we know that $(\psi, \psi^{e_s d_s})$ is optimal in the dual problem, so

$$\int e^*_{s^*} d_{s^*} d\pi = \int \psi^{e_s d_s} d\nu - \int \psi d\mu.$$

Now, by the marginal condition we may rewrite this as

$$\int \psi^{e_s d_s} - \psi - e^*_{s^*} d_{s^*} d\pi = 0,$$

and by remark 43 on page 31 we can write,

$$\int \psi(y) - \psi(x) - e^*_{s^*} d_{s^*} d\pi = 0$$

We know the integrand is non-negative so π must be concentrated on the pairs (x, y) for which

$$\psi(y) - \psi(x) - e^*_{s^*} d_{s^*}(x, y) = 0.$$

But this is the sub-differential of ψ, so since π and ψ are arbitrary, any optimal plan must be concentrated on the sub-differential of any optimal ψ. Thus, if Γ is defined as the intersection of all sub-differentials of optimal functions ψ, then Γ also contains the support of of all optimal plans.

For the converse, consider an arbitrary transfer plan $\tilde{\pi} \in \Pi(\mu, \nu)$ concentrated on Γ,then

$$\begin{aligned} \int e^*_{s^*} d_{s^*} d\tilde{\pi} &= \int \left[\psi^{e_s d_s} - \psi\right] d\tilde{\pi} \\ &= \int \psi^{e_s d_s} d\nu - \int \psi d\mu. \end{aligned}$$

So $\tilde{\pi}$ is optimal. Similarly, if $\tilde{\psi}$ is a $e_s d_s-$convex function such that $\partial_{e_s d_s} \tilde{\psi}$ contains Γ, then by the previous estimates $\tilde{\psi}$ and $\psi^{\widetilde{e_s d_s}}$ are integrable against μ and ν respectively, and

$$\begin{aligned}\int e^*_{s^*} d_{s^*} d\pi &= \int \left[\tilde{\psi}^{e_s d_s} - \tilde{\psi}\right] d\pi \\ &= \int \tilde{\psi}^{e_s d_s} d\nu - \int \tilde{\psi} d\mu \\ &= \int \tilde{\psi}(y) d\nu - \int \tilde{\psi}(x) d\mu\end{aligned}$$

We may conclude that $\tilde{\psi}(x)$ is optimal. □

5.6 Some Equivalent Statements About Optimal Transmission Methods.

The proof of the duality in the previous section would work just as well if d_s were merely a semi-metric on X_s, i.e. a function with all the properties of a metric but not necessarily the triangle inequality. However, in that case Steif's metric is no longer a metric[7] but a semi-metric. Regardless, in the next theorem that is all that is assumed.

Theorem 46. *If d_s is a real valued semi-metric on X_s for all $s \in S$ and $\bar{d}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \sup_{s \in S} \int d_s d\pi$ is finite, then there is a closed measurable $(S, e_s d_s)-$cyclically monotone set $\Gamma \subset \mathcal{X} \times \mathcal{X}$ such that for any $\pi \in \Pi(\mu, \nu)$ the following five statements are equivalent:*

1. *π is optimal*
2. *π is $(S, e_s d_s)-$cyclically monotone;*
3. *There is a $e_s d_s-$convex ψ such that, $\pi-$almost surely $\psi^{e_s d_s}(y) - \psi(x) = \sum_s e^*_s d_{s^*}$;*
4. *There exist $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $\phi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$, such that $\phi(y) - \psi(x) \leq \sum_s e^*_s d_{s^*}$ for all (x, y), with equality $\pi-$almost surely;*

[7] Which is equally true of the 1-Wassertstein distance, but the Kantorovich duality holds either way.

5. π is concentrated on Γ

Proof. Firstly assume $\bar{d}(\mu,\nu)$ is finite and $\forall s \in S,\ d_s \in \mathbb{R}$. We establish the truth of the above claim via the following sequence of implications

$$(1) \implies (2) \implies (3) \implies (4) \implies (1) \implies (5) \implies (2).$$

The implication $(1) \implies (3)$ must be done before $(1) \implies (5)$ as the former is used in the proof of the latter.

Setup: By step 4 of the proof of theorem 45 we can find $(\pi_k, \phi_k, \psi_k, (d_{s^*})_k)_{k\in\mathbb{N}}$ such that ψ_k is bounded and $e_s d_s-$convex, $\phi_k = (\psi_k)^c$, and

$$\int (e^*_{s^*} d_{s^*})_k d\pi_k(x,y) = \int \phi_k(y) d\nu(y) - \int \psi_k(x) d\mu(x). \tag{5.17}$$

$(1) \implies (2)$: Since the optimal transmission error is finite by assumption, the cost function $e^*_{s^*} d_{s^*}$ belongs to $L^1(\pi)$. From (5.17) and the marginal property of π,

$$\int \left[e^*_{s^*} d_{s^*}(x,y) - \phi_k(y) + \psi_k(x)\right] d\pi(x,y) \underset{k\to\infty}{\longrightarrow} 0,$$

so $e^*_{s^*} d_{s^*}(x,y) - \phi_k(y) + \psi_k(x)$ converges to 0 in $L^1(\pi)$ as $k \to \infty$. We may assume that up to a subsequence

$$\phi_k(y_i) - \psi_k(x_i) \underset{k\to\infty}{\longrightarrow} e^*_{s^*} d_{s^*}(x_i, y_i) \quad \pi(dx_i, dy_i) - a.s$$

By passing to the limit in the inequality

$$\begin{aligned}\sum_{i=1}^{M} e^{*}_{s^*} d_{s^*}(x_i, y_{i+1}) &\geq \sum_{i=1}^{M} \left[\phi_k(y_{i+1}) - \psi_k(x_i)\right] \\ &= \sum_{i=1}^{M} \left[\phi_k(y_i) - \psi_k(x_i)\right]\end{aligned}$$

(where by convention $y_{M+1} = y_1$)we see that, $\pi^{\otimes M} - a.s$

$$\sum_{i=1}^{M} e^{*}_{s^*} d_{s^*}(x_i, y_{i+1}) \geq \sum_{i=1}^{M} e^{*}_{s^*} d_{s^*}(x_i, y_i).$$

At this point we know that $\pi^{\otimes M}$ is concentrated on some set $\Gamma_M \subset (\mathcal{X} \times \mathcal{X})^M$, such that Γ_M consists of $M-$tuples $((x_1, y_1), ..., (x_M, y_M))$ satisfying (5.2). Let $\mathrm{proj}_k((x_i, y_i)_{1\leq i\leq M}) := (x_k, y_k)$ be the projection onto the k^{th} factor of $(\mathcal{X} \times X)^M$. *One can check that* $\Gamma := \cap_{1\leq k\leq M} \mathit{proj}\ _k(\Gamma_M)$ *is a* $(S, e_s d_s)-$*cyclically monotone set which has full* $\pi-$*measure; so* π *is indeed* $(S, e_s d_s)-$*cyclically monotone.*

(2) $\implies$ (3) : Let π be a cyclically monotone transmission method. The function ψ can be constructed just as in Step 3 of the proof of duality - but with some key differences. First, Γ is not necessarily closed; it is just a Borel set such that $\pi[\Gamma] = 1$. (If it is not Borel we can make it Borel by modifying on a negligible set.) Now define as in step 3 of the proof of theorem 45,

$$\begin{aligned}\psi(x) := \inf_{e\in\mathcal{E}} \sup_{m\in\mathbb{N}} \sup\{&\left[e_{s^*} d_{s^*}(x^0, y^0) - e_{s^*} d_{s^*}(x^1, y^0)\right] + \left[e_{s^*} d_{s^*}(x^1, y^1) - e_{s^*} d_{s^*}(x^2, y^1)\right] + \\ &\cdots + \left[e_{s^*} d_{s^*}(x^m, y^m) - e_{s^*} d_{s^*}(x, y^m)\right]; \quad \left(x^0, y^0\right), \ldots, (x^m, y^m) \in \Gamma\}.\end{aligned}$$

From its definition, for any $x \in \mathcal{X}$,

$$\psi(x) \geq e^{*}_{s^*} d_{s^*}(x^0, y^0) - e^{*}_{s^*} d_{s^*}(x, y^0) > -\infty.$$

Then we proceed just as in step 3, showing that $\psi(x^0) = 0$, ψ is $e_s d_s-$convex and π is concentrated on $\partial_{e_s d_s}\psi$.

(3) $\implies$ (4) : Just let $\phi = \psi^{e_s d_s}$.

(4) $\implies$ (1) : Let (ψ, ϕ) be a pair of admissible functions, and let π be a transmission method such that $\phi - \psi = \sum_{s\in S} e_s^* d_{s^*}$, $\pi-$almost surely. The goal is to show that π is optimal. The main difficulty lies in the fact that ψ and ϕ need not be separately integrable. This problem will be circumvented by a careful truncation procedure. For $n \in \mathbb{N}$, $w \in \mathbb{R}_{\geq 0} \cup \{+\infty\}$, define

$$T_n(w) = \begin{cases} w & \text{if } w \leq n \\ n & \text{if } w > n, \end{cases}$$

and

$$\xi(x, y) := \phi(y) - \psi(x); \qquad \xi_n(x, y) := T_n(\phi(y)) - T_n(\psi(x)).$$

In particular, $\xi_0 = 0$. It is easily checked that ξ_n converges monotonically to ξ; more precisely,

- $\xi_n(x, y)$ remains equal to 0 if $\xi(x, y) = 0$;
- $\xi_n(x, y)$ increases to $\xi(x, y)$ if the latter quantity is positive;

As a consequence, $\xi_n \leq (\xi_n)_+ \leq \xi_+ \leq \sum_{s\in S} e_s^* d_{s^*}$. So $(T_n\phi, T_n\psi)$ is a plausible strategy in the dual problem, and

$$\int \xi_n d\pi = \int (T_n\phi) d\nu - \int (T_n\psi) d\mu \leq \sup_{\phi' - \psi' \leq e_s d_s} \left(\int \phi' d\mu - \int \psi' d\nu \right). \tag{5.18}$$

On the other hand, by monotone convergence and since ξ coincides with outside of a $\pi-$negligible set,

$$\int_{\xi\geq 0} \xi_n d\pi \xrightarrow[n\to\infty]{} \int_{\xi\geq 0} \xi d\pi = \int \sum_{s\in S} e_s^* d_{s^*} d\pi;$$

This and (5.18) imply that

$$\int \sum_{s\in S} e_s^* d_{s^*} d\pi \leq \sup_{\phi'-\psi'\leq e_s d_s} \left(\int \phi' d\mu - \int \psi' d\nu \right);$$

So π is optimal. Before completing the chain of equivalences, we should first construct the set Γ. By theorem 4.1 there is at least one optimal transmission method, say $\tilde{\pi}$. By the implication (1) $\implies$ (3), there is some $\tilde{\psi}$ such that $\tilde{\pi}$ is concentrated on $\partial_{e_s d_s}\tilde{\psi}$; just choose $\Gamma := \partial_{e_s d_s}\tilde{\psi}$.

(1) $\implies$ (5) : Let $\tilde{\pi}$ be the optimal plan used to construct Γ, and let $\psi = \tilde{\psi}$ be associated $e_s d_s-$convex function; let $\phi = \psi^c$. Then let π be another optimal plan. Since π and $\tilde{\pi}$ have the same cost and same marginals,

$$\begin{aligned} \int \sum_{s\in S} e_s^* d_{s^*} d\pi = \int \sum_{s\in S} e_s^* d_{s^*} d\tilde{\pi} &= \lim_{n\to\infty} \int (T_n\phi - T_n\psi) d\tilde{\pi} \\ &= \lim_{n\to\infty} \int (T_n\phi - T_n\psi) d\pi, \end{aligned}$$

where T_n is the truncation operator that was used in the proof of (4) $\implies$ (1). So

$$\int \left[\sum_{s\in S} e_s^* d_{s^*}(x,y) - T_n\phi(y) + T_n\psi(x) \right] d\pi(x,y) \xrightarrow[n\to\infty]{} 0. \tag{5.19}$$

As before, define $\xi(x,y) := \phi(y) - \psi(x)$; then by monotone convergence,

$$\int_{\xi\geq 0} \left[\sum_{s\in S} e_s^* d_{s^*} - T_n\phi + T_n\psi \right] d\pi \xrightarrow[n\to\infty]{} \int_{\xi\geq 0} (\sum_{s\in S} e_s^* d_{s^*} - \xi) d\pi.$$

Since $\xi \leq \sum_{s\in S} e_s^* d_{s^*}$, the integrands here are nonnegative and both integrals make sense in $[0, +\infty]$. So by adding the two limits and using (5.19) we get

$$\int (\sum_{s\in S} e_s^* d_{s^*} - \xi) d\pi = \lim_{n\to\infty} \int \left[\sum_{s\in S} e_s^* d_{s^*} - T_n\phi + T_n\psi \right] = 0.$$

Since $\xi \leq \sum_{s\in S} e_s^* d_{s^*}$, this proves that $\sum_{s\in S} e_s^* d_{s^*}$ coincides $\pi-$almost surely with ξ, which was the desired conclusion.

$(5) \implies (2)$: This is obvious since Γ is cyclically monotone by assumption. □

6 Conclusion

In this dissertation I have proven the equality of two metrics, extending on work by R.S. MacKay et al in (MacKay & Diakonova, 2011), (MacKay, 2011) and (MacKay, 2019). The proof was adapted from that of the Kantorovich duality in (Villani, 2009). I hope that the result will be useful in the classification of complex systems in terms of the parameters corresponding to their metastable states. It remains quite difficult to actually calculate the distances on complex systems. To this end I think the place to start is with algorithms for calculating maximal multi-commodity flows, for example in (Williamson, 2019).

7 Bibliographic Notes

The section I wrote to provide context was made up of three sections. The part on IPS was taken from the introduction in (Ligget, 2005) and (Toom et. al., 1990). The description of PCA was mainly based on (Toom et. al., 1990) and (Pierre-Yves, Nardi and Fernandez, 2018). Finally the subsection on complexity and emergence was based on (Hoekstra, Kroc and Sloot, 2013) along with (MacKay & Diakonova, 2011). the latter being primarily for information on emergence and the former on complexity.

In the next section the mathematical description of PCA is taken from (Pierre-Yves, Nardi and Fernandez, 2018) and the NEC voter model PCA is as described in (MacKay & Diakonova, 2011). Remark 10, which describes the product $\sigma-$algebra is from (Folland, 2009) while Dobrushin's metric was introduced by R. MacKay in (MacKay, 2011). Steif's metric is an extension by R. MacKay in (MacKay, 2018) of a metric introduced by Steif in (Steif, 1988) and (Steif, 1991). The calculation of Dobrushin's metric is from (MacKay & Diakonova, 2011).

All of section four is from chapter 5 of (Villani, 2009).

In section six I follow Villani's proofs quite closely, I had to include Theorem 29 from (Cohn, 1980) to ensure certain parts worked. Proving that the optimal coupling was a fairly trivial extension of proofs in (Villani, 2009). The definition I supply of cyclical monotonicity is a non-trivial extension of the one Villani provides as is the definition e_sd_s-convexity. Lemma 36 is the combination of a proof sketched for me by Robert MacKay in a meeting in semester 2 of 2019/2020 and a proof in (Follmer & Horst, 2001). The illustration of the optimisation problem based on information theory is mine, but it was inspired by applications discussed in (Dobrushin, 1972). The proof of the duality follows Villani's proof of the Kantorovich duality in (Villani, 2009) but uses the extended definitions I introduced. The first three steps in particular required significant extensions, much of the rest however seemed to follow naturally after that.

Finally the proof, in the appendix, that Steif's metric is complete on $\mathcal{P}(\mathcal{X})$ is based on the one in (Steif, 1988). I had to extend it slightly, especially Lemma 51 on page 54 which I had to extend from $\{0,1\}$ to all compact Polish spaces.

8 Appendix

8.1 $\bar{d}(\mu, \nu)$ is a complete metric on $\mathcal{P}(\mathcal{X})$.

Remark 47 (Notation). Denote by Δ the symmetric difference of two sets i.e. the set of elements which are in the union of the sets but not in their intersection.

Proposition 48. *$\bar{d}(\mu, \nu)$ is a metric*

To prove proposition 48 we will need the gluing lemma. For a proof one can consult (Dudley, 1999; Theorem 1.1.10), but the proof goes too far off track, involves stochastic processes and requires new definitions not really relevant to anything else I discuss.

Lemma 49. *Let $(\mathcal{X}_i, \mu_i)$, $i = 1, 2, 3$, be Polish probability spaces. If (X_1, X_2) is a coupling of (μ_1, μ_2) and (X_2, X_3) is a coupling of (μ_2, μ_3), then one can construct a triple of random variables (Z_1, Z_2, Z_3) such that (Z_1, Z_2) has the same law as (X_1, X_2) and (Z_2, Z_3) has the same law as (X_2, X_3).*

Proof of proposition. We must prove symmetry (1), identity of indiscernible (2) and the triangle inequality (3).

1. Symmetry is obvious

2. If $\bar{d}(\mu, \nu) = 0$ then for all $s \in S$ we have $\int_{X_s \times X_s} d_s(x_s, y_s) d\pi^* = 0$ which implies that $x = y \ \pi^* - a.s$ so $\mathbb{P}^{\pi^*}(x = y) = 1$ which implies that $\mu = \nu$. Conversely if $\mu = \nu$ we can trivially couple them with the identity map, then $x = y \ \pi - a.s$ and hence $\bar{d}(\mu, \nu) = 0$.

3. Let $(\mathcal{X}_i, \mu_i)$, $i = 1, 2, 3$, be Polish probability spaces. Let X_1, X_2, X_3 and (Z_1, Z_2, Z_3) be as in Lemma 49. Choose $\text{Law}(X_1, X_2) = \pi_1$ such that

$$\sup_{s \in S} \int_{\mathcal{X}_1 \times \mathcal{X}_2} d_s(x, y) d\pi_1 \leq \bar{d}(\mu_1, \mu_2) + \epsilon$$

and likewise choose Law$(X_2, X_3) = \pi_2$ such that

$$\sup_{s\in S} \int_{\mathcal{X}_2\times\mathcal{X}_3} d_s(x,y)d\pi_2 \le \bar{d}(\mu_2,\mu_3) + \epsilon.$$

By Lemma 49 we can choose a measure $\tilde{\pi}$ on $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ whose projection onto the first two factors is π_1 and whose projection onto the second two factors is π_2. Let π_3 be the projection of $\tilde{\pi}$ onto the first and third factors then π_3 is a (μ_1, μ_3) coupling and

$$\begin{aligned}
\int_{\mathcal{X}_1\times\mathcal{X}_3} d_s(x,z)d\pi_3 &= \int_{\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3} d_s(x,z)d\tilde{\pi} \\
&\le \int_{\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3} d_s(x,y) + d_s(y,z)d\tilde{\pi} \\
&\le \int_{\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3} d_s(x,y)d\tilde{\pi} + \int_{\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3} d_s(y,z)d\tilde{\pi} \\
&= \int_{\mathcal{X}_1\times\mathcal{X}_2} d_s(x,y)d\pi_1 + \int_{\mathcal{X}_2\times\mathcal{X}_3} d_s(y,z)d\pi_2
\end{aligned}$$

Let $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3$ then we have

$$\begin{aligned}
\bar{d}(\mu_1,\mu_3) &= \sup_{s\in S} \int_{\mathcal{X}\times\mathcal{X}} d_s(x,z)d\pi_3 \\
&\le \sup_{s\in S} \left(\int_{\mathcal{X}\times\mathcal{X}} d_s(x,y)d\pi_1 + \int_{\mathcal{X}\times\mathcal{X}} d_s(y,z)d\pi_2 \right) \\
&\le \sup_{s\in S} \int_{\mathcal{X}\times\mathcal{X}} d_s(x,y)d\pi_1 + \sup_{s\in S} \int_{\mathcal{X}\times\mathcal{X}} x_s d_s(y,z)d\pi_2 \\
&= \bar{d}(\mu_1,\mu_2) + \bar{d}(\mu_2,\mu_3) + 2\epsilon.
\end{aligned}$$

Since ϵ can be taken to be arbitrarily small this gives the result.

□

Proposition 50. *$\bar{d}(\mu,\nu)$ is complete.*

Before proving Proposition 28 we need the following Lemma.

Lemma 51. *There exists a metric ρ on $\mathcal{P}(\mathcal{X})$ such that $\rho \leq \bar{d}$ and the topology induced from ρ is the weak* topology on $\mathcal{P}(\mathcal{X})$.*

Proof. Let $S = \{s_1, s_2, \dots\}$ and let (X_s, d_s) be compact and Polish. Let $\sigma(s_1, \dots, s_N)$ be the sub $\sigma-$field of $\mathcal{X}$ generated by the coordinates $\{s_1, \dots, s_N\}$. Now define ρ in the following way

$$\rho(\mu, \nu) = \sum_{i=1}^{\infty} \frac{1}{i(2K)^i} \sum_{F \in \mathcal{A}(\sigma(s_1, \dots, s_i))} |\mu(F) - \nu(F)|.$$

Where $\mathcal{A}(\sigma(s_1, ..., s_i))$ is the collection of subsets corresponding to the sub $\sigma-$field $\sigma(s_1, ..., s_i)$. By Tychnoff's theorem $\mathcal{X}$ is compact and every open covering of $\mathcal{X}$ has a finite subcover. This implies that the Borel $\sigma-$algebra $\sigma(\mathcal{X})$ is finite. Hence, the second sum is finite since the cardinality of $\mathcal{A}$ is finite. If $\sup_s |\sigma(X_s)| = K$ then the second term is bounded above by K^i. It's obvious that ρ is a metric, we prove now that ρ corresponds to the weak* topology. Assume that $\mu_n \underset{w^*}{\longrightarrow} \mu$. Let $\epsilon > 0$ be given and choose M so that

$$\sum_{i=M+1}^{\infty} \frac{1}{i2^i} \leq \frac{\epsilon}{2}.$$

Next, choose N such that for all $n \geq N$, μ_n and μ agree to within $\frac{\epsilon}{2}$ on all sets in the $\sigma-$field generated by $\{s_1, ..., s_M\}$. Then if $n \geq N$

$$\begin{aligned}
\rho(\mu_n, \mu) &= \sum_{i=1}^{M} \frac{1}{i(2K)^i} \sum_{F \in \mathcal{A}(\sigma(s_1, \dots s_i))} |\mu_n(F) - \mu(F)| + \sum_{i=1+M}^{\infty} \frac{1}{i(2K)^i} \sum_{F \in \mathcal{A}(\sigma(s_1, \dots s_i))} |\mu_n(F) - \mu(F)| \\
&\leq \sum_{i=1}^{M} \frac{1}{i2^i} \frac{\epsilon}{2} + \sum_{i=M+1}^{\infty} \frac{1}{i2^i} \\
&< \epsilon
\end{aligned}$$

so $\mu_n \underset{\rho}{\rightarrow} \mu$.

Conversely, assume $\mu_n \underset{\rho}{\rightarrow} \mu$. Let F be an arbitrary element of the Borel $\sigma-$algebra. Then

$$F \in \mathcal{A}(\sigma(s_i, ..., s_i))$$

for some i sufficiently large. Choose N sufficiently large so that for all $n \geq N$, $\rho(\mu_n, \mu) < \frac{\epsilon}{i(2K)^i}$. Then for all $n \geq N$,

$$\frac{|\mu_n(F) - \mu(F)|}{i(2K)^i} \leq \rho(\mu_n, \mu) \leq \frac{\epsilon}{i(2K)^i}$$

and so $|\mu(F) - \mu_n(F)| \leq \epsilon$ and $\mu_n \underset{w^*}{\rightarrow} \mu$. Hence the weak topology and the weak$^*-$topology coincide. We need to show that $\rho \leq \bar{d}$. Let π be an arbitrary μ, ν coupling. It suffices to show that $\rho(\mu, \nu) \leq \sup_s \int d_s(x, y) d\pi$.

$$\begin{aligned}
\rho(\mu, \nu) &= \sum_{i=1}^{\infty} \frac{1}{i(2K)^i} \sum_{F \in \mathcal{A}(\sigma(s_1, \ldots, s_i))} |\int_F f d\mu - \int_F f d\nu| \\
&= \sum_{i=1}^{\infty} \frac{1}{i(2K)^i} \sum_{F \in \mathcal{A}(\sigma(s_1, \ldots, s_i))} |\int_{F \times \mathcal{X}} d_s(x, y) d\pi - \int_{\mathcal{X} \times F} d_s(x, y) d\pi| \\
&\leq \sum_{i=1}^{\infty} \frac{1}{i(2K)^i} \sum_{F \in \mathcal{A}(\sigma(s_1, \ldots, s_i))} \int_{(F \times \mathcal{X}) \Delta (\mathcal{X} \times F)} d_s(x, y) d\pi \\
&\leq \sum_{i=1}^{\infty} \frac{1}{i(2K)^i} \sum_{F \in \mathcal{A}(\sigma(s_1, \ldots, s_i))} \sum_{j=1}^{i} \int_{\mathcal{X} \times \mathcal{X}} d_s(x, y) d\pi \\
&\leq \sum_{i=1}^{\infty} \frac{1}{2^i} \int_{\mathcal{X} \times \mathcal{X}} d_s(x, y) d\pi \\
&= \int_{\mathcal{X} \times \mathcal{X}} d_s(x, y) d\pi \\
&\leq \sup_{s \in S} \int_{\mathcal{X} \times \mathcal{X}} d_s(x, y) d\pi
\end{aligned}$$

as required. □

Proof of Proposition 28. Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a $\bar{d}-$Cauchy sequence. By the previous lemma it is also $\rho-$Cauchy and hence converges weakly to some μ. We show that $\mu_n \underset{\bar{d}}{\longrightarrow} \mu$. To do this consider a coupling of μ_n and μ_m, say $\pi_{n,m}$, such that

$$\sup_s \int d_s(x,y)d\pi_{n,m} \leq 2\bar{d}(\mu_n,\mu_m).$$

We can use a standard diagonalization argument and the weak* compactness of the collection of probability measures (since $\mathcal{X}\times\mathcal{X}$ is compact) we can choose $m_k \to \infty$ such that, for all n

$$\pi_{n,m_k} \underset{w^*}{\longrightarrow} \pi_n \in \mathcal{P}(\mathcal{X}\times\mathcal{X}),\ k\to\infty.$$

This is a coupling of μ_n, μ and we now want to show that these $\{\pi_n\}_{n\in\mathbb{N}}$ yield good μ_n, μ couplings.

Let $\epsilon > 0$ and choose N such that $\forall n, m \geq N,\ \bar{d}(\mu_n,\mu_m) \leq \epsilon$. Now, if $n, m_k \geq N$ and $s \in S$,

$$\begin{aligned}\int d_s(x,y)d\pi_{n,m_k} &\leq \sup_s \int d_s(x,y)d\pi_{n,m_k}\\ &\leq 2\bar{d}(\mu_n,\mu_{m_k})\\ &< 2\epsilon.\end{aligned}$$

Letting $k\to\infty$ together with $\pi_{n,m_k} \underset{w^*}{\longrightarrow} \pi_n$ and compactness of $\mathcal{X}\times\mathcal{X}$ we get that

$$\sup_s \int d_s(x,y)d\pi_n \leq 2\epsilon$$

and we can conclude that $\bar{d}(\mu_n,\mu) \leq 2\epsilon$ if $n \geq N$ and so $\mu_n \underset{\bar{d}}{\longrightarrow} \mu$ as required. □

References

[1] Cohn, Donald. L. “Chapter 8: Polish Spaces and Analytic Sets.” In Measure Theory .., 251–54. Berlin: Springer, 1980.

[2] Diakonova, M., and R. S. Mackay. “Mathematical Examples Of Space-Time Phases.” International Journal of Bifurcation and Chaos 21, no. 08 (2011): 2297–2304. https://doi.org/10.1142/s0218127411029793.

[2] Dobrushin, R. L., Kriukov V. I., A. L. Toom, N. B. Vasilyev, O. N. Stavskaya, L. G. Mityushin, G. L. Kurdyumov, and S. A. Pirogov. “Part 1: Discrete Local Markov Systems.” In Stochastic Cellular Systems: Ergodicity, Memory, Morphogenesis, 1–33. Manchester: Manchester University Press, 1990.

[3] Dobrushin, R. “Survey of Soviet Research in Information Theory.” IEEE Transactions on Information Theory 18, no. 6 (1972): 703–24. https://doi.org/10.1109/tit.1972.1054923.

[4] Dudley, R. M. “1.1 Empirical Processes: the Classical Case.” Essay. In Uniform Central Limit Theorems, 7–9. Cambridge: Cambridge University Press, 1999.

[5] Dudley, R. M. “Chapter 11: Convergence Laws on Separable Metric Spaces .” Essay. In Real Analysis and Probability, 399–405. Cambridge: Cambridge University Press, 2002.

[6] Folland, Gerald B. “Chapter 2: Measure and Integration - A General Theory.” Essay. In A Guide to Advanced Real Analysis, 21–40. Washington, D.C.: Mathematical Association of America, 2009.

[7] Föllmer, Hans, and Ulrich . “Convergence of Locally and Globally Interacting Markov Chains.” Stochastic Processes and their Applications 96, no. 1 (2001): 99–121. https://doi.org/10.1016/s0304-4149(01)00110-7.

[8] Hoekstra, Alfons G., Jiri Kroc, and Peter M. A. Sloot. “Chapter 1: Introduction to Modelling of Complex Systems Using Cellular Automata.” In Simulating Complex Systems by Cellular Automata, 1–16. Berlin: Springer Berlin, 2013.

[9] Liggett, T.M. Interacting Particle Systems. Berlin: Springer, 2005.

[10] Louis, Pierre-Yves, Francesca R. Nardi, and Roberto . “Chapter 1: Overview - PCA Models and Issues .” In Probabilistic Cellular Automata: Theory, Applications and Future Perspectives, 1–30. Cham, Switzerland: Springer, 2018.

[11] Mackay, R S. “Management of Complex Dynamical Systems.” Nonlinearity 31, no. 2 (2018). https://doi.org/10.1088/1361-6544/aa952d.

[12] Mackay, R. S. “Robustness of Markov Processes on Large Networks.” Journal of Difference Equations and Applications 17, no. 8 (2011): 1155–67. https://doi.org/10.1080/10236190902976889.

[13] Maes, Christian. “Coupling Interacting Particle Systems.” Reviews in Mathematical Physics 05, no. 03 (1993): 457–75. https://doi.org/10.1142/s0129055x93000139.

[14] Steif, Jeffrey E. “Convergence to Equilibrium and Space—Time Bernoullicity for Spin Systems in the M< ϵ Case.” Ergodic Theory and Dynamical Systems 11, no. 3 (1991): 547–75. https://doi.org/10.1017/s0143385700006337.

[15] Steif, Jeffrey E. “The Ergodic Structure of Interacting Particle Systems,” 1988.

[16] Tsetlin, M L. “Finite Automata And Models Of Simple Forms Of Behaviour.” Russian Mathematical Surveys 18, no. 4 (1963): 1–27. https://doi.org/10.1070/rm1963v018n04abeh001139.

[17] Villani Cédric. “Chapters 1-6.” In Optimal Transport: Old and New, 1–113. Berlin: Springer, 2009.

[18] Williamson, David P. “Chapter 7: Multicommodity Flow Algorithms.” In Network Flow Algorithms. Cambridge: Cambridge University Press, 2019.