Risk Measurement under Dependence Structure Ambiguity

by

Harris Chen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Quantitative Finance

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, we work on a generalization of the entropy regularized optimal transport problem, with the objective function being (spectral) risk measures. We accomplish three goals: to present the corresponding dual problem and prove Kantorovich duality, to prove stability of the optimal value under the weak convergence of marginals, the reference measure and the regularization threshold, and to explore an efficient numerical algorithm for a solution of the optimization problem.

The analogue of the Kantorovich duality is proved using techniques from convex analysis. Stability and convergence of approximating optimization problems are studied using the techniques of Gamma convergence, combined with recent results on shadow couplings. For the numerical solution of the optimization problem, a variation on Sinkhorn's algorithm is developed, which improves on a naive linear programming implementation significantly, in terms of both running time and storage requirements.

Acknowledgements

I would like to thank my supervisor and my teacher, professor David Saunders, for his guidance not only on this thesis, but also on the precious lessons from which I gradually learn how to be a researcher. Without his mentorship, I could not have gone this far.

I would like to thank my teacher, professor Mario Ghoussoub, for all his support and encouragement for me to do research work throughout my past years. We have many constructive conversations which I will always remember.

Also I would like to thank professor Serge D'Alessio who I met on my first calculus course here at University of Waterloo. His advice of the mathematical finance program and the mentioning of the MQF program opens a window for me to explore more aspects of mathematics.

I would like to thank professor David Saunders, Mario Ghoussoub, Yi Shen, Alexander Schied, Christiane Lemieux, Alexandru Nica, Stephen New, Ricardo Fukasawa, etc. for their excellent and interesting courses on mathematics that inspire many passionate math students like me.

Most importantly I would like to thank my families who respect my decisions I make for myself and who are always their for me. I would like to in particular thank my aunt Linda who supports me along the way.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

The topic of optimal transport has been explored during the past decades, which originates from the study of optimization in areas of resources allocation, route planning, etc. Recently the topic of entropy regularized optimal transportation draws the interest of academics, as well as the connection to machine learning. In this thesis, we combine the theory of (entropy regularized) optimal transportation and the spectral risk measure, which could serve for the general interest in risk measurement, and contribute to both the theoretical and computational aspect of the problem.

1.1.1 Risk Measures

A risk measure R maps a subset $\mathcal{R}$ of the set of random variables on some probability space $(\Omega, \mathcal{F}, \pi)$, interpreted as the set of discounted net financial positions, into real numbers. Coherent risk measures have been particularly well-studied (see, e.g. Artzner et al. (1999), Delbaen (2002), Riedel (2004)), and are characterized by the following properties:

- Monotonicity: $R(X) \leq R(Y)$, for all $X, Y \in \mathcal{R}$ such that $X \leq Y$, π-a.s.
- Positive Homogeneity: $R(\lambda X) = \lambda R(X)$, for all $X \in \mathcal{R}$ and all $\lambda \in \mathbb{R}_+$
- Cash Invariance: $R(X + c) = R(X) + c$, for all $X \in \mathcal{R}$ and $c \in \mathbb{R}$
- Subadditivity: $R(X + Y) \leq R(X) + R(Y)$ for all $X, Y \in \mathcal{R}$

One coherent risk measure (see Artzner et al. (1999)), Expected Shortfall (ES), also known as Conditional Value-at-Risk (CVaR), is often used in the area of quantitative risk management in the banking and insurance industries (see Basel Committee on Banking Supervision (2019)).

For continuous loss variables, it is the expected loss given that the losses exceed a prescribed quantile. Let $\mathcal{W}$ be a Polish space, let $\mathcal{P}(\mathcal{W})$ be the set of Borel probability measures on $\mathcal{W}$. For a confidence level $\alpha \in (0,1)$, a probability measure $\pi \in \mathcal{P}(\mathcal{W})$, and a bounded random variable $L : \mathcal{W} \to \mathbb{R}$, the Expected Shortfall (ES) of L is defined to be:

$$ES_{\alpha,\pi}(L) = (1-\alpha)^{-1} \int_{\alpha}^{1} F^{\leftarrow}_{L,\pi}(u)\, du,$$

where $F^{\leftarrow}_{L,\pi}$ is the generalized inverse of the cumulative distribution function of the random variable L (given the probability measure π on $\mathcal{W}$):

$$F^{\leftarrow}_{L,\pi}(u) = \inf\left\{x \in \mathbb{R} : \pi(L^{-1}(-\infty, x]) \geq u\right\}.$$

1.1.2 Optimal Transportation Theory

Let's begin with an example of the optimal transportation theory: consider the problem of matching a number of workers and firms. Each worker can only work for one firm, and each firm can only hire one worker. A matching between all workers and all firms produces an economic utility. A natural question to ask is does there exist an optimal matching such that the utility coming out of the matching is maximized; and if so, what can we conclude for the optimal matching and the corresponding optimal value, etc.

There is a large literature on the optimal transport problem and its applications (e.g., Rachev and Rüschendorf (1998) or Villani (2003) or Villani (2008), and the referneces therein). The applications to economics are discussed in Galichon (2016) and to risk measures in Rüschendorf (2013). Optimal transport, and the related martingale optimal transport problem have also been applied to the problem of determining bounds for prices of financial instruments (e.g., Beiglböck, Henry-Labordère, and Penkner (2013) or Henry-Labordère (2017)), and on risk measures for particular choices of the function L (see McNeil, Frey, and Embrechts (2015)).

Here we give a basic introduction to the optimal transport problem. For more details, see Galichon (2016) and Villani (2003).

Given two measure spaces $\mathcal{X}$ and $\mathcal{Y}$, let $\mu \in \mathcal{X}$ represent the vectors of characteristics of workers, and let $\nu \in \mathcal{Y}$ represent the vectors of characteristics of firms,

and we can assume that the workers and firms are in equal mass, so we may normalize μ and ν to 1, hence we can let μ and ν representing probability measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, the cost function c acting on (X, Y) gives the cost/profit of assigning worker X to firm Y. Any possible pairing (X, Y), called a coupling, can be represented as an element in $\Pi(\mu, \nu)$, where $\Pi(\mu, \nu)$ is defined to be the set of all possible joint distribution with fixed marginals μ and ν.

The Monge problem assumes that every worker of type x get assigned to the same type of firm, hence the coupling (X, Y) becomes deterministic in terms of X, i.e. $Y = T(X)$ where we call T in this case a pure assignment. The Monge problem reads:

$$\max_{\{T\colon X \to Y \mid T\#\mu = \nu\}} \mathbb{E}_\mu(c(X, T(X)))$$

Important advances have been made by the mathematician and economist Leonid Kantorovich. A key approach to the Monge problem is to use the idea of relaxation: rather than optimizing over the pure assignments, optimizing over any assignment (or over all possible joint distributions):

$$\max_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi(c(X, Y))$$

The following result is fundamental to the optimal transport.

Theorem 1.1.1 (Kantorovich duality)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces, let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{+\infty\}$ be a lower semi-continuous cost function.*

Whenever $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and $(\varphi, \psi) \in L^1(d\mu) \times L^1(d\nu)$, define

$$I(\pi) = \int c\, d\pi, \quad J(\varphi, \psi) = \int \varphi\, d\mu + \int \psi\, d\nu.$$

Define $\Pi(\mu, \nu)$ to be the set of all Borel probability measures π on $\mathcal{X} \times \mathcal{Y}$ such that for all measurable subsets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$,

$$\pi(A \times \mathcal{Y}) = \mu(A), \quad \pi(\mathcal{X} \times B) = \nu(B),$$

and define Φ_c to be the set of all measurable functions $(\varphi, \psi) \in L^1(d\mu) \times L^1(d\nu)$ satisfying

$$\varphi(x) + \psi(y) \leq c(x, y)$$

for $d\mu$-almost all $x \in \mathcal{X}$, $d\nu$-almost all $y \in \mathcal{Y}$.

Then

$$\inf_{\Pi(\mu,\nu)} I(\pi) = \sup_{\Phi_c} J(\varphi, \psi) \tag{1.1.1}$$

Moreover, the infimum in the left-hand side of (1.1.1) *is attained. Furthermore, it does not change the value of the supremum in the right-hand side of* (1.1.1) *if one restricts the definition of* Φ_c *to those functions* (φ, ψ) *which are bounded and continuous.*

Proof. See Villani (2003) for more details. □

1.1.3 Entropy Regularization

Recent literature (e.g. Glasserman and Yang (2018) or Nutz and Wiesel (2021) or Genevay (2019)) combines the original optimal transport problem with an entropy regularization term. One meaningful interpretation among others is that the optimizer should not deviate from our prior experience (which is represented as a reference measure) to a large extent, and we want to control the distance between the two measures.

A common choice for the regularization term is the relative entropy, or Kullback–Leibler divergence, which is used in economics, for example in areas of max-min expected utility theory and robust control theory (see e.g. Hansen and Sargent (2001) and Hansen et al. (2006)), and more recent literature on fast numerical algorithms and solvers (e.g. Mérigot and Thibert (2021) and Tenetov, Wolansky, and Kimmel (2018)). Given a reference measure π_{ref}, the relative entropy is given by:

$$\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi$$

where $\pi \ll \pi_{\text{ref}}$.

However other choices for the entropy on the probability space are possible to study. In the terminology of economics and decision theory, Maccheroni, Marinacci, and Rustichini (2006) characterizes the KL-divergence as one of the multiplier preferences (in the language of Hansen and Sargent, see Hansen and Sargent (2001)) which in turn belong to a bigger class of preferences, called divergence preferences that are in general smooth, which is an important feature for applications. The class of divergence preferences also include a third type of preferences, called the mean-variance preferences of Markowitz (Markowitz (1952)) and Tobin (Tobin (1958)). An example of mean-variance preferences is the relative Gini

concentration index (or χ^2 divergence) given by

$$\int_{\mathcal{X}\times\mathcal{Y}} 2^{-1}\left(\frac{d\pi}{d\pi_{\text{ref}}}-1\right)^2 d\pi$$

where $\pi \ll \pi_{\text{ref}}$.

Given a reference measure π_{ref}, the (relative) entropy regularized optimal transport is:

$$\max_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c\, d\pi - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi \tag{1.1.2}$$

where the problem is defined for all $\pi \ll \pi_{\text{ref}}$, and c_0 is the regularization parameter in $\mathbb{R}$.

One benefit for adding a regularization term comes from the numerical aspect, such as smoothness, existence of a gradient in gradient descent methods, and an improved sampling complexity (see e.g. Genevay et al. (2016) and Cuturi, Teboul, and Vert (2019)). In particular, the entropic regularization, as in our case, combined with the Sinkhorn's algorithm could be implemented at large scale using parallel computing (see Sinkhorn and Knopp (1967) and Knight (2008)).

For large $c_0 > 0$, the penalty term in (1.1.2) is negligible, and solving (1.1.2) is essentially approximating the unregularized version of the problem by empowering the benefit of numerical algorithms (see e.g. Weed (2018) and Altschuler, Weed, and Stromme (2021)). On the other hand, for small $c_0 > 0$, the penalty term is not negligible, and the problem (1.2.3) itself is of its own interest due to its interpretation in measuring the divergence of distribution, as is in our case. The magnitude of c_0 can also reflect the degree of ambiguity aversion in economics and decision theory (see Maccheroni, Marinacci, and Rustichini (2006)).

1.2 Problem Formulation

The original motivation for the formulation of the problem arose from applications in counterparty credit risk, in which the two factors correspond to sets of factors for market risk and credit risk (e.g., Garcia-Cespedes et al. (2010) and Rosen and Saunders (2012)). Bounding the Credit Valuation Adjustment (CVA) (i.e. the price of counterparty credit risk losses) given known market and credit risk factor distributions assumes the form of an optimal transport problem: the distribution of the exposures and the distribution of the counterparty's default time are treated as given marginals, and the expected credit loss for a counterparty portfolio is a function that depends on both marginals, which provides another way to build the correlation between these factors in addition to the copula approach.

Memartoluie, Saunders, and Wirjanto (2012) considered the problem of bounding ES given the distributions of the market and credit factors, and showed that in the case of finite sample spaces, the problem is equivalent to a linear program. Memartoluie (2017) further investigated into the issues on the bounds on CVA contributions with given marginals, counterparty credit risk and bounds on ES.

Throughout the thesis, let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish (i.e., complete, separable, metric) spaces. Whenever we define measures and random variables, all the spaces are equipped with their respective Borel sigma algebras. For a Polish space $\mathcal{W}$, and for sequences $\pi_n \in \mathcal{P}(\mathcal{W})$, $\pi_n \to \pi$ denotes weak convergence of probability measures:

$$\int f d\pi_n \to \int f d\pi, \quad \forall f \in C_b(\mathcal{W})$$

where $C_b(\mathcal{W})$ denotes the space of bounded continuous functions from $\mathcal{W}$ to $\mathbb{R}$. We can metrize this convergence using (for example) the Prokhorov metric on $\mathcal{P}(\mathcal{W})$ defined by:

$$d_{\mathcal{P}}(P,Q) = \inf_{\pi \in \Pi(P,Q)} \inf \left\{ \varepsilon > 0 : \pi[(x,y) : d_{\mathcal{W}(x,y)} \geq \varepsilon] \leq \varepsilon \right\},$$

where $\Pi(P,Q)$ denotes the set of joint probability measures on $\mathcal{V}$ with marginals P and Q.

There are two useful equivalent characterizations of ES. The first one is based on Rockafellar and Uryasev (2000) (see also Föllmer and Schied (2016) lemma 4.46):

$$ES_{\alpha,\pi}(L) = \min_{\beta \in \mathbb{R}} (\beta + (1-\alpha)^{-1} E_\pi[(L-\beta)_+]).$$

The second one is the so-called dual representation of ES as a coherent risk measure:

$$ES_{\alpha,\pi}(L) = \max_{\Theta \in G_\alpha(\pi)} E_\Theta[L],$$

where $G_\alpha(\pi)$ is the set of all probability measures Θ absolutely continuous with respect to π with density satisfying $\frac{d\Theta}{d\pi} \leq (1-\alpha)^{-1}$, π - almost surely:

$$G_\alpha(\pi) := \left\{ \Theta \in \mathcal{P}(\mathcal{W}) \mid \Theta \ll \pi, \frac{d\Theta}{d\pi} \leq (1-\alpha)^{-1} \right\},$$

with the inequality holding π-almost surely.

Furthermore, it is known (see Föllmer and Schied (2016)) that the above maximum is attained by the probability measure $\Theta_0 \in G_\alpha(\pi)$ with density:

$$\frac{d\Theta_0}{d\pi} = \frac{1}{1-\alpha}(1_{\{L>q\}} + \kappa 1_{\{L=1\}}),$$

where q is an α-quantile of L, and where κ is defined as:

$$\kappa = \begin{cases} 0, & \text{if } \pi(L=q)=0 \\ \frac{(1-\alpha)-\pi(L>q)}{\pi(L=q)}, & \text{otherwise.} \end{cases}$$

Ghossoub, Hall, and Saunders (2020) introduces the Maximum Expected Shortfall (MES) based on the above framework and by interpreting $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ as the distributions of risk factors whose marginal distributions are known, but whose joint distribution is unknown. Formally, the Maximum Expected Shortfall (MES) consistent with the prescribed marginals is:

$$\begin{aligned} \text{MES}_\alpha(L) &:= \sup_{\pi \in \Pi(\mu,\nu)} \text{ES}_{\alpha,\pi}(L) \\ &= \sup_{\pi \in \Pi(\mu,\nu)} \min_{\beta \in \mathbb{R}} (\beta + (1-\alpha)^{-1} \mathbb{E}_\pi[(L-\beta)_+]) \\ &= \sup_{\pi \in \Pi(\mu,\nu), \Theta \in G_\alpha(\pi)} \mathbb{E}_\Theta[L]. \end{aligned}$$

Föllmer and Schied (2016) proposition 4.19 states that, as a supremum of suitably well-behaved risk measures, MES is itself a coherent risk measure on the set of bounded functions on $\mathcal{W}$.

In the notation above, we have expressed the dependence of the MES on marginals μ and ν. In order to emphasize this dependence, we denote the MES as follows:

$$\begin{aligned} V_\alpha(\mu,\nu) &:= \sup_{\pi \in \Pi(\mu,\nu)} \left\{ \min_{\beta \in \mathbb{R}} \left\{ \beta + (1-\alpha)^{-1} \mathbb{E}_\pi[(L-\beta)_+] \right\} \right\} \\ &= \sup_{(\pi,\Theta) \in F_\alpha(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} L(x,y)\, d\Theta(x,y). \end{aligned} \tag{1.2.1}$$

where the correspondence $F_\alpha : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \Rightarrow \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \times \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is defined as

$$F_\alpha(\mu,\nu) := \left\{ (\pi,\Theta) \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y}) \times \mathcal{M}_+(\mathcal{X}\times\mathcal{Y}) \,|\, \pi \in \Pi(\mu,\nu), \Theta \ll \pi, \frac{d\Theta}{d\pi} \leq (1-\alpha)^{-1} \right\}, \tag{1.2.2}$$

with the inequality holding π-almost surely.

In practice, we often employ a probability measure that represents our belief of the past experience. We call this the reference probability in our setting. As Glasserman and Yang (2018) presents, we may use the relative entropy (also called Kullback–Leibler divergence) to measure the divergence between the prior probability measure and the reference measure. By adding the penalty to the original

problem (1.2.1), we formulate the penalized version of the primal problem, with a (given) penalty parameter c_0 and a (given) reference probability measure π_{ref}:

$$\begin{aligned} V_{\alpha,c_0}(\mu,\nu) &:= \sup_{(\pi,\Theta)\in F_\alpha(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi \qquad (1.2.3)\\ &= \sup_{(\pi,\Theta)\in F_\alpha(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) \frac{d\pi}{d\pi_{\text{ref}}}\, d\pi_{\text{ref}}\\ &= \sup_{(\pi,\Theta)\in F_\alpha(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \Gamma\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi_{\text{ref}} \end{aligned}$$

and where $\Gamma : \mathbb{R}^+ \to \mathbb{R}$ is defined as

$$\Gamma(r) := \begin{cases} r\log r & \text{if } r > 0\\ 0 & \text{if } r = 0. \end{cases}$$

1.3 Research Contributions

The work by Ghossoub, Hall, and Saunders (2020) considers the problem of bounding the ES of a loss $L(X, Y)$ given that the marginal distributions of the factors X and Y are known, and generalizes the setting by relaxing the assumptions on the loss functions L (e.g. relaxing some assumptions on the shape of L which is done in McNeil, Frey, and Embrechts (2015) and Rüschendorf (2013)) as well as relaxing the spaces from being $\mathbb{R}$-valued to complete separable metric spaces where the risk factors lie.

The resulting optimization problem has the same constraint set as the optimal transport problem, but its objective function is more general. Ghossoub, Hall, and Saunders (2020) presents an analogue of the Kantorovich duality, the continuity of the optimal value and optimizers with respect to the perturbation of the marginal distributions, and the limiting behavior of the optimal value with marginals being simulated from finite sample spaces.

In this thesis, we extend the work by Ghossoub, Hall, and Saunders (2020) to the entropic regularization of optimal transport problem, and contribute to the following aspects:

- We present a proof of an analogue of the Kantorovich duality by utilizing existing results on the Kantorovich duality for optimal transport.
- We prove a convergence of the optimal values and a subsequence of optimal solutions for the primal problem with respect to the Wasserstein convergence of the marginal distributions, the reference measure (assumed to be the product measure $\mu \otimes \nu$), and the penalty parameter.
- We modify an efficient numerical algorithm in solving the entropy regularized optimal transport problem, i.e. Sinkhorn's algorithm, to solve the problem of interest in our case: use the minimax theorem to enable an iterative calling of the Sinkhorn's algorithm and the one dimensional problem solver

The remainder of this thesis is structured as follows: in Chapter 2 we derive the dual problem and prove Kantorovich duality; in Chapter 3 we study the qualitative stability of the primal problem; in Chapter 4 we focus on the numerical simulation of the primal problem; and we conclude the thesis in Chapter 5. In Appendix A we provide some background materials in convex analysis as well as some relevant and important theorems and definitions.

Chapter 2

Kantorovich Duality

In this chapter, we discuss the Kantorovich duality (including both weak and strong duality) between the primal problem (1.2.3) and the dual problem (2.1.4).

Note that for the standard primal problem in many literature:

$$\max_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c \, d\pi,$$

its associated dual problem reads:

$$\begin{aligned} \min_{\varphi,\psi} \quad & \int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{Y}} \psi \, d\nu \\ \text{s.t.} \quad & \varphi(x) + \psi(y) \geq c(x,y), \quad \text{for almost all } (x,y) \in (\mathcal{X},\mathcal{Y}) \end{aligned}$$

where the infimum is over measurable and integrable functions φ and ψ.

The interpretation of the dual problem, in the worker and firm example, is as follows: the primal problem calculates the total utility generated from an optimal matching between workers and firms; the dual problem essentially breaks down the total utility at the individual level. Here $\varphi(x)$ represents the payoff that worker x can generate, similarly $\psi(y)$ represents the payoff that firm y can generate, hence the dual problem represents the sum of each one integrated against its distribution.

The weak duality essentially says that the value of the primal problem is less than or equal to that of the dual problem, which is reasonable since the total utility returned by any pairing cannot exceed the sum of the workers' utility (the maximum utility the workers can possibly achieve) and firms' utility (the maximum utility the firms can possibly achieve). The strong duality says that the worker and firm system reach an equilibrium if there is no gap between the optimal values of the two problems.

2.1 Dual Problem

In this section, we first employ the standard dual derivation method as in the nonlinear programming in order to motivate the form of the dual problem. Then we write down the actual dual problem that we consider in the thesis, for which the proofs on Kantorovich duality are presented in later chapters.

We start from the penalized primal problem (1.2.3):

$$\begin{aligned}\sup_{(\pi,\Theta)\in F_\alpha(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\mathrm{ref}}}\right)\frac{d\pi}{d\pi_{\mathrm{ref}}}\, d\pi_{\mathrm{ref}} \\ = \sup_{(\pi,\Theta)\in F_\alpha(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \Gamma\left(\frac{d\pi}{d\pi_{\mathrm{ref}}}\right) d\pi_{\mathrm{ref}} \quad (2.1.1)\end{aligned}$$

Introduce the Lagrange multipliers $\varphi \in L^1(\mu)$, $\psi \in L^1(d\nu)$, $\beta \in \mathbb{R}$, $\rho \in L^1(d\pi)$ and $\rho \geq 0$, and we get the Lagrangian

$$\begin{aligned}\mathcal{L}(\pi,\Theta,\varphi,\psi,\beta,\rho) &= \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \Gamma\left(\frac{d\pi}{d\pi_{\mathrm{ref}}}\right) d\pi_{\mathrm{ref}} \\ &\quad + \left(\int_{\mathcal{X}} \varphi\, d\mu - \int_{\mathcal{X}\times\mathcal{Y}} \varphi\, d\pi\right) + \left(\int_{\mathcal{Y}} \psi\, d\nu - \int_{\mathcal{X}\times\mathcal{Y}} \psi\, d\pi\right) \\ &\quad + \int_{\mathcal{X}\times\mathcal{Y}} \rho\left((1-\alpha)^{-1} d\pi - d\Theta\right) + \beta\left(1 - \int_{\mathcal{X}\times\mathcal{Y}} d\Theta\right) \\ &= \int_{\mathcal{X}} \varphi\, d\mu + \int_{\mathcal{Y}} \psi\, d\nu + \beta + \int_{\mathcal{X}\times\mathcal{Y}} (L - \rho - \beta)\, d\Theta \\ &\quad + \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \left(c_0((1-\alpha)^{-1}\rho - \varphi - \psi)\frac{d\pi}{d\pi_{\mathrm{ref}}} - \Gamma\left(\frac{d\pi}{d\pi_{\mathrm{ref}}}\right)\right) d\pi_{\mathrm{ref}}\end{aligned}$$

The dual Lagrange function is then given by

$$\begin{aligned}\mathcal{L}_{\mathrm{dual}}(\varphi,\psi,\beta,\rho) &= \sup_{(\pi,\Theta)\in F_\alpha(\mu,\nu)} \mathcal{L}(\pi,\Theta,\varphi,\psi,\beta,\rho) \\ &= \int_{\mathcal{X}} \varphi\, d\mu + \int_{\mathcal{Y}} \psi\, d\nu + \beta + \sup_{\Theta\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \left\{\int_{\mathcal{X}\times\mathcal{Y}} (L - \rho - \beta)\, d\Theta\right\} \\ &\quad + \sup_{\pi\in\Pi(\mu,\nu)} \left\{\frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \left(c_0((1-\alpha)^{-1}\rho - \varphi - \psi)\frac{d\pi}{d\pi_{\mathrm{ref}}} - \Gamma\left(\frac{d\pi}{d\pi_{\mathrm{ref}}}\right)\right) d\pi_{\mathrm{ref}}\right\} \quad (2.1.2)\end{aligned}$$

Define

$$\mathfrak{C} := \left\{ g \in L^1(d\pi_{\text{ref}}) \mid g = \frac{d\pi}{d\pi_{\text{ref}}} \text{ for some } \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \right\}$$

Then if we restrict the set for taking supremum in (2.1.2), we have by Bhattacharya and Dykstra (1995) and Bhattacharya (2006):

$$\begin{aligned} &\sup_{g \in \mathfrak{C}} \left\{ \frac{1}{c_0} \int_{\mathcal{X} \times \mathcal{Y}} \left(c_0((1-\alpha)^{-1}\rho - \varphi - \psi)g - \Gamma(g) \right) d\pi_{\text{ref}} \right\} \\ &= \log \int_{\mathcal{X} \times \mathcal{Y}} e^{c_0\left((1-\alpha)^{-1}\rho - \varphi - \psi\right)} d\pi_{\text{ref}} \end{aligned} \tag{2.1.3}$$

which relates to the Donsker-Varadhan variational formula as shown in Lemma 2.3.1.

Here we make a reasonable conjecture for the dual problem defined as follows:

$$\begin{aligned} D_{\alpha,c_0}(\mu,\nu) &:= \inf_{(\varphi,\psi,\beta) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \times \mathbb{R}} \mathcal{L}_{\text{dual}}(\varphi,\psi,\beta,\rho) \\ &= \inf_{(\varphi,\psi,\beta) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \times \mathbb{R}} \left\{ \int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{Y}} \psi \, d\nu + \beta \right. \\ &\quad \left. + \frac{1}{c_0} \log \int_{\mathcal{X} \times \mathcal{Y}} e^{c_0\left((1-\alpha)^{-1}(L-\beta)_+ - \varphi - \psi\right)} d\pi_{\text{ref}} \right\} \end{aligned} \tag{2.1.4}$$

Note that here we restrict the feasible set from L^1 integrable functions to those of the continuous and bounded functions. Such change will not affect the theorems and proofs in later chapters, as shown in Proposition 5.

2.2 Weak Duality

In this section, we prove an analogue of the Kantorovich duality for the primal problem (1.2.3) and dual problem (2.1.4), i.e., we show that $D_{\alpha,c_0}(\mu,\nu) \leq V_{\alpha,c_0}(\mu,\nu)$. Throughout this section, $\alpha \in (0,1)$ is fixed. We begin with the attainment of the supremum of the entropic regularized primal problem.

Proposition 1 (Existence of Primal Optimizer)**.** *Let L be bounded above and upper-semicontinuous. Then the supremum in (1.2.3) is attained. That is, there exists a pair $(\pi^*,\Theta^*) \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \times \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ such that $(\pi^*,\Theta^*) \in F_\alpha(\mu,\nu)$ and $\int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta^* - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi^*}{d\pi_{\text{ref}}}\right) d\pi^* = V_{\alpha,c_0}(\mu,\nu)$.*

Proof. First note that $F_\alpha(\mu,\nu)$ is nonempty (e.g., $\pi = \Theta = \mu\otimes\nu$). Next, since L is bounded from above, the first integral term is finite; and the second term is bounded below by 0:

$$\int \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi = -\int \log\left(\frac{d\pi_{\text{ref}}}{d\pi}\right) d\pi \geq -\log\int \frac{d\pi_{\text{ref}}}{d\pi} d\pi = 0, \text{ by Jensen's Inequality}$$

we have that the supremum is finite.

Let $\{(\pi_n,\Theta_n)\}_n$ be a sequence in $F_\alpha(\mu,\nu)$ with

$$\int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta_n - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi_n}{d\pi_{\text{ref}}}\right) d\pi_n \to V_{\alpha,c_0}(\mu,\nu).$$

First note that the set $\Pi(\mu,\nu)$ is tight on $\mathcal{X}\times\mathcal{Y}$: given $\mathcal{X}$ and $\mathcal{Y}$ being Polish spaces, since $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, by Ulam's Lemma (see appendix A.2.1), both μ and ν are tight. Then for any $\epsilon > 0$, there exists compact sets $K_\epsilon \subset \mathcal{X}$ and $L_\epsilon \subset \mathcal{Y}$ such that

$$\mu(K_\epsilon^c \leq \epsilon/2), \quad \nu(L_\epsilon^c \leq \epsilon/2).$$

then for any $\pi \in \Pi(\mu,\nu)$,

$$\pi((K_\epsilon \times L_\epsilon)^c) \leq \mu(K_\epsilon^c) + \nu(L_\epsilon^c) \leq \epsilon.$$

By Prokhorov's Theorem (see appendix theorem A.3.1), the set $\{\pi_n\}$ is sequentially compact: there exists a subsequence $\{\pi_{n_k}\}$ of $\{\pi_n\} \subset \Pi(\mu,\nu)$ and $\pi^* \in \Pi(\mu,\nu)$ such that $\pi_{n_k} \to \pi^*$. Then for each $\varepsilon > 0$ there exists a compact K_ε such that $\pi_{n_k}(K_\varepsilon) \leq (1-\alpha)\varepsilon$ for all $k \in \mathbb{N}$.

Let $\Theta_{n_k} \in G_\alpha(\pi_{n_k})$, i.e. $\Theta_{n_k} \ll \pi_{n_k}$ and $\frac{d\Theta_{n_k}}{d\pi_{n_k}} \leq (1-\alpha)^{-1}$. Then for all $k \in \mathbb{N}$:

$$\Theta_{n_k}(K_\varepsilon) = \int_{K_\varepsilon} \frac{d\Theta_{n_k}}{d\pi_{n_k}}\, d\pi_{n_k} \leq \varepsilon$$

Therefore, the set $\{\Theta_{n_k}\}$ is tight. Again by Prokhorov's Theorem, the set $\{\Theta_{n_k}\}$ has a convergent subsequence $\{\Theta_{n_{k_j}}\}$ tending to some Θ^*.

Hence we obtain a (further sub)sequence $\{(\pi_m, \Theta_m)\} \in F_\alpha(\mu, \nu)$ with $\pi_m \in \Pi(\mu, \nu)$, $(\pi_m, \Theta_m) \to (\pi^*, \Theta^*)$. Lemma 6 from Ghossoub, Hall, and Saunders (2020) yields that $\Theta^* \ll \pi^*$ and $\frac{d\Theta^*}{d\pi^*} \leq (1-\alpha)^{-1}$.

Since L is upper-semicontinuous, and $\pi_m \to \pi^*$, by Villani (2008) Lemma 4.3,

$$\limsup_{m\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta_m \leq \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta^*.$$

Since the relative entropy function is lower-semicontinuous (Dupuis and Ellis (1997) Lemma 1.4.3),

$$\lim_{m\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi_m}{d\pi_{\text{ref}}}\right) d\pi_m \geq \liminf_{m\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi_m}{d\pi_{\text{ref}}}\right) d\pi_m \geq \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi^*}{d\pi_{\text{ref}}}\right) d\pi^*.$$

Hence combing the above two inequalities,

$$\begin{aligned} V_{\alpha,c_0}(\mu,\nu) &= \lim_m \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta_m - \frac{1}{c_0} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi_m}{d\pi_{\text{ref}}}\right) d\pi_m \\ &\leq \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta^* - \frac{1}{c_0} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi^*}{d\pi_{\text{ref}}}\right) d\pi^*. \end{aligned}$$

Hence (π^*, Θ^*) is an optimal solution in (1.2.3), as desired.

□

Proposition 2 (Weak Duality)**.** *Let L be bounded above and upper-semicontinuous. Let (π, Θ) be feasible for the problem* (1.2.3) *and (φ, ψ, β) be feasible for the dual problem* (2.1.4)*. Then:*

$$\int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\frac{d\pi}{d\pi_{\text{ref}}}\, d\pi \le \int_{\mathcal{X}} \varphi\, d\mu + \int_{\mathcal{Y}} \psi\, d\nu + \beta + \frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}} e^{c_0\left((1-\alpha)^{-1}(L-\beta)_+ - \varphi - \psi\right)}\, d\pi_{\text{ref}} \tag{2.2.1}$$

Proof. Since (π, Θ) is feasible, we have $\frac{d\pi}{d\pi_{\text{ref}}} \ge 0$ well-defined. Following our discussion on (2.1.3), and we have

$$\begin{aligned}
&\frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}} e^{c_0\left((1-\alpha)^{-1}(L-\beta)_+ - \varphi - \psi\right)}\, d\pi_{\text{ref}} \\
&\ge \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \left(c_0((1-\alpha)^{-1}(L-\beta)_+ - \varphi - \psi)\frac{d\pi}{d\pi_{\text{ref}}} - \frac{d\pi}{d\pi_{\text{ref}}}\log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right)\right) d\pi_{\text{ref}} \\
&= \int_{\mathcal{X}\times\mathcal{Y}} ((1-\alpha)^{-1}(L-\beta)_+ - \varphi - \psi)\, d\pi - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi \\
&= \int_{\mathcal{X}\times\mathcal{Y}} (1-\alpha)^{-1}(L-\beta)_+\, d\pi - \int_{\mathcal{X}} \varphi\, d\mu - \int_{\mathcal{Y}} \psi\, d\nu - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi
\end{aligned}$$

Then for any $\beta \in \mathbb{R}$,

$$\begin{aligned}
\beta + \frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}} e^{c_0\left((1-\alpha)^{-1}(L-\beta)_+ - \varphi - \psi\right)}\, d\pi_{\text{ref}} \ge \beta &+ \int_{\mathcal{X}\times\mathcal{Y}} (1-\alpha)^{-1}(L-\beta)_+\, d\pi \\
&- \int_{\mathcal{X}} \varphi\, d\mu - \int_{\mathcal{Y}} \psi\, d\nu - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi
\end{aligned}$$

which implies

$$\begin{aligned}
&\int_{\mathcal{X}} \varphi\, d\mu + \int_{\mathcal{Y}} \psi\, d\nu + \beta + \frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}} e^{c_0\left((1-\alpha)^{-1}(L-\beta)_+ - \varphi - \psi\right)}\, d\pi_{\text{ref}} \\
&\ge \beta + \int_{\mathcal{X}\times\mathcal{Y}} (1-\alpha)^{-1}(L-\beta)_+\, d\pi - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi \\
&\ge \min_{\beta\in\mathbb{R}} \left\{\beta + \int_{\mathcal{X}\times\mathcal{Y}} (1-\alpha)^{-1}(L-\beta)_+\, d\pi\right\} - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi
\end{aligned}$$

as needed. □

From the above proposition, we obtain $V_{\alpha,c_0}(\mu,\nu) \le D_{\alpha,c_0}(\mu,\nu)$ (feasibility of the dual problem is obvious by taking $\varphi = \psi = 0$, and $\beta = 0$).

2.3 Strong Duality when $\mathcal{X}$ and $\mathcal{Y}$ are Compact

In this section, we aim at proving the strong duality $D_{\alpha,c_0}(\mu,\nu) = V_{\alpha,c_0}(\mu,\nu)$ when both $\mathcal{X}$ and $\mathcal{Y}$ are compact.

We first note the following lemma.

Lemma 2.3.1 (Donsker-Varadhan variational formula)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. The relative entropy $H(\cdot \mid \cdot)$ has the following properties.*

1. *Denote by $C_b(\mathcal{X})$ the space of continuous bounded functions mapping $\mathcal{X}$ to $\mathbb{R}$. Then for each γ and θ in $\mathcal{P}(\mathcal{X})$,*

$$\int_{\mathcal{X}} \log\left(\frac{d\gamma}{d\theta}\right) d\gamma =: H(\gamma|\theta) = \sup_{g\in C_b(\mathcal{X})} \left\{ \int_{\mathcal{X}} g\, d\gamma - \log \int_{\mathcal{X}} e^g\, d\theta \right\}$$

2. *$H(\gamma|\theta)$ is a convex, lower-semicontinuous function of $(\gamma,\theta) \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \times \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, and of each variables γ or θ separately.*

Proof. See Dupuis and Ellis (1997) page 29-32. □

Hence by Lemma 2.3.1, we can rewrite the relative entropy in the form of the variational formula:

$$H(\pi|\pi_{\text{ref}}) = \sup_{g\in C_b(\mathcal{X}\times\mathcal{Y})} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} g\, d\pi - \log \int_{\mathcal{X}\times\mathcal{Y}} e^g\, d\pi_{\text{ref}} \right\} \tag{2.3.1}$$

which is convex and lower-semicontinuous of $(\pi,\pi_{\text{ref}}) \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \times \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, and of each variables π or π_{ref} separately. Then $-H(\pi|\pi_{\text{ref}})$ is concave and upper-semicontinuous.

Note that the supremum is obtained iff g is chosen to be

$$g := \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right)$$

which makes $\int e^g\, d\pi_{\text{ref}} = 1$. Given that $\log(t) \le t-1$ and the equality holds when $t=1$, we have another representation of the KL-divergence:

$$H(\pi|\pi_{\text{ref}}) = \sup_{g\in C_b(\mathcal{X}\times\mathcal{Y})} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} g\, d\pi + 1 - \int_{\mathcal{X}\times\mathcal{Y}} e^g\, d\pi_{\text{ref}} \right\}. \tag{2.3.2}$$

Lemma 2.3.2. *Suppose that $\mathcal{X}, \mathcal{Y}$ are compact complete separable metric spaces, $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, and $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Then*

$$V_{\alpha,c_0}(\mu,\nu)$$
$$= \sup_{\pi\in\Pi(\mu,\nu)} \left\{ \inf_{\beta\in\mathbb{R}, g\in C_b(\mathcal{X}\times\mathcal{Y})} \left\{ \beta + \int_{\mathcal{X}\times\mathcal{Y}} \left((1-\alpha)^{-1}(L-\beta)_+ - \frac{1}{c_0} g \right) d\pi + \frac{1}{c_0} \log \int_{\mathcal{X}\times\mathcal{Y}} e^g \, d\pi_{\text{ref}} \right\} \right\}$$

Proof.

$$V_{\alpha,c_0}(\mu,\nu)$$
$$= \sup_{(\pi,\Theta)\in F_\alpha(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} L \, d\Theta - \frac{1}{c_0} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) \frac{d\pi}{d\pi_{\text{ref}}} \, d\pi_{\text{ref}} \quad \text{by (1.2.3)}$$
$$= \sup_{\pi\in\Pi(\mu,\nu)} \left\{ \min_{\beta\in\mathbb{R}} \left\{ \beta + (1-\alpha)^{-1} \int_{\mathcal{X}\times\mathcal{Y}} (L-\beta)_+ \, d\pi \right\} \right.$$
$$\left. - \frac{1}{c_0} \sup_{g\in C_b(\mathcal{X}\times\mathcal{Y})} \left\{ \int g \, d\pi - \log \int_{\mathcal{X}\times\mathcal{Y}} e^g \, d\pi_{\text{ref}} \right\} \right\}$$

by Lemma 2.3.1

$$= \sup_{\pi\in\Pi(\mu,\nu)} \left\{ \min_{\beta\in\mathbb{R}} \left\{ \beta + (1-\alpha)^{-1} \int_{\mathcal{X}\times\mathcal{Y}} (L-\beta)_+ \, d\pi \right\} \right.$$
$$\left. + \frac{1}{c_0} \inf_{g\in C_b(\mathcal{X}\times\mathcal{Y})} \left\{ - \int g \, d\pi + \log \int_{\mathcal{X}\times\mathcal{Y}} e^g \, d\pi_{\text{ref}} \right\} \right\}$$
$$= \sup_{\pi\in\Pi(\mu,\nu)} \left\{ \inf_{\beta\in\mathbb{R}, g\in C_b(\mathcal{X}\times\mathcal{Y})} \left\{ \beta + \int_{\mathcal{X}\times\mathcal{Y}} \left((1-\alpha)^{-1}(L-\beta)_+ - \frac{1}{c_0} g \right) d\pi + \frac{1}{c_0} \log \int_{\mathcal{X}\times\mathcal{Y}} e^g \, d\pi_{\text{ref}} \right\} \right\}$$

□

We need the following important theorem in order to continue the argument.

Lemma 2.3.3 (Minimax Theorem)**.** *Let X be a locally convex space, Y be a linear space (vector space), $A \subset X$ be a nonempty convex compact set and $B \subset Y$ be a nonempty convex set. Let also $f : A \times B \to \mathbb{R}$ be a function with the property that $f(\cdot, y)$ is concave and upper semicontinuous for every $y \in B$, and $f(x, \cdot)$ is convex for every $x \in A$. Then*

$$\max_{x\in A} \min_{y\in B} f(x,y) = \inf_{y\in B} \max_{x\in A} f(x,y)$$

Proof. See Zălinescu (2002) page 144-146. □

Definition 2.3.1 (Locally Convex Space)**.** A **topological vector space** over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$ is a vector space with a Hausdorff topology such that the product and sum maps $p : \mathbb{F} \times U \to U$ and $s : U \times U \to U$, given by $p(t, x) = tx$ and $s(x, y) = x + y$, are both continuous (where $F \times U$ and $U \times U$ use the product topology). When U is a topological vector space, the continuous dual of U are the spaces

$$U^* = \{f : U \to \mathbb{F} \mid f \text{ is linear and continuous}\}.$$

A topological vector space is said to be **locally convex** when its topology has a basis which consists of convex sets.

Theorem 2.3.1 (Strong Duality under Compactness)**.** *Suppose that $\mathcal{X}, \mathcal{Y}$ are compact complete separable metric spaces, $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, and $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous and bounded. Then $D_{\alpha,c_0}(\mu, \nu) = V_{\alpha,c_0}(\mu, \nu)$.*

Proof. In order to apply the Lemma 2.3.3 in our case, we need to carefully examine the following assumptions:

- X (which is $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$) is locally convex
- Y (which is $\mathbb{R} \times C_b(\mathcal{X} \times \mathcal{Y})$) is a linear space
- A (which is $\Pi(\mu, \nu)$) is nonempty, convex and compact
- B (which is $\mathbb{R} \times C_b(\mathcal{X} \times \mathcal{Y})$) is nonempty and convex
- $f(\cdot, y)$ is concave and upper semicontinuous for every $y \in B$, and $f(x, \cdot)$ is convex for every $x \in A$

where

$$f(\pi, (\beta, g)) := \beta + \int_{\mathcal{X}\times\mathcal{Y}} \left((1-\alpha)^{-1}(L - \beta)_+ - \frac{1}{c_0} g \right) d\pi + \frac{1}{c_0} \log \int_{\mathcal{X}\times\mathcal{Y}} e^g \, d\pi_{\text{ref}}.$$

In the statement, we assume both $\mathcal{X}$ and $\mathcal{Y}$ are compact, following our notes in section 1.2, as in Theorem A.4.1 the space of continuous and bounded functions on a locally compact Hausdorff space, $C_b(\mathcal{X} \times \mathcal{Y})$, has a dual space $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$, the set of Radon measures on $\mathcal{X} \times \mathcal{Y}$ with bounded variation (the norm being total variation of the measure). $C_b(\mathcal{X} \times \mathcal{Y})$ induces a weak topology on $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$, which defines the weak convergence of measures:

$$\mu_n \to \mu \quad \text{iff} \quad \int f \, d\mu_n \to \int f \, d\mu, \quad \forall f \in C_b(\mathcal{X} \times \mathcal{Y})$$

Then the weak* topology $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ locally convex (see Brezis (2010) Proposition 3.12 that gives a basis consisting of convex sets).

The space $\mathbb{R} \times C_b(\mathcal{X} \times \mathcal{Y})$ is obviously convex, and linear (i.e. a vector space).

The set $\Pi(\mu, \nu)$ is nonempty (the product measure $\mu \otimes \nu$ is contained in the set) and convex: let $\pi_1 \in \Pi(\mu, \nu)$ and $\pi_2 \in \Pi(\mu, \nu)$, let $\alpha \in (0, 1)$, then $\alpha\pi_1 + (1-\alpha)\pi_2 \in \Pi(\mu, \nu)$. Indeed,

$$\begin{aligned}
\alpha\pi_1 + (1-\alpha)\pi_2 &\geq 0, \\
\alpha\pi_1(\mathcal{X} \times \mathcal{Y}) + (1-\alpha)\pi_2(\mathcal{X} \times \mathcal{Y}) &= 1, \\
\alpha\pi_1(A \times \mathcal{Y}) + (1-\alpha)\pi_2(A \times \mathcal{Y}) &= \alpha\mu(A) + (1-\alpha)\mu(A) = \mu(A), \quad \forall A \subset \mathcal{X}, \\
\alpha\pi_1(\mathcal{X} \times B) + (1-\alpha)\pi_2(\mathcal{X} \times B) &= \alpha\nu(B) + (1-\alpha)\nu(B) = \nu(B), \quad \forall B \subset \mathcal{Y}.
\end{aligned}$$

Furthermore, $\Pi(\mu, \nu) \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \subset \mathcal{M}(\mathcal{X} \times \mathcal{Y})$, which is the set of joint probability measures with given marginals μ and ν, is (weakly) compact as discussed in Villani (2003) page 49-50.

Finally, for each $(\beta, g) \in \mathbb{R} \times C_b(\mathcal{X} \times \mathcal{Y})$, $f(\pi, (\beta, g))$ is concave and continuous in $\pi \in \Pi(\mu, \nu)$ by Dupuis and Ellis (1997) pages 29-30. And for each $\pi \in \Pi(\mu, \nu)$, $f(\pi, (\beta, g))$ is convex in (β, g) (see Simon (2011) pages 12-13).

Now we are ready to finalize the proof of theorem 2.3.1.

$$
\begin{aligned}
&V_{\alpha,c_0}(\mu,\nu)\\
&= \sup_{\pi\in\Pi(\mu,\nu)}\left\{\inf_{\beta\in\mathbb{R},g\in C_b(\mathcal{X}\times\mathcal{Y})}\left\{\beta+\int_{\mathcal{X}\times\mathcal{Y}}\left((1-\alpha)^{-1}(L-\beta)_+-\frac{1}{c_0}g\right)\,d\pi+\frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}}e^g\,d\pi_{\text{ref}}\right\}\right\}\\
&\quad\text{by Lemma 2.3.2}\\
&= \inf_{\beta\in\mathbb{R},\,g\in C_b(\mathcal{X}\times\mathcal{Y})}\left\{\sup_{\pi\in\Pi(\mu,\nu)}\left\{\beta+\int_{\mathcal{X}\times\mathcal{Y}}\left((1-\alpha)^{-1}(L-\beta)_+-\frac{1}{c_0}g\right)\,d\pi+\frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}}e^g\,d\pi_{\text{ref}}\right\}\right\} \qquad (2.3.3)\\
&= \inf_{\beta\in\mathbb{R},\,g\in C_b(\mathcal{X}\times\mathcal{Y})}\left\{\beta+\sup_{\pi\in\Pi(\mu,\nu)}\left\{\int_{\mathcal{X}\times\mathcal{Y}}\left((1-\alpha)^{-1}(L-\beta)_+-\frac{1}{c_0}g\right)\,d\pi\right\}+\frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}}e^g\,d\pi_{\text{ref}}\right\}\\
&= \inf_{\beta\in\mathbb{R},\,g\in C_b(\mathcal{X}\times\mathcal{Y})}\left\{\beta+\inf_{\substack{\varphi\in C_b(\mathcal{X}),\psi\in C_b(\mathcal{Y}),\\ \varphi+\psi\geq(1-\alpha)^{-1}(L-\beta)_+-\frac{1}{c_0}g}}\left\{\int_{\mathcal{X}}\varphi\,d\mu+\int_{\mathcal{Y}}\psi\,d\nu\right\}+\frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}}e^g\,d\pi_{\text{ref}}\right\} \qquad (2.3.4)\\
&= \inf_{\substack{\beta\in\mathbb{R},\,g\in C_b(\mathcal{X}\times\mathcal{Y}),\\ \varphi\in C_b(\mathcal{X}),\psi\in C_b(\mathcal{Y}),\\ \varphi+\psi\geq(1-\alpha)^{-1}(L-\beta)_+-\frac{1}{c_0}g}}\left\{\beta+\int_{\mathcal{X}}\varphi\,d\mu+\int_{\mathcal{Y}}\psi\,d\nu+\frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}}e^g\,d\pi_{\text{ref}}\right\}\\
&= \inf_{\varphi\in C_b(\mathcal{X}),\psi\in C_b(\mathcal{Y}),\beta\in\mathbb{R}}\left\{\int_{\mathcal{X}}\varphi\,d\mu+\int_{\mathcal{Y}}\psi\,d\nu+\beta+\frac{1}{c_0}\log\int_{\mathcal{X}\times\mathcal{Y}}e^{c_0\left((1-\alpha)^{-1}(L-\beta)_+-\varphi-\psi\right)}\,d\pi_{\text{ref}}\right\} \qquad (2.3.5)
\end{aligned}
$$

where in (2.3.3) we apply the minimax theorem; in (2.3.4) we apply the Kantorovich Duality; and in (2.3.5) we take $g = c_0\left((1-\alpha)^{-1}(L-\beta)_+-\varphi-\psi\right)$ since the term $\log\int e^g\,d\pi_{\text{ref}}$ is increasing/non-decreasing in g. □

2.4 General Strong Duality

In this section, we aim at proving the strong duality result when $\mathcal{X}$ and $\mathcal{Y}$ are Polish, following Villani (2003) pages 31-32.

Proposition 3. *Suppose that $\mathcal{X}, \mathcal{Y}$ are complete separable metric spaces, $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, and $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is uniformly continuous and bounded. Then $D_{\alpha,c_0}(\mu,\nu) = V_{\alpha,c_0}(\mu,\nu)$.*

Proof. Main idea: In light of the weak duality in Proposition 2, we only need to prove $D_{\alpha,c_0}(\mu,\nu) \leq V_{\alpha,c_0}(\mu,\nu)$. We also know that on compact spaces, by the above Theorem 4.1 we have already obtained the strong duality. Hence we aim at constructing a compact subset in the product space, and relate the inequality between $D_{\alpha,c_0}(\mu,\nu)$, $V_{\alpha,c_0}(\mu,\nu)$ and their value restricted on the compact subset.

Given $\delta_1 > 0$, by tightness, there exist compact $X_1^0 \subset \mathcal{X}$ and $Y_1^0 \subset \mathcal{Y}$ such that $\mu[(X_1^0)^c] \leq \delta_1$, $\nu[(Y_1^0)^c] \leq \delta_1$ and therefore

$$\forall \pi \in \Pi(\mu,\nu), \quad \pi[(X_1^0 \times Y_1^0)^c] \leq 2\delta_1. \tag{2.4.1}$$

Given $\delta_2 > 0$, by tightness again, there exists a compact $K \subset \mathcal{X} \times \mathcal{Y}$ such that

$$\pi_{\text{ref}}[K^c] \leq 2\delta_2 \tag{2.4.2}$$

Now let $X^0 \times Y^0 := K \cup (X_1^0 \times Y_1^0)$.

Also note that since $\log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right)$ is measurable w.r.t. $d\pi$,

$$\forall \epsilon > 0, \quad \exists\, \delta_3 > 0 \quad \text{s.t.} \quad \pi[(X^0 \times Y^0)^c] \leq 2\delta_3 \;\Rightarrow\; \left|\int_{(X^0\times Y^0)^c} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi\right| < \epsilon \tag{2.4.3}$$

Hence we can choose $\delta := \min\{\delta_1, \delta_2, \delta_3\} > 0$, such that all (2.4.1) and (2.4.2) and (2.4.3) hold.

Let (π_*, Θ_*) be optimal for the primal problem on $\mathcal{X} \times \mathcal{Y}$ (existence of an optimal solution is guaranteed by Proposition 1), and let π_*^0 be the normalized restriction of π_* to $X^0 \times Y^0$, i.e.,

$$\pi_*^0 = \frac{1_{X^0\times Y^0}}{\pi_*(X^0 \times Y^0)}\pi_*$$

which is a probability measure on $X^0 \times Y^0$. Let the marginal distributions of π_*^0 be denoted by μ^0 and ν^0.

Let

$$V^0_{\alpha,c_0}(\mu_0,\nu_0) := \sup_{(\pi^0,\Theta^0)\in F^0_\alpha(\mu^0,\nu^0)} \int_{X^0\times Y^0} L\, d\Theta^0 - \frac{1}{c_0}\int_{X^0\times Y^0} \log\left(\frac{d\pi^0}{d\pi_{\text{ref}}}\right) d\pi^0$$

where:

$$F^0_\alpha(\mu^0,\nu^0) := \{(\pi^0,\Theta^0) | \pi^0 \in \Pi^0(\mu^0,\nu^0), \Theta^0 \ll \pi^0, \frac{d\Theta^0}{d\pi^0} \leq (1-\alpha)^{-1}\}$$

and let $(\tilde{\pi}^0, \tilde{\Theta}^0)$ attain this maximum (existence is guaranteed by Proposition 1). Define:

$$\tilde{\pi} := \pi_*(X^0\times Y^0)\tilde{\pi}^0 + 1_{(X^0\times Y^0)^c}\pi_* \quad \text{on } \mathcal{X}\times\mathcal{Y}$$

Lemma 2.4.1. *By our definition,* $\tilde{\pi} \in \Pi(\mu,\nu)$.

Proof. First it is easy to see that $\tilde{\pi} \geq 0$. We also have that

$$\begin{aligned}
\int_{\mathcal{X}\times\mathcal{Y}} d\tilde{\pi} &= \int_{\mathcal{X}\times\mathcal{Y}} d(\pi_*(X^0\times Y^0)\tilde{\pi}^0 + 1_{(X^0\times Y^0)^c}\pi_*) \\
&= \int_{X^0\times Y^0} d(\pi_*(X^0\times Y^0)\tilde{\pi}^0) + \int_{(X^0\times Y^0)^c} d(1_{(X^0\times Y^0)^c}\pi_*) \\
&= \pi_*(X^0\times Y^0)\int_{X^0\times Y^0} d\tilde{\pi}^0 + \int_{(X^0\times Y^0)^c} d\pi_* \\
&= \pi_*(X^0\times Y^0) + \pi_*((X^0\times Y^0)^c) \\
&= 1
\end{aligned}$$

Hence $\tilde{\pi}$ is a probability measure on $\mathcal{X}\times\mathcal{Y}$.

Let A be a measurable subset of $\mathcal{X}$, then

$$\begin{aligned}
\tilde{\pi}[A \times \mathcal{Y}] &= \int_{A\times\mathcal{Y}} d\tilde{\pi} \\
&= \int_{A\cap X^0\times\mathcal{Y}} d\tilde{\pi} + \int_{A\cap (X^0)^c\times\mathcal{Y}} d\tilde{\pi} \\
&= \int_{A\cap X^0\times Y^0} d\tilde{\pi} + \int_{A\cap X^0\times (Y^0)^c} d\tilde{\pi} + \int_{A\cap (X^0)^c\times Y^0} d\tilde{\pi} + \int_{A\cap (X^0)^c\times (Y^0)^c} d\tilde{\pi} \\
&= \int_{A\cap X^0\times Y^0} d(\pi_*(X^0\times Y^0)\tilde{\pi}^0) + \int_{A\cap X^0\times (Y^0)^c} d(\mathtt{1}_{(X^0\times Y^0)^c}\pi_*) \\
&\quad + \int_{A\cap (X^0)^c\times Y^0} d(\mathtt{1}_{(X^0\times Y^0)^c}\pi_*) + \int_{A\cap (X^0)^c\times (Y^0)^c} d(\mathtt{1}_{(X^0\times Y^0)^c}\pi_*) \\
&= \pi_*(X^0\times Y^0)\int_{A\cap X^0\times Y^0} d\tilde{\pi}^0 + \int_{A\cap (X^0)^c\times (Y^0)^c} d\pi_* \\
&= \pi_*(X^0\times Y^0)\tilde{\pi}^0(A\cap X^0\times Y^0) + \pi_*(A\cap (X^0)^c\times (Y^0)^c) \\
&= \pi_*(X^0\times Y^0)\mu^0(A\cap X^0) + \pi_*(A\cap (X^0)^c\times (Y^0)^c) \\
&= \pi_*(A\cap X^0\times Y^0) + \pi_*(A\cap (X^0)^c\times (Y^0)^c) \\
&= \pi_*(A\cap \mathcal{X}\times\mathcal{Y}) \\
&= \mu(A)
\end{aligned}$$

Similarly, we have $\tilde{\pi}[\mathcal{X}\times B] = \nu(B),\ \forall$ measurable $B \subset \mathcal{Y}$. Hence by definition, $\tilde{\pi} \in \Pi(\mu,\nu)$. □

Define $\tilde{\Theta}$ by:

$$\begin{aligned}
\frac{d\tilde{\Theta}}{d\tilde{\pi}} &:= \frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0} + \mathtt{1}_{(X^0\times Y^0)^c} \le (1-\alpha)^{-1} \quad \text{on } \mathcal{X}\times\mathcal{Y} \\
&= \begin{cases} \frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0} & \text{on } X^0\times Y^0 \\ 1 & \text{on } (X^0\times Y^0)^c \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Lemma 2.4.2. *Based on our definition, we have* $\tilde{\Theta} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

Proof.

$$\begin{aligned}
\int_{\mathcal{X}\times\mathcal{Y}} d\tilde{\Theta} &= \int_{\mathcal{X}\times\mathcal{Y}} \frac{d\tilde{\Theta}}{d\tilde{\pi}}\, d\tilde{\pi} \\
&= \int_{\mathcal{X}\times\mathcal{Y}} (\frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0} + \mathtt{1}_{(X^0\times Y^0)^c})\, d(\pi_*(X^0 \times Y^0)\tilde{\pi}^0 + \mathtt{1}_{(X^0\times Y^0)^c}\pi_*) \\
&= \int_{X^0\times Y^0} \frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0}\, d(\pi_*(X^0 \times Y^0)\tilde{\pi}^0) + \int_{(X^0\times Y^0)^c} \mathtt{1}_{(X^0\times Y^0)^c}\, d(\mathtt{1}_{(X^0\times Y^0)^c}\pi_*) \\
&= \pi_*(X^0 \times Y^0) \int_{X^0\times Y^0} \frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0}\, d\tilde{\pi}^0 + \int_{(X^0\times Y^0)^c} d\pi_* \\
&= \pi_*(X^0 \times Y^0) \int_{X^0\times Y^0} d\tilde{\Theta}^0 + \pi_*((X^0 \times Y^0)^c) \\
&= \pi_*(X^0 \times Y^0) + \pi_*((X^0 \times Y^0)^c) = 1
\end{aligned}$$

as desired. □

Now that we have constructed compact subsets and the associated probability measures, we first aim at relating the penalized primal problem value $V_{\alpha,c_0}(\mu,\nu)$ with its restriction on the compact subset $V^0_{\alpha,c_0}(\mu,\nu)$.

Note that

$$\begin{aligned}
V_{\alpha,c_0}(\mu,\nu) &= \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta_* - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi_*}{d\pi_{\text{ref}}}\right) d\pi_* \\
&\geq \int_{\mathcal{X}\times\mathcal{Y}} L\, d\tilde{\Theta} - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right) d\tilde{\pi} \\
&= \int_{X^0\times Y^0} L\, d\tilde{\Theta} - \frac{1}{c_0}\int_{X^0\times Y^0} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right) d\tilde{\pi} \\
&\quad + \int_{(X^0\times Y^0)^c} L\, d\tilde{\Theta} - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right) d\tilde{\pi} \\
&= T_1 + T_2
\end{aligned} \tag{2.4.4}$$

where

$$\begin{aligned}
T_1 &:= \int_{X^0\times Y^0} L\, d\tilde{\Theta} - \frac{1}{c_0}\int_{X^0\times Y^0} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right) d\tilde{\pi} \\
T_2 &:= \int_{(X^0\times Y^0)^c} L\, d\tilde{\Theta} - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right) d\tilde{\pi}
\end{aligned}$$

Our goal is to give a lower bound of the above inequality (2.4.4). In order to improve clarity of the proof, we deviate a little bit here to present two lemmas that aim at working on (2.4.4).

Lemma 2.4.3.

$$T_1 = \pi_*(X^0 \times Y^0) V^0_{\alpha,c_0}(\mu_0, \nu_0) - \pi_*(X^0 \times Y^0) \log(\pi_*(X^0 \times Y^0)) \frac{1}{c_0}$$

Proof.

$$\begin{aligned}
T_1 &= \int_{X^0 \times Y^0} L \frac{d\tilde{\Theta}}{d\tilde{\pi}} \, d\tilde{\pi} - \frac{1}{c_0} \int_{X^0 \times Y^0} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right) d\tilde{\pi} \\
&= \int_{X^0 \times Y^0} L \left(\frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0} + \mathtt{1}_{(X^0 \times Y^0)^c}\right) d\tilde{\pi} - \frac{1}{c_0} \int_{X^0 \times Y^0} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right) d\tilde{\pi} \\
&= \int_{X^0 \times Y^0} L \frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0} \, d\tilde{\pi} - \frac{1}{c_0} \int_{X^0 \times Y^0} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right) d\tilde{\pi} \\
&= \int_{X^0 \times Y^0} L \frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0} \, d(\pi_*(X^0 \times Y^0)\tilde{\pi}^0 + \mathtt{1}_{(X^0 \times Y^0)^c} \pi_*) \\
&\quad - \frac{1}{c_0} \int_{X^0 \times Y^0} \log\left(\frac{d(\pi_*(X^0 \times Y^0)\tilde{\pi}^0 + \mathtt{1}_{(X^0 \times Y^0)^c}\pi_*)}{d\pi_{\text{ref}}}\right) d(\pi_*(X^0 \times Y^0)\tilde{\pi}^0 + \mathtt{1}_{(X^0 \times Y^0)^c}\pi_*) \\
&= \int_{X^0 \times Y^0} L \frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0} \, d(\pi_*(X^0 \times Y^0)\tilde{\pi}^0) - \frac{1}{c_0} \int_{X^0 \times Y^0} \log\left(\frac{d(\pi_*(X^0 \times Y^0)\tilde{\pi}^0)}{d\pi_{\text{ref}}}\right) d(\pi_*(X^0 \times Y^0)\tilde{\pi}^0) \\
&= \pi_*(X^0 \times Y^0) \int_{X^0 \times Y^0} L \frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0} \, d\tilde{\pi}^0 - \pi_*(X^0 \times Y^0) \frac{1}{c_0} \int_{X^0 \times Y^0} \log\left(\frac{d\tilde{\pi}^0}{d\pi_{\text{ref}}} \pi_*(X^0 \times Y^0)\right) d\tilde{\pi}^0 \\
&= \pi_*(X^0 \times Y^0) \int_{X^0 \times Y^0} L \, d\tilde{\Theta}^0 \\
&\quad - \pi_*(X^0 \times Y^0) \frac{1}{c_0} \int_{X^0 \times Y^0} \log\left(\frac{d\tilde{\pi}^0}{d\pi_{\text{ref}}}\right) d\tilde{\pi}^0 \\
&\quad - \pi_*(X^0 \times Y^0) \frac{1}{c_0} \int_{X^0 \times Y^0} \log(\pi_*(X^0 \times Y^0)) \, d\tilde{\pi}^0 \\
&= \pi_*(X^0 \times Y^0) V^0_{\alpha,c_0}(\mu_0, \nu_0) - \pi_*(X^0 \times Y^0) \log(\pi_*(X^0 \times Y^0)) \frac{1}{c_0}
\end{aligned}$$

□

Lemma 2.4.4.

$$T_2 = \int_{(X^0\times Y^0)^c} L\, d\pi_* - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\pi_*}{d\pi_{\text{ref}}}\right)\, d\pi_*$$

Proof.

$$\begin{aligned}
T_2 &= \int_{(X^0\times Y^0)^c} L\frac{d\tilde{\Theta}}{d\tilde{\pi}}\, d\tilde{\pi} - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right)\, d\tilde{\pi} \\
&= \int_{(X^0\times Y^0)^c} L(\frac{d\tilde{\Theta}^0}{d\tilde{\pi}^0} + 1_{(X^0\times Y^0)^c})\, d\tilde{\pi} - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right)\, d\tilde{\pi} \\
&= \int_{(X^0\times Y^0)^c} L1_{(X^0\times Y^0)^c}\, d\tilde{\pi} - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right)\, d\tilde{\pi} \\
&= \int_{(X^0\times Y^0)^c} L\, d\tilde{\pi} - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right)\, d\tilde{\pi} \\
&= \int_{(X^0\times Y^0)^c} L\, d(\pi_*(X^0\times Y^0)\tilde{\pi}^0 + 1_{(X^0\times Y^0)^c}\pi_*) \\
&\quad - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d(\pi_*(X^0\times Y^0)\tilde{\pi}^0 + 1_{(X^0\times Y^0)^c}\pi_*)}{d\pi_{\text{ref}}}\right)\, d(\pi_*(X^0\times Y^0)\tilde{\pi}^0 + 1_{(X^0\times Y^0)^c}\pi_*) \\
&= \int_{(X^0\times Y^0)^c} L\, d(1_{(X^0\times Y^0)^c}\pi_*) - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d(1_{(X^0\times Y^0)^c}\pi_*)}{d\pi_{\text{ref}}}\right)\, d(1_{(X^0\times Y^0)^c}\pi_*) \\
&= \int_{(X^0\times Y^0)^c} L\, d\pi_* - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\pi_*}{d\Theta_{ref}}\right)\, d\pi_*
\end{aligned}$$

□

Now we go back to the main proof of Proposition 4. By combining the above two lemmas, we can finally bound $V_{\alpha,c_0}(\mu,\nu)$ in (2.4.4) from below:

$$\begin{aligned}
V_{\alpha,c_0}(\mu,\nu) &\geq \int_{X^0\times Y^0} L\, d\tilde{\Theta} - \frac{1}{c_0}\int_{X^0\times Y^0} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right)\, d\tilde{\pi} \\
&\quad + \int_{(X^0\times Y^0)^c} L\, d\tilde{\Theta} - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\tilde{\pi}}{d\pi_{\text{ref}}}\right)\, d\tilde{\pi} \\
&\geq \pi_*(X^0\times Y^0)V^0_{\alpha,c_0}(\mu_0,\nu_0) - \pi_*(X^0\times Y^0)\log(\pi_*(X^0\times Y^0))\frac{1}{c_0} \\
&\quad + \int_{(X^0\times Y^0)^c} L\, d\pi_* - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\pi_*}{d\pi_{\text{ref}}}\right)\, d\pi_*
\end{aligned}$$

Since $1 \geq \pi_*(X^0 \times Y^0) \geq 1 - 2\delta$, as $\delta \searrow 0$ we have $\pi_*(X^0 \times Y^0) \to 1$.

We still need an additional argument to make sure that the last term in the above equation $\int_{(X^0\times Y^0)^c} L\, d\pi_* - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\pi_*}{d\pi_{\text{ref}}}\right) d\pi_*$ vanishes as $\delta \searrow 0$, as shown in the below lemma.

Lemma 2.4.5. *$\int_{(X^0\times Y^0)^c} L\, d\pi_* - \frac{1}{c_0}\int_{(X^0\times Y^0)^c} \log\left(\frac{d\pi_*}{d\pi_{\text{ref}}}\right) d\pi_*$ is finite as a linear order in δ.*

Proof.

$$\begin{aligned} \left|\int_{(X^0\times Y^0)^c} L\, d\pi_*\right| &\leq \|L\|_\infty\, \pi_*((X^0 \times Y^0)^c) \\ &\leq 2\delta \|L\|_\infty \end{aligned}$$

$$\frac{1}{c_0}\left|\int_{(X^0\times Y^0)^c} \log\left(\frac{d\pi_*}{d\pi_{\text{ref}}}\right) d\pi_*\right| \leq \frac{1}{c_0}\epsilon \quad \text{by our choice of } \delta$$

where $\epsilon \searrow 0$ as $\delta \searrow 0$.

□

Hence

$$\delta \searrow 0, \quad V_{\alpha,c_0}(\mu,\nu) \geq V^0_{\alpha,c_0}(\mu,\nu) \tag{2.4.5}$$

which completes the first part of our final goal.

Next, we aim at relating the penalized dual problem value $D_{\alpha,c_0}(\mu,\nu)$ with its value $D^0_{\alpha,c_0}(\mu,\nu)$ restricted on the compact subset.

Now define, with the obvious notation:

$$J(w) := \int_{\mathcal{X}} \varphi d\mu + \int_{\mathcal{Y}} \psi d\nu + \beta + \frac{1}{c_0}\log \int_{\mathcal{X}\times\mathcal{Y}} e^{c_0\left((1-\alpha)^{-1}(L-\beta)_+ - \varphi - \psi\right)} d\pi_{\text{ref}}$$

$$J^0(w^0) := \int_{X^0} \varphi^0 d\mu^0 + \int_{Y^0} \psi^0 d\nu^0 + \beta^0 + \frac{1}{c_0}\log \int_{X^0\times Y^0} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \varphi^0 - \psi^0\right)} d\pi_{\text{ref}}$$

There exists $w^0 = (\varphi^0, \psi^0, \beta^0) \in C_b(X^0) \times C_b(Y^0) \times \mathbb{R})$ such that

$$J^0(w^0) \leq \inf J^0 + \delta =: D^0_{\alpha,c_0}(\mu^0,\nu^0) + \delta$$

where

$$
\begin{aligned}
D^0_{\alpha,c_0}(\mu^0,\nu^0) &:= \inf_{w^0} J^0(w^0) \\
&= \inf_{(\varphi^0,\psi^0,\beta^0)\in C_b(X^0)\times C_b(Y^0)\times\mathbb{R}} \Big\{ \int_{X^0} \varphi^0 d\mu^0 + \int_{Y^0} \psi^0 d\nu^0 + \beta^0 \\
&\quad + \frac{1}{c_0}\log \int_{X^0\times Y^0} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \varphi^0 - \psi^0\right)}\, d\pi_{\text{ref}} \Big\}
\end{aligned}
$$

Noting that $(0,0,\|L\|_\infty)$ is feasible, we obtain $\inf J^0 \le \|L\|_\infty$, and therefore:

$$
\begin{aligned}
J^0(\nu^0) &= \int_{X^0} \varphi^0 d\mu^0 + \int_{Y^0} \psi^0 d\nu^0 + \beta^0 + \frac{1}{c_0}\log \int_{X^0\times Y^0} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \varphi^0 - \psi^0\right)}\, d\pi_{\text{ref}} \\
&\le \|L\|_\infty + \delta \qquad (2.4.6)
\end{aligned}
$$

Since X^0, Y^0 are compact subsets of Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, by Tietze extension theorem (see appendix A.6.1), we can extend our choice $(\varphi^0,\psi^0,\beta^0) \in C_b(X^0)\times C_b(Y^0)\times\mathbb{R}$ to $(\bar\varphi^0,\bar\psi^0,\beta^0) \in C_b(\mathcal{X})\times C_b(\mathcal{Y})\times\mathbb{R}$ such that $\bar\varphi^0 = \varphi^0$ on X^0, $\bar\psi^0 = \psi^0$ on Y^0.

Proceeding as in Villani (2003) page 30, one can then show that

$$
\begin{aligned}
D_{\alpha,c_0}(\mu,\nu) &\le \int_{\mathcal{X}} \bar\varphi^0\, d\mu + \int_{\mathcal{Y}} \bar\psi^0\, d\nu + \beta^0 + \frac{1}{c_0}\log \int_{\mathcal{X}\times\mathcal{Y}} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar\varphi^0 - \bar\psi^0\right)}\, d\pi_{\text{ref}} \\
&= \int_{X^0} \bar\varphi^0\, d\mu + \int_{Y^0} \bar\psi^0\, d\nu + \beta^0 + \frac{1}{c_0}\log \int_{X^0\times Y^0} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar\varphi^0 - \bar\psi^0\right)}\, d\pi_{\text{ref}} \\
&\quad + \int_{(X^0)^c} \bar\varphi^0\, d\mu + \int_{(Y^0)^c} \bar\psi^0\, d\nu + \frac{1}{c_0}\log \int_{(X^0\times Y^0)^c} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar\varphi^0 - \bar\psi^0\right)}\, d\pi_{\text{ref}} \\
&= \int_{X^0} \varphi^0\, d\mu^0 + \int_{Y^0} \psi^0\, d\nu^0 + \beta^0 + \frac{1}{c_0}\log \int_{X^0\times Y^0} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar\varphi^0 - \bar\psi^0\right)}\, d\pi_{\text{ref}} \\
&\quad + \int_{(X^0)^c} \bar\varphi^0\, d\mu + \int_{(Y^0)^c} \bar\psi^0\, d\nu + \frac{1}{c_0}\log \int_{(X^0\times Y^0)^c} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar\varphi^0 - \bar\psi^0\right)}\, d\pi_{\text{ref}} \\
&\le D^0_{\alpha,c_0}(\mu^0,\nu^0) + \delta \\
&\quad + \int_{(X^0)^c} \bar\varphi^0\, d\mu + \int_{(Y^0)^c} \bar\psi^0\, d\nu + \frac{1}{c_0}\log \int_{(X^0\times Y^0)^c} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar\varphi^0 - \bar\psi^0\right)}\, d\pi_{\text{ref}} \\
&\le D^0_{\alpha,c_0}(\mu^0,\nu^0) \\
&\quad + \delta\cdot\left(\|\bar\varphi^0\|_\infty + \|\bar\psi^0\|_\infty\right) + \frac{1}{c_0}\log \int_{(X^0\times Y^0)^c} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar\varphi^0 - \bar\psi^0\right)}\, d\pi_{\text{ref}}
\end{aligned}
$$

Note that since for all $x > 0$, $\log x < x$, we have that as $\delta \searrow 0$,

$$\begin{aligned}
&\log \int_{(X^0\times Y^0)^c} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar{\varphi}^0 - \bar{\psi}^0\right)}\, d\pi_{\mathrm{ref}} \\
&< \int_{(X^0\times Y^0)^c} e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar{\varphi}^0 - \bar{\psi}^0\right)}\, d\pi_{\mathrm{ref}} \\
&\leq \pi_{\mathrm{ref}}[(X^0 \times Y^0)^c] \left\| e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar{\varphi}^0 - \bar{\psi}^0\right)} \right\|_\infty \\
&\leq 2\delta \cdot \left\| e^{c_0\left((1-\alpha)^{-1}(L-\beta^0)_+ - \bar{\varphi}^0 - \bar{\psi}^0\right)} \right\|_\infty \to 0
\end{aligned}$$

Hence we have

$$\delta \searrow 0 \quad \Rightarrow \quad D_{\alpha,c_0}(\mu,\nu) \leq D^0_{\alpha,c_0}(\mu,\nu) \tag{2.4.7}$$

which completes the last part of our final goal.

Combining (2.4.5), (2.4.7) and the strong duality on compact sets (i.e. Theorem 4.1), we obtain

$$D_{\alpha,c_0}(\mu,\nu) \leq D^0_{\alpha,c_0}(\mu,\nu) = V^0_{\alpha,c_0}(\mu,\nu) \leq V_{\alpha,c_0}(\mu,\nu)$$

which, together with weak duality, yields the intended result. □

Proposition 4. *Suppose that $\mathcal{X}, \mathcal{Y}$ are complete separable metric spcaes, $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, and $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is upper semi-continuous and bounded above. Then $D_{\alpha,c_0}(\mu,\nu) = V_{\alpha,c_0}(\mu,\nu)$.*

Proof. Again, we can follow Villani (2003) pp.31-32.

Since L is upper semi-continuous and bounded above in the Polish space, one can find a sequence of nonincreasing uniformly continuous functions L_n such that $L = \inf L_n$. By replacing L_n by $\sup\{L_n, n\}$, we can assume that each L_n is bounded from above.

Denote by $D^n_{\alpha,c_0}(\mu,\nu) = V^n_{\alpha,c_0}(\mu,\nu)$ the optimal value with loss function L_n. It is immediate that $H_{\alpha,c_0}(L_n) \subset H_{\alpha,c_0}(L)$ and therefore $D^n_{\alpha,c_0}(\mu,\nu) \geq D_{\alpha,c_0}(\mu,\nu)$. Furthermore, $V^n_{\alpha,c_0}(\mu,\nu)$ is decreasing in n and $V^n_{\alpha,c_0}(\mu,\nu) \geq V_{\alpha,c_0}(\mu,\nu)$. Let (π_n, Θ_n) be optimal for $V^n_{\alpha,c_0}(\mu,\nu)$ and, applying Prokhorov's Theorem as above (and passing to a subsequence if necessary), $(\pi_n, \Theta_n) \to (\pi, \Theta)$. When $n \geq m$, $\int L_n d\Theta_n - \frac{1}{c_0}\int \log\left(\frac{d\pi_n}{d\pi_{\text{ref}}}\right) d\pi_n \leq \int L_m d\Theta_n - \frac{1}{c_0}\int \log\left(\frac{d\pi_n}{d\pi_{\text{ref}}}\right) d\pi_n$, and therefore by the Portmanteau Theorem and lower-semicontinuity of the Kullback-Leibler distance:

$$\begin{aligned}
\lim_{n\to\infty} V^n_{\alpha,c_0}(\mu,\nu) &= \lim_{n\to\infty} \int L_n d\Theta_n - \frac{1}{c_0}\int \log\left(\frac{d\pi_n}{d\pi_{\text{ref}}}\right) d\pi_n \\
&\leq \limsup_{n\to\infty} \int L_m \, d\Theta_n - \frac{1}{c_0}\int \log\left(\frac{d\pi_n}{d\pi_{\text{ref}}}\right) d\pi_n \\
&\leq \int L_m d\Theta - \frac{1}{c_0}\int \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi
\end{aligned}$$

Monotone convergence then gives $\lim_{n\to\infty} V^n_{\alpha,c_0}(\mu,\nu) \leq \int L \, d\Theta - \frac{1}{c_0}\int \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi \leq V_{\alpha,c_0}(\mu,\nu)$. Thus $\lim_{n\to\infty} V^n_{\alpha,c_0}(\mu,\nu) = V_{\alpha,c_0}(\mu,\nu)$, and we have $V_{\alpha,c_0}(\mu,\nu) \geq D_{\alpha,c_0}(\mu,\nu)$, and the result follows by weak duality. □

Proposition 5. *It does not change the value of the infimum in* (2.1.4) *if one restricts the dual feasible set to those functions* (φ, ψ, β) *which are bounded and continuous.*

Proof. We wish to emphasize the distinction between these definitions, hence here we informally write $H_{\alpha,c_0}(L) \cap C_b$ for restriction to continuous bounded functions, and $H_{\alpha,c_0}(L)$ for the dual feasible set containing L^1 integrable functions.

Define

$$J(\varphi, \psi, \beta, \rho) = \int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{Y}} \psi \, d\nu + \beta + \frac{1}{c_0} \log \int_{\mathcal{X}\times\mathcal{Y}} e^{c_0\left((1-\alpha)^{-1}(L-\beta)_+ - \varphi - \psi\right)} \, d\pi_{\text{ref}}$$

$$I(\pi, \Theta) = \int_{\mathcal{X}\times\mathcal{Y}} L \, d\Theta - \frac{1}{c_0} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right) d\pi$$

We claim that

$$\inf_{\varphi\in C_b(\mathcal{X}),\, \psi\in C_b(\mathcal{Y}),\, \beta\in\mathbb{R}} J(\varphi, \psi, \beta, \rho) \geq \inf_{\varphi\in L^1(d\mu),\, \psi\in L^1(d\nu),\, \beta\in\mathbb{R}} J(\varphi, \psi, \beta, \rho) \geq \sup_{(\pi,\Theta)\in F_\alpha(\mu,\nu)} I(\pi, \Theta)$$

Indeed, the left inequality is a direct consequence of $C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \times \mathbb{R} \subset L^1(d\mu) \times L^1(d\nu) \times \mathbb{R}$, and the right inequality is shown in Proposition 2.

Hence if we manage to show the strong duality

$$\inf_{\varphi\in C_b(\mathcal{X}),\, \psi\in C_b(\mathcal{Y}),\, \beta\in\mathbb{R}} J(\varphi, \psi, \beta, \rho) = \sup I(\pi, \Theta),$$

it automatically implies

$$\inf_{\varphi\in L^1(d\mu),\, \psi\in L^1(d\nu),\, \beta\in\mathbb{R}} J(\varphi, \psi, \beta, \rho) = \inf_{\varphi\in C_b(\mathcal{X}),\, \psi\in C_b(\mathcal{Y}),\, \beta\in\mathbb{R}} J(\varphi, \psi, \beta, \rho).$$

□

Hence we can work with (2.1.4).

Chapter 3

Stability of the Entropy Penalized Maximum Expected Shortfall

3.1 Introduction

In this chapter, we study the convergence of the objective value, and the set of optimal solutions of the entropy penalized problem (1.2.3) with respect to weak convergence of the marginal distributions μ, ν, and the reference measure π_{ref} that parametrize the problem.

Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be two Polish (i.e., complete, separable, metric) spaces, and $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$ be the product space. When $p \in [1, \infty]$ is given, we will employ the following product metric on $\mathcal{W}$, where $w = (x, y)$, and $w' = (x', y')$

$$d_{\mathcal{W},p}(w, w') = \begin{cases} (d_{\mathcal{X}}(x, x')^p + d_{\mathcal{Y}}(y, y')^p)^{1/p}, & p \in [1, \infty), \\ \max(d_{\mathcal{X}}(x, x'), d_{\mathcal{Y}}(y, y')), & p = \infty. \end{cases} \tag{3.1.1}$$

The spaces are equipped with their respective Borel sigma algebras. For a Polish space $\mathcal{V}$, let $\mathcal{P}(\mathcal{V})$ denote the set of all Borel probability measures. For sequences $\pi_n \in \mathcal{P}(\mathcal{V})$, $\pi_n \to \pi$ denotes weak convergence of probability measures:

$$\int f d\pi_n \to \int f d\pi, \quad \forall f \in C_b(\mathcal{V}).$$

Let $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, and let $\Pi(\mu, \nu)$ be the set of joint probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. Assume $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous and bounded (we will need additional assumptions on L later on).

Suppose that we have sequences $\mu_n \to \mu$ in $\mathcal{P}(\mathcal{X})$, $\nu_n \to \nu$ in $\mathcal{P}(\mathcal{Y})$, and $\pi_{\text{ref},n} \in \Pi(\mu_n, \nu_n) \to \pi_{\text{ref}} \in \Pi(\mu, \nu)$. Let c_n be sequence of strictly positive numbers, with $c_n \to c > 0$.

Define the following functions on $\mathcal{P}(\mathcal{W})$:

$$I_n(\pi_n) := \begin{cases} \mathrm{ES}_{\alpha,\pi}(L) - \frac{1}{c_n} H(\pi_n \,|\, \pi_{\text{ref},n}), & \pi_n \in \Pi(\mu_n, \nu_n) \\ -\infty, & \pi_n \notin \Pi(\mu_n, \nu_n) \end{cases}$$
$$I(\pi) := \begin{cases} \mathrm{ES}_{\alpha,\pi}(L) - \frac{1}{c} H(\pi \,|\, \pi_{\text{ref}}), & \pi \in \Pi(\mu, \nu) \\ -\infty, & \pi \notin \Pi(\mu, \nu) \end{cases}$$

Alternatively, we can write:

$$I_n(\pi_n) := \begin{cases} \min_{\beta \in \mathbb{R}} \left\{ \beta + (1-\alpha)^{-1} \int (L-\beta)_+ \, d\pi_n \right\} - \frac{1}{c_n} H(\pi_n \,|\, \pi_{\text{ref},n}), & \pi_n \in \Pi(\mu_n, \nu_n) \\ -\infty, & \pi_n \notin \Pi(\mu_n, \nu_n) \end{cases}$$
$$I(\pi) := \begin{cases} \min_{\beta \in \mathbb{R}} \left\{ \beta + (1-\alpha)^{-1} \int (L-\beta)_+ \, d\pi \right\} - \frac{1}{c} H(\pi \,|\, \pi_{\text{ref}}), & \pi \in \Pi(\mu, \nu) \\ -\infty, & \pi \notin \Pi(\mu, \nu) \end{cases}$$

Furthermore, let $V_n = \sup I_n$, and $V = \sup I$, and note that by Proposition 1, both suprema are attained.

Then

$$V_n := \sup_{\pi_n \in \Pi(\mu_n, \nu_n)} I_n(\pi_n) = I_n(\pi_n^*).$$

Also define

$$V := \sup_{\pi \in \Pi(\mu, \nu)} I(\pi) = I(\pi^*).$$

We are interested in two fundamental questions:

- Does $V_n \to V$ hold?
- Does a subsequence $\pi_{n_k}^*$ of $\{\pi_n^*\} \subset \Pi(\mu_n, \nu_n)$ exist such that $\pi_{n_k}^*$ converges to $\pi^* \in \Pi(\mu, \nu)$, which solves the limiting optimization problem?

3.2 Epi-Convergence

In order to answer the questions raised above, we employ the notion of Gamma convergence, also called epi-convergence, which we recall here (see, for example Borwein and Zhu (2005)).

Definition 3.2.1 (Epi-convergence)**.** Let $\mathcal{W}$ be a metric space, and let $f_n : \mathcal{W} \to [-\infty, \infty]$ be a sequence of lower semi-continuous functions. We say that f_n epi-converges to the function $f : \mathcal{W} \to \mathbb{R}$ if and only if at each point $w \in \mathcal{W}$ we have:

- $\liminf_{n\to\infty} f_n(w_n) \geq f(w)$ for every sequence $w_n \to w$
- $\limsup_{n\to\infty} f_n(w_n) \leq f(w)$ for some sequence $w_n \to w$

The importance of this notion of convergence is emphasized by the following result (Borwein and Zhu (2005), Theorem 5.1.14). Here, $\mathrm{dom} f := \{v \,|\, f(v) < \infty\}$, and for a sequence of sets A_n in a metric space $\mathcal{V}$, $\limsup A_n = \{v \in \mathcal{V} \,|\, \liminf d(A_n, v) = 0\}$.

Theorem 3.2.1. *Let $\mathcal{V}$ be a metric space, and let $f_n : \mathcal{V} \to [-\infty, \infty]$ be a sequence of lower semi-continuous functions. Suppose that f_n epi-converges to f, and that $\mathrm{dom} f, \mathrm{dom} f_n \subseteq K$, $n = 1, 2, \ldots$ for some compact subset K of $\mathcal{V}$. Then*

$$\lim_{n\to\infty} (\inf f_n) = \inf f \tag{3.2.1}$$

and

$$\limsup(\arg\min f_n) \subset \arg\min f. \tag{3.2.2}$$

We note that epi-convergence is designed for an application to problems of minimization, whereas we are considering a problem of maximization. Consequently, we seek to prove epi-convergence of the sequence $f_n = -I_n$ to the function $f = -I$, i.e. we want to prove the following two inequalities:

$$\limsup_{n\to\infty} I_n(\pi_n) \leq I(\pi) \quad \text{for every sequence } \pi_n \to \pi \tag{3.2.3}$$

$$\liminf_{n\to\infty} I_n(\pi_n) \geq I(\pi) \quad \text{for some sequence } \pi_n \to \pi \tag{3.2.4}$$

The first inequality is (relatively) easy, and we present its proof here.

Proposition 6. *Suppose that* $\pi_n \to \pi$. *Then* $\limsup_{n\to\infty} I_n(\pi_n) \leq I(\pi)$.

Proof. If $\pi \notin \Pi(\mu, \nu)$, then there is nothing to prove. Consider a subsequence π_{n_k} with $\lim_{k\to\infty} I_{n_k}(\pi_{n_k}) = \limsup_{n\to\infty} I_n(\pi_n)$. The desired inequality is immediate unless $\pi_{n_k} \in \Pi(\mu_{n_k}, \nu_{n_k})$ for k large enough.

Let Θ_{n_k} be such that $\Theta_{n_k} \ll \pi_{n_k}$, $\frac{d\Theta_{n_k}}{d\pi_{n_k}} \leq (1-\alpha)^{-1}$, and $\mathbb{E}_{\Theta_{n_k}}[L] = \mathrm{ES}_{\alpha,\pi_{n_k}}[L]$ (Föllmer and Schied (2016), Theorem 4.52 and Remark 4.53). Since $\pi_{n_k} \to \pi$, the set $\{\pi_{n_k}, k \in \mathbb{N}\}$ is compact. Therefore, by Prokhorov's Theorem, for each $\varepsilon > 0$ there exists a compact K_ε such that $\pi_{n_k}(K_\varepsilon) \leq (1-\alpha)\varepsilon$ for all $k \in \mathbb{N}$. Thus, for all $k \in \mathbb{N}$:

$$\Theta_{n_k}(K_\varepsilon) = \int_{K_\varepsilon} \frac{d\Theta_{n_k}}{d\pi_{n_k}}\, d\pi_{n_k} \leq \varepsilon.$$

Therefore, again by Prokhorov's Theorem, the set Θ_{n_k} has a convergent subsequence $\Theta_{n_{k_j}}$ tending to some Θ. Lemma 6 from Ghossoub, Hall, and Saunders (2020) yields that $\Theta \ll \pi$ and $\frac{d\Theta}{d\pi} \leq (1-\alpha)^{-1}$, and therefore $\mathbb{E}_\Theta[L] \leq \mathrm{ES}_{\alpha,\pi}(L)$. Thus:

$$\begin{aligned}
\limsup_n I_n(\pi_n) &= \lim_j E_{\Theta_{n_{k_j}}}[L] - \frac{1}{c_{n_{k_j}}} H(\pi_{n_{k_j}} \,|\, \pi_{n_{k_j},\mathrm{ref}}) \\
&= \mathbb{E}_\Theta[L] - \lim_j \frac{1}{c_{n_{k_j}}} H(\pi_{n_{k_j}} \,|\, \pi_{n_{k_j},\mathrm{ref}}) \\
&\leq I(\pi), \qquad (3.2.5)
\end{aligned}$$

using that $\mathbb{E}_\Theta[L] \leq \mathrm{ES}_{\alpha,\pi}(L)$, and the joint lower semicontinuity of relative entropy (see Dupuis and Ellis (1997)). □

Corollary 3.2.1.1. *The functions* $f = -I$, *and* $f_n = -I_n$ *are lower semi-continuous.*

Proof. Lower semi-continuity of f follows from the above Proposition by taking $\mu_n \equiv \mu$, $\nu_n \equiv \nu$, $c_n \equiv c$, and $\pi_{n,\mathrm{ref}} \equiv \pi$. Lower semi-continuity of the f_n then follows immediately. □

3.3 Required Background Material from Probability Theory

For the proof of the second part of the epi-convergence (3.2.4), we need to introduce some additional concepts from probability theory.

Definition 3.3.1 (Kernel)**.** Suppose that $(S, \mathscr{S})$ and $(T, \mathscr{T})$ are measurable spaces. A **kernel** from $(S, \mathscr{S})$ to $(T, \mathscr{T})$ is a function $K : S \times \mathscr{T} \to [0, \infty]$ such that

1. $x \mapsto K(x, A)$ is a measurable function from S into $[0, \infty]$ for each $A \in \mathscr{T}$
2. $A \mapsto K(x, A)$ is a positive measure on $\mathscr{T}$ for each $x \in S$

If $(T, \mathscr{T}) = (S, \mathscr{S})$, then K is said to be a kernel on $(S, \mathscr{S})$. If $K(x, T) = 1$ for every $x \in S$, then K is called a probability kernel.

Definition 3.3.2 (Products of Kernels and Measures)**.** Let $(S, \mathscr{S}, \eta)$ be a probability space, and $(T, \mathscr{T})$ a measurable space.

1. By $\eta \otimes K$ we denote the measure on $\mathscr{S} \otimes \mathscr{T}$ defined for sets of the form $A \times B$, $A \in \mathscr{S}$, $B \in \mathscr{T}$ as:
$$\int_A K(x, B)\, d\eta(x),$$
and by extension to $\mathscr{S} \otimes \mathscr{T}$.
2. By ηK, we denote the second marginal of $\eta \otimes K$, i.e. the measure on $(T, \mathscr{T})$ defined by:
$$\eta K(B) = \int_S K(x, B)\, d\eta(x), \quad B \in \mathscr{T}.$$

The following fundamental result allows us to decompose a measure on $S \times T$ into its marginal distribution on S and a kernel K. It is proved in Kallenberg (2021) (it is a special case of Theorem 3.4 there).

Theorem 3.3.1 (Disintegration Theorem)**.** *Let $(S, \mathscr{S})$ be a measurable space, $(T, \mathscr{T})$ be a Borel space,*[1] *and m be a probability measure on $\mathscr{S} \otimes \mathscr{T}$. Let η be the marginal distribution of m on S. Then $m = \eta \otimes K$ for a probability kernel K.*

1. Measurable spaces $(T_1, \mathscr{T}_1)$ and $(T_2, \mathscr{T}_2)$ are Borel isomorphic if there exists a bijection $f : T_1 \to T_2$ such that both f and f^{-1} are measurable. A space is Borel if it is Borel isomorphic to a Borel set in $[0, 1]$. Kallenberg (2021), Theorem 1.8 shows that any Polish space equipped with its Borel sigma-algebra is a Borel space.

Definition 3.3.3 (Product of Kernels)**.** Suppose that $(S_i, \mathscr{S}_i)$ and $(T_i, \mathscr{T}_i)$, $i = 1, 2$ are measurable spaces, and that $K_i : S_i \times \mathscr{T}_i \to [0, \infty]$ are kernels. Then the product kernel $K_1 \otimes K_2 : (S_1 \times S_2) \times (\mathscr{T}_1 \otimes \mathscr{T}_2)$ is defined for $s_i \in S_i$ and sets of the form $B_1 \times B_2$, $B_i \in \mathscr{T}_i$ as:

$$K_1 \otimes K_2((s_1, s_2), (B_1 \times B_2)) = K_1(s_1, B_1) \cdot K_2(s_2, B_2)$$

and by extension to $\mathscr{T}_1 \otimes \mathscr{T}_2$.

We will need the following important distances on subsets of the set of probability measures on a metric space $\mathcal{V}$.

Definition 3.3.4 (p-Wasserstein distance)**.** Let $(\mathcal{V}, d_{\mathcal{V}})$ be a Polish space, and $p \in [1, \infty]$. Let $\mathcal{P}_p(\mathcal{V}) \subseteq \mathcal{P}(\mathcal{V})$ be the set of probability measures η with finite p-th moment, i.e., $\int d_{\mathcal{V}}(v, v_0)^p \, \eta(dv) < \infty$ for some $v_0 \in \mathcal{V}$. For $p = \infty$, we define $\mathcal{P}_\infty(\mathcal{V})$ as the measures with a bounded support. Define the p-Wasserstein distance $W_p(\eta, \eta')$ between $\eta, \eta' \in \mathcal{P}_p(\mathcal{V})$ as:

$$W_p(\eta, \eta') := \inf_{\pi \in \Pi(\eta, \eta')} \int d_{\mathcal{V}}(v, v')^p \pi(dv, dv'), \quad p \in [1, \infty),$$

$$W_\infty(\eta, \eta') := \inf_{\pi \in \Pi(\eta, \eta')} \operatorname*{ess\,sup}_{(V, V') \sim \pi} d_{\mathcal{V}}(V, V').$$

The following is proven in Theorem 7.12 in Villani (2003).

Theorem 3.3.2. *Let $(\mathcal{V}, d_{\mathcal{V}})$ be a Polish space, $p \in [1, \infty)$. Let $(\eta_n)_{n \in \mathbb{N}}$ be a sequence in $\mathcal{P}_p(\mathcal{V})$, and $\eta \in \mathcal{P}(\mathcal{V})$. Then $W_p(\eta_n, \eta) \to 0$ if and only if*

$$\int f \, d\eta_n \to \int f \, d\eta \tag{3.3.1}$$

for all continuous functions f such that there exist $C \in \mathbb{R}_+$ and $v_0 \in \mathcal{V}$ such that:

$$|f(v)| \leq C[1 + d(v_0, v)^p], \quad \forall v \in \mathcal{V}. \tag{3.3.2}$$

In particular, we see that $W_p(\eta_n, \eta) \to 0$ implies that $\eta_n \to \eta$.

Kernels, disintegration, and Wasserstein distances come together nicely in the following definition due to Eckstein and Nutz (2021).

Definition 3.3.5 (Shadow)**.** Fix $N \in \mathbb{N}$ and let (X_i, d_{X_i}), $i = 1, \ldots, N$ be Polish probability spaces with measures $\mu_i \in \mathcal{P}(X_i)$. Let $p \in [1, \infty]$, and $\mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i), i = 1, \ldots, N$. Let $\kappa_i \in \Pi(\mu_i, \tilde{\mu}_i)$ be a coupling attaining $W_p(\mu_i, \tilde{\mu}_i)$ and $\kappa_i = \mu_i \otimes K_i$ a disintegration. Given $\pi \in \Pi(\mu_1, \ldots, \mu_N)$, its **shadow** $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \ldots, \tilde{\mu}_N)$ is defined as the second marginal of $\pi \otimes K \in \mathcal{P}(X \times X)$, where the kernel $K : X \to \mathcal{P}(X)$ is defined as $K(x) = K_1(x_1) \otimes \ldots \otimes K_N(x_N)$.

p-Wasserstein distances on $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$ will always be understood to be taken with respect to $d_{\mathcal{W},p}$. Furthermore, following Eckstein and Nutz (2021), we define, for $\mu, \tilde{\mu} \in \mathcal{P}(\mathcal{X})$ and $\nu, \tilde{\nu} \in \mathcal{P}(\mathcal{Y})$

$$W_p(\mu, \nu; \tilde{\mu}, \tilde{\nu}) = \begin{cases} (W_p(\mu, \tilde{\mu})^p + W_p(\nu, \tilde{\nu})^p)^{1/p}, & p \in [1, \infty) \\ \max(W_\infty(\mu, \tilde{\mu}), W_\infty(\nu, \tilde{\nu})), & p = \infty. \end{cases} \tag{3.3.3}$$

For our purposes, the main importance of shadows is due to the following result of Eckstein and Nutz (2021).

Lemma 3.3.1. *Let $p \in [1, \infty]$ and $\mu, \tilde{\mu} \in \mathcal{P}(\mathcal{X})$ and $\nu, \tilde{\nu} \in \mathcal{P}(\mathcal{Y})$. Given $\pi \in \Pi(\mu, \nu)$, its shadow $\tilde{\pi} \in \Pi(\tilde{\mu}, \tilde{\nu})$ satisfies:*

$$W_p(\pi, \tilde{\pi}) = W_p(\mu, \nu; \tilde{\mu}, \tilde{\nu}), \tag{3.3.4}$$

$$H(\tilde{\pi} \,|\, \tilde{\mu} \otimes \tilde{\nu}) \leq H(\pi, \mu \otimes \nu). \tag{3.3.5}$$

3.4 Completion of the Proof of Epi-Convergence under Additional Assumptions

To complete the proof of epi-convergence of the functions $f_n = -I_n$ to $f = -I$, we require some additional assumptions on the loss function L and the reference measures $\pi_{\text{ref}}, \pi_{\text{ref},n}$. In particular, in order to use the results on shadows from Eckstein and Nutz (2021), we will from now on need to assume that $\pi_{\text{ref}} = \mu \otimes \nu$, and $\pi_{\text{ref},n} = \mu_n \otimes \nu_n$. We will furthermore assume that the cost function is Hölder continuous, in order to employ the following result from Pichler (2013).

Proposition 7. *Suppose that $L : \mathcal{W} \to \mathbb{R}$ is bounded and Hölder continuous, with constant $C_{L,\beta}$, i.e. $|L(w) - L(w')| \leq C_{L,\beta} \cdot d_{\mathcal{W}}(w, w')^\beta$, for some $\beta \leq 1$, and let $\pi, \pi' \in \mathcal{P}_p(\mathcal{W})$, $p \in [1, \infty)$. Then:*

$$|ES_{\alpha,\pi}(L) - ES_{\alpha,\pi'}(L)| \leq C_{L,\beta} \cdot W_p(\pi, \pi') \cdot (1 - \alpha)^{-\frac{\beta}{p}}. \tag{3.4.1}$$

Proposition 8. *Let $p \in [1, \infty)$, and suppose that:*

- *L is bounded and Hölder continuous.*
- *$W_p(\mu_n, \mu) \to 0$ and $W_p(\nu_n, \nu) \to 0$.*
- *$\pi_{\text{ref}} = \mu \otimes \nu$, and $\pi_{\text{ref},n} = \mu_n \otimes \nu_n$.*

Then exists a sequence $\pi_n \to \pi$ such that $\liminf_{n\to\infty} I_n(\pi_n) \geq I(\pi)$.

Proof. Suppose that $\pi \notin \Pi(\mu,\nu)$, so that $I(\pi) = -\infty$. Then the constant sequence $\pi_n \equiv \pi$ will have a subsequence π_{n_k} with $\pi_{n_k} \notin \Pi(\mu_{n_k},\nu_{n_k})$, and therefore the result will hold.

Now suppose that $\pi \in \Pi(\mu,\nu)$. Let $\pi_n \in \Pi(\mu_n,\nu_n)$ be the shadow of π. By Lemma (3.3.1), we have that $W_p(\pi_n,\pi) = W_p(\mu_n,\nu_n;\mu,\nu) \to 0$, and therefore, by Proposition 7, $ES_{\alpha,\pi_n}(L) \to ES_{\alpha,\pi}(L)$. Then:

$$\liminf_{n\to\infty} I_n(\pi_n) = \liminf_{n\to\infty}(ES_{\alpha,\pi_n}(L) - c_n^{-1}H(\pi_n|\pi_{n,\text{ref}})) \tag{3.4.2}$$

$$= ES_{\alpha,\pi}(L) - \limsup_{n\to\infty} c_n^{-1}H(\pi_n|\pi_{n,\text{ref}}) \tag{3.4.3}$$

$$\geq ES_{\alpha,\pi}(L) - c^{-1}H(\pi|\pi_{\text{ref}}) = I(\pi), \tag{3.4.4}$$

where, in the final line we have again used Lemma 3.3.1. □

Corollary 3.4.0.1. *Under the hypotheses of the proposition, the sequence of functions f_n epi-converges to f.*

3.5 Convergence of the Optimal Values and (Subsequences of) Optimal Solutions

We have shown epi-convergence of the f_n to f. To apply Theorem 3.2.1, it remains to show that there exists a compact set $K \subseteq \mathcal{W}$ such that $\text{dom}f, \text{dom}f_n \subseteq K$.

Proposition 9. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}(\mathcal{X})$, with $\mu_n \to \mu$, and $(\nu_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}(\mathcal{Y})$ with $\nu_n \to \nu$. Then:*

$$\Pi^* = \Pi(\mu,\nu) \cup \left(\bigcup_{n=1}^{\infty} \Pi(\mu_n,\nu_n)\right) \tag{3.5.1}$$

is relatively compact.

Proof. We first note that the sets $A = \{\mu,\mu_1,\mu_2,\ldots\}$ and $B = \{\nu,\nu_1,\nu_2,\ldots\}$ are compact in $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ respectively. Let $\varepsilon > 0$. By Prokhorov's Theorem, there exist compact sets $K_1 \subseteq \mathcal{X}$ and $K_2 \subseteq \mathcal{Y}$ such that $m_1(K_1^c) < \frac{\varepsilon}{2}$, and $m_2(K_2^c) < \frac{\varepsilon}{2}$ for any $m_1 \in A$, and $m_2 \in B$. Note that $K_1 \times K_2$ is compact, and that $(K_1 \times K_2)^c \subseteq (K_1^c \times \mathcal{Y}) \cup (\mathcal{X} \times K_2^2)$, so that for any $\pi \in \Pi^*$:

$$\pi(K_1 \times K_2) \leq \pi(K_1^c \times \mathcal{Y}) + \pi(\mathcal{X} \times K_2^c) \leq \varepsilon. \tag{3.5.2}$$

Another application of Prokhorov's Theorem implies the result. □

We can now give the main result of this chapter.

Theorem 3.5.1. *Let $p \in [1, \infty)$, and suppose that:*

- *L is bounded and Hölder continuous.*
- *$W_p(\mu_n, \mu) \to 0$ and $W_p(\nu_n, \nu) \to 0$.*
- *$\pi_{\text{ref}} = \mu \otimes \nu$, and $\pi_{\text{ref},n} = \mu_n \otimes \nu_n$.*

Then $V_n \to V$. Furthermore, $\limsup(\arg\max I_n) \subseteq \arg\max I$.

Proof. We have that $f_n = -I_n$ epi-converges to $f = -I$, and that $\text{dom} f, \text{dom} f_n$ are all contained in the (compact) closure of Π^*. The hypotheses of Theorem 3.2.1 are then satisfied, yielding the result. □

Chapter 4

Numerical Simulation

In this section, we aim at efficiently computing the entropy penalized primal problem (1.2.3). Throughout this section, we assume $|\mathcal{X}| = N_{\mathcal{X}}$ and $|\mathcal{Y}| = N_{\mathcal{Y}}$.

The penalized primal in the discrete setting is as follows:

$$\begin{aligned}
\max_{\substack{\pi\in\mathbb{R}^{N_{\mathcal{X}}}\times\mathbb{R}^{N_{\mathcal{Y}}}\\ \Theta\in\mathbb{R}^{N_{\mathcal{X}}}\times\mathbb{R}^{N_{\mathcal{Y}}}}} \quad & \sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}} L_{ij}\Theta_{ij} - \frac{1}{c_0}\sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}\pi_{ij}\log\left(\frac{\pi_{ij}}{(\pi_{\mathrm{ref}})_{ij}}\right)\\
\text{s.t.} \quad & \sum_{j=1}^{N_{\mathcal{Y}}}\pi_{ij} = \mu_i, && i = 1,\dots,N_{\mathcal{X}},\\
& \sum_{j=1}^{N_{\mathcal{X}}}\pi_{ij} = \nu_j, && j = 1,\dots,N_{\mathcal{Y}},\\
& \Theta_{ij} \le (1-\alpha)^{-1}\pi_{ij}, && i = 1,\dots,N_{\mathcal{X}}, \quad j = 1,\dots,N_{\mathcal{Y}},\\
& \sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}\Theta_{ij} = 1,\\
& \pi_{ij},\Theta_{ij} \ge 0, && i = 1,\dots,N_{\mathcal{X}}, \quad j = 1,\dots,N_{\mathcal{Y}}
\end{aligned} \tag{4.0.1}$$

and the penalized dual problem (although we do not consider computing it):

$$\min_{\substack{\varphi\in C_b(\mathcal{X}),\\ \psi\in C_b(\mathcal{Y}),\beta\in\mathbb{R}}} \quad \sum_i \varphi_i\mu_i + \sum_j \psi_j\nu_j + \beta + \frac{1}{c_0}\log\sum_i\sum_j e^{c_0\left((1-\alpha)^{-1}(L_{ij}-\beta)_+ - \varphi_i - \psi_j\right)}(\pi_{\mathrm{ref}})_{ij}.$$

A direct implementation of the entropy penalized primal problem (4.0.1) is both time and memory consuming: computing a single distance between a pair

of histograms of high dimensions can take more than a few seconds, not to even mention adding more constraints and complicating objective functions. For the problem in the space $\mathbb{R}^d \times \mathbb{R}^d$, with the use of the standard algorithms in optimization, such as network simplex and interior point methods, the complexity scales at least in $O(d^3 \log d)$ and is super-cubic in practice (see Pele and Werman (2009)).

4.1 Switching Min and Max

We apply the minimax theorem, Lemma 2.3.3, again to the primal problem, note that here we have equivalence between sup and max because of the finite dimensions of $\mathcal{X}$ and $\mathcal{Y}$ which guarantees the compactness of $\Pi(\mu, \nu)$:

$$\begin{aligned}
&\sup_{(\pi,\Theta)\in F_\alpha(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} L\, d\Theta - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right)\frac{d\pi}{d\pi_{\text{ref}}}\, d\pi_{\text{ref}} \\
&= \sup_{\pi\in\Pi(\mu,\nu)} \left\{ \min_{\beta\in\mathbb{R}} \left\{ \beta + \int_{\mathcal{X}\times\mathcal{Y}} (1-\alpha)^{-1}(L-\beta)_+\, d\pi \right\} - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right)\frac{d\pi}{d\pi_{\text{ref}}}\, d\pi_{\text{ref}} \right\} \\
&= \max_{\pi\in\Pi(\mu,\nu)} \left\{ \min_{\beta\in\mathbb{R}} \left\{ \beta + \int_{\mathcal{X}\times\mathcal{Y}} (1-\alpha)^{-1}(L-\beta)_+\, d\pi \right\} - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right)\frac{d\pi}{d\pi_{\text{ref}}}\, d\pi_{\text{ref}} \right\} \\
&= \inf_{\beta\in\mathbb{R}} \left\{ \beta + \sup_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} (1-\alpha)^{-1}(L-\beta)_+\, d\pi - \frac{1}{c_0}\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\pi_{\text{ref}}}\right)\frac{d\pi}{d\pi_{\text{ref}}}\, d\pi_{\text{ref}} \right\} \right\},
\end{aligned} \tag{4.1.1}$$

or in the discrete case,

$$\inf_{\beta\in\mathbb{R}} \left\{ \beta + \max_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}} (1-\alpha)^{-1}(L_{i,j}-\beta)_+\pi_{i,j} - \frac{1}{c_0}\sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}} \log\left(\frac{\pi_{i,j}}{\pi_{\text{ref}\,i,j}}\right)\pi_{i,j} \right\} \right\}. \tag{4.1.2}$$

4.2 Sinkhorn Algorithm - Computation of Max

An efficient algorithm for computing the above maximum part is Sinkhorn-Knopp matrix scaling algorithm, or Sinkhorn algorithm in short (see Sinkhorn and Knopp (1967) and Knight (2008)). The Sinkhorn algorithm employs the technique of vectorization and is amenable to large scale computation on parallel platforms such as GPGPU (see Cuturi (2013), Benamou et al. (2015)). Recent applications connect the computational optimal transport to the field of machine learning and partial differential equations (see Mena et al. (2017), Berman (2020)).

The optimal value of the maximum part inside (4.1.2), i.e.

$$
\begin{aligned}
\xi(\beta) :&= \max_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}(1-\alpha)^{-1}(L_{i,j}-\beta)_+\pi_{i,j} - \frac{1}{c_0}\sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}\log\left(\frac{\pi_{i,j}}{\pi_{\mathrm{ref}\,i,j}}\right)\pi_{i,j} \right\} \\
&= \max_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}(1-\alpha)^{-1}(L_{i,j}-\beta)_+\pi_{i,j} - \frac{1}{c_0}\sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}\log(\pi_{i,j})\pi_{i,j} \right. \\
&\qquad \left. + \frac{1}{c_0}\sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}\log(\pi_{\mathrm{ref}\,i,j})\pi_{i,j} \right\} \\
&= \max_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}\left((1-\alpha)^{-1}(L_{i,j}-\beta)_+ + \frac{1}{c_0}\log(\pi_{\mathrm{ref}\,i,j})\right)\pi_{i,j} \right. \\
&\qquad \left. - \frac{1}{c_0}\sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}\log(\pi_{i,j})\pi_{i,j} \right\} \\
&= -\min_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}} -\left((1-\alpha)^{-1}(L_{i,j}-\beta)_+ + \frac{1}{c_0}\log(\pi_{\mathrm{ref}\,i,j})\right)\pi_{i,j} \right. \\
&\qquad \left. + \frac{1}{c_0}\sum_{i=1}^{N_{\mathcal{X}}}\sum_{j=1}^{N_{\mathcal{Y}}}\log(\pi_{i,j})\pi_{i,j} \right\}
\end{aligned}
$$

can be efficiently computed using Sinkhorn-Knopp algorithm (see Cuturi (2013)):

Algorithm 1 Sinkhorn-Knopp's fixed point iteration (written in Matlab syntax)

Input: $M = -\left((1-\alpha)^{-1}(L_{i,j}-\beta)_+ + \frac{1}{c_0}\log(\pi_{\text{ref}\,i,j})\right)$, $\lambda = c_0$, $\mu \in \mathbb{R}^{N_\mathcal{X}}$, $\nu \in \mathbb{R}^{N_\mathcal{Y}}$

$I \leftarrow (\mu > 0)$

$\mu \leftarrow \mu(I)$

$M \leftarrow M(I,:)$

$K \leftarrow \exp(-\lambda * M)$

$x \leftarrow \text{ones}(\text{length}(\mu), \text{size}(\nu, 2))/\text{length}(\mu)$

while x changes **do** $x \leftarrow diag(1./\mu) * K * (\nu. * (1./(K^{'} * (1./x))))$

$y \leftarrow 1./x$

$z \leftarrow \nu. * (1./(K^{'} * y))$

Output: OptimalValue $= \text{sum}(y. * ((K. * M) * z))$, Optimizer $= \text{diag}(y) * e^{-\lambda M} * \text{diag}(z)$

4.3 Computation of Min

Now that we can solve $\xi(\beta)$ efficiently using the Sinkhorn's algorithm, we need to solve

$$\min_{\beta\in\mathbb{R}} \tau(\beta)$$

where

$$\tau(\beta) = \beta + \xi(\beta).$$

For $\beta \geq \|L\|_\infty$, $\xi(\beta) = 0$ and it holds when $\pi = \pi_{\text{ref}}$ solves the maximum problem. For $\beta \leq \min_{i,j} L_{i,j}$, we have $(L-\beta)_+ = L - \beta \geq 0$ and then

$$\begin{aligned}
\tau(\beta) &= \beta + \xi(\beta) \\
&= \beta + \max_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{i=1}^{N_\mathcal{X}} \sum_{j=1}^{N_\mathcal{Y}} \left((1-\alpha)^{-1}(L_{i,j}-\beta)_+ + \frac{1}{c_0}\log(\pi_{\text{ref}\,i,j}) \right) \pi_{i,j} \right. \\
&\qquad \left. - \frac{1}{c_0} \int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{N_\mathcal{X}} \sum_{j=1}^{N_\mathcal{Y}} \log(\pi_{i,j})\pi_{i,j} \right\} \\
&= \beta + \max_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{i=1}^{N_\mathcal{X}} \sum_{j=1}^{N_\mathcal{Y}} \left((1-\alpha)^{-1}(L_{i,j}-\beta) + \frac{1}{c_0}\log(\pi_{\text{ref}\,i,j}) \right) \pi_{i,j} \right. \\
&\qquad \left. - \frac{1}{c_0} \int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{N_\mathcal{X}} \sum_{j=1}^{N_\mathcal{Y}} \log(\pi_{i,j})\pi_{i,j} \right\} \\
&= \beta(1-(1-\alpha)^{-1}) \\
&\quad + \max_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{i=1}^{N_\mathcal{X}} \sum_{j=1}^{N_\mathcal{Y}} \left((1-\alpha)^{-1} L_{i,j} + \frac{1}{c_0}\log(\pi_{\text{ref}\,i,j}) \right) \pi_{i,j} - \frac{1}{c_0} \int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{N_\mathcal{X}} \sum_{j=1}^{N_\mathcal{Y}} \log(\pi_{i,j})\pi_{i,j} \right\}
\end{aligned}$$

whose value decreases as β increases.

Hence it is enough to only consider the following problem:

$$\min_{\min_{i,j} L_{i,j} \leq \beta \leq \max_{i,j} L_{i,j}} \tau(\beta) = \min_{\min_{i,j} L_{i,j} \leq \beta \leq \max_{i,j} L_{i,j}} \beta + \xi(\beta)$$

which can be solved efficiently using one dimensional algorithms, such as golden section method and Brent's method.

Note that the subdifferential of τ at β is $[\partial\tau_-(\beta), \partial\tau_+(\beta)]$ where

$$\partial\tau_-(\beta) = 1 - \min_{\pi\in\Pi^*(\beta)} \sum_{i,j} \pi_{i,j} \mathbf{1}\{L_{i,j} \geq \beta\}$$

$$\partial\tau_+(\beta) = 1 - \min_{\pi\in\Pi^*(\beta)} \sum_{i,j} \pi_{i,j}\mathbf{1}\{L_{i,j} > \beta\}$$

where $\Pi^*(\beta)$ is the set of optimizers for $\xi(\beta)$. Hence the optimal β is obtained when $0 \in [\partial\tau_-(\beta), \partial\tau_+(\beta)]$.

It is worth mentioning in Cuturi (2013) that when $\pi_{\text{ref}} = \mu \otimes \nu$, the product measure, then if one is interested in the following problem:

$$\begin{aligned} \max_{\pi\in\Pi(\mu,\nu)} \quad & \sum_{i,j} (L_{ij} - \beta)_+ \pi_{ij} \\ \text{s.t.} \quad & \sum_{i,j} \pi_{ij} \log \frac{\pi_{ij}}{(\mu \otimes \nu)_{ij}} \leq \eta \end{aligned}$$

then Sinkhorn's algorithm can be iteratively applied by changing the values of c_0 until $\int \log(d\pi)\, d\pi = \int \log(d\mu)\, d\mu + \int \log(d\nu)\, d\nu + \eta$ where $\pi = \arg\max \xi(\beta)$.

4.4 Numerical Results

In this section we give numerical examples as in Ghossoub, Hall, and Saunders (2020) part 6, but with an enlarged sample space due to the power of the Sinkhorn's algorithm.

Example 1 (Linear Loss with Gaussian Marginals)**.** There exists known results for the case where the objective function being linear and the marginals being Gaussian. According to Manistre and Hancock (2005), the limiting distribution of the ES based on empirical samples will be Gaussian.

Now let $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be defined as $L(x, y) = x + y$. Let $\mu \in \mathcal{P}(\mathcal{X})$ follows $N(0, 1)$, and let $\nu \in \mathcal{P}(\mathcal{Y})$ follows $N(0, 1)$. We sample d points from $(\mathcal{X}, \mu)$ and d points from $(\mathcal{Y}, \nu)$, set $\alpha = 0.9$, choose a penalty parameter $c_0 = 5$, and a reference measure $\pi_{\text{ref}} \sim \text{Unif}(\mathcal{X} \times \mathcal{Y})$. For $|\mathcal{X}| = |\mathcal{Y}| = d = 200$, the results are displayed in Figure 4.1 and Figure 4.2. For $|\mathcal{X}| = |\mathcal{Y}| = d = 300$, the results are displayed in Figure 4.3 and Figure 4.4.

Test for Normality	Shapiro-Wilk Test	D'Agostino and Pearson Test
d = 200	0.546	0.700
d = 300	0.546	0.599

Table 4.1: Statistical Test for Normality at 5% Significance Level - Example 1

In both cases we cannot reject the distribution of the optimal values being Gaussian at 5% significance level, which validate the known results. It is worth noting that as the sample size increases, the variance of the distribution decreases, which is also expected since the sample variance is negatively related to the sample size.

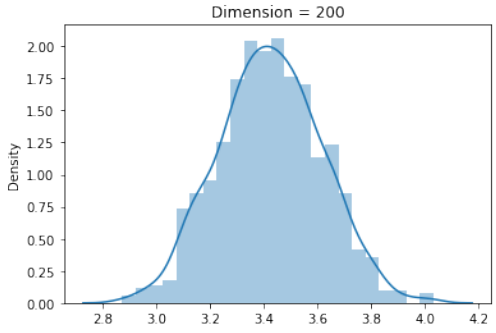

Figure 4.1: Histogram of the entropy regularized primal problem based on 1000 simulations.

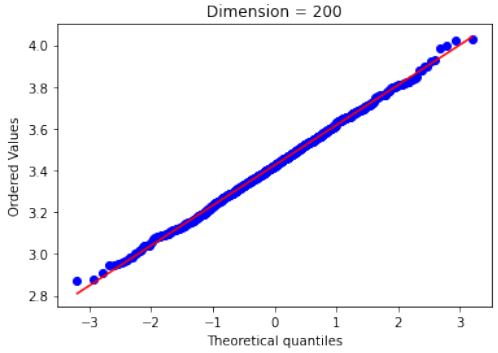

Figure 4.2: qq-plot of the entropy regularized problem based on 1000 simulations.

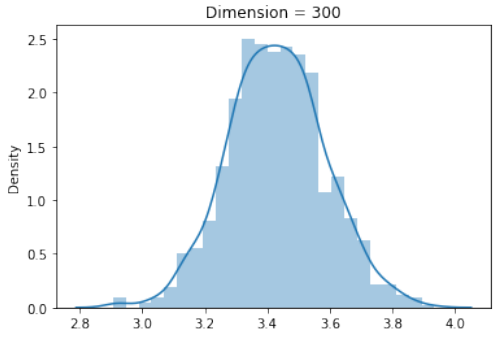

Figure 4.3: Histogram of the entropy regularized primal problem based on 1000 simulations.

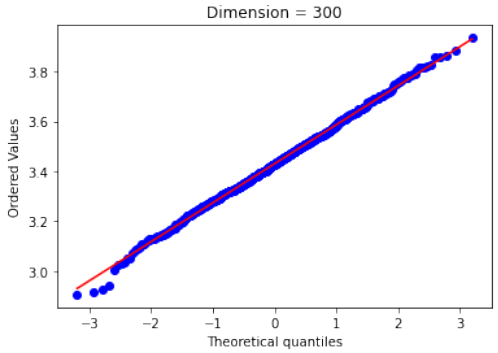

Figure 4.4: qq-plot of the entropy regularized problem based on 1000 simulations.

Example 2 (Counterparty Credit Risk)**.** A loss function related to counterparty credit risk (see Memartoluie (2017)) is as follows:

$$L(X,Y) = \max(Y_1,0)\cdot\Phi\left(\frac{\Phi^{-1}(PD_1)-\sqrt{\rho_1}X}{\sqrt{1-\rho_1}}\right)+\max(Y_2,0)\cdot\Phi\left(\frac{\Phi^{-1}(PD_2)-\sqrt{\rho_2}X}{\sqrt{1-\rho_2}}\right)$$

where Φ is the standard normal cumulative distribution function. This gives the systematic credit losses in the Vasicek model, with systematic credit factor $X \sim N(0,1)$, for a portfolio consisting of two counterparties with probabilities of default PD_1 and PD_2, systematic credit factor loadings ρ_1 and ρ_2, and counterparty portfolio values Y_1 and Y_2 (i.e. the distributions of exposures to two different counterparties of the bank), where we assume for simplicity that

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^1 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

We simulated 1000 values from each of X and Y, and set $\alpha = 0.9$, with the model parameters $PD_1 = PD_2 = 0.02$, $\rho_1 = \rho_2 = 0.2$, $r = 0.5$, $\mu_1 = 10, \mu_2 = -10$, $\sigma_1 = \sigma_2 = 5$. The histogram of 1000 realized optimal values is given in Figure 4.5, the qq plot is given in Figure 4.6.

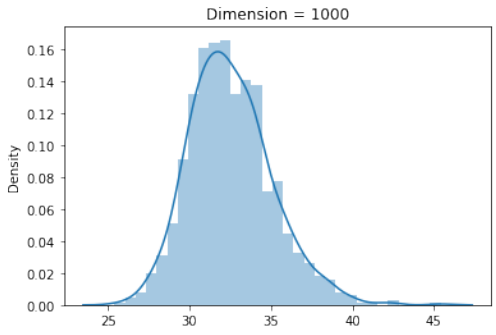

Figure 4.5: Histogram of the entropy regularized primal problem based on 1000 simulations.

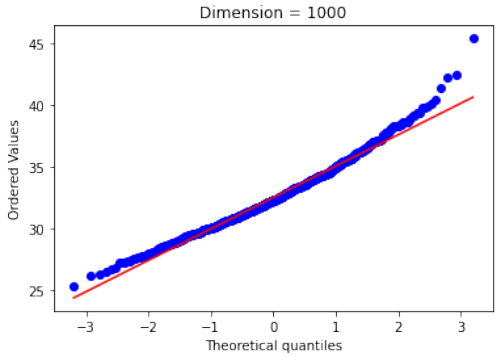

Figure 4.6: qq-plot of the entropy regularized problem based on 1000 simulations.

Test for Normality	Shapiro-Wilk Test	D'Agostino and Pearson Test
d = 1000	1.24e-09	1.828e-15

Table 4.2: Statistical Test for Normality at 5% significance level - Example 2

Note that we reject the resulting distribution of the optimal values being normal at 5% significance level, based on the plots and statistical tests.

Chapter 5

Conclusions

In this thesis, we start with an entropy penalized maximum expected shortfall problem, and call it our primal problem in the context of duality theory. We present the associated dual problem based on a conjecture using the standard convex analysis approach.

Next we prove the Kantorovich duality, including both weak duality and strong duality. For the strong duality, we first prove the case when the underlying spaces are not only Polish spaces but also compact, by utilizing the minimax theorem and the variational formula to rewrite and relative entropy. For the general strong duality, i.e. without the compactness of the underlying spaces, we follow the strategy in Villani (2003).

Then we prove the convergence of the optimal value of the primal problem in terms of the Wasserstein convergence of the marginals, the penalty threshold, and the reference measure (assumed to be the product measure), when the cost function is bounded and Hölder continuous. The key strategy we employed is the epi-convergence, or Gamma convergence. One direction of the epi-convergence comes relatively easily due to the lower-semicontinuity property of the Kullback-Leibler divergence.

For the other direction, we invoke the results from Eckstein and Nutz (2021) including the concepts of shadow and stochastic kernels. In particular, we need to make additional assumption: the loss function (begin bounded and Hölder continuous), the reference measure simply being the product measures of marginals, and the p-Wasserstein metricizing the weak convergence. Then we are able to construct a shadow of the probability measure and succeed in proving this direction.

We end with the discussion of the numerical simulation of the primal problem. A direct implementation of the problem would be costly and impractical. Hence

we apply the minimax theorem again and use Sinkhorn's algorithm to handle the max part in our problem, and then solve the min part which is a one dimensional optimization problem where we also discuss the interval for the optimal solution. Two numerical examples are presented, including the linear loss with gaussian marginals for which we have known results of the limiting distribution, and the counterparty credit risk example.

5.1 Directions for Future Research

There are a number of possible future research directions based on this thesis:

- As mentioned in the introduction, there are other (actually wider classes of) choices for the penalty term, and the result in this paper could be potentially further generalized.
- For the stability under weak convergence part, we assume really strong conditions on the loss function, i.e. Hölder continuity, as well as the p-Wasserstein distance that metricizes weak convergence. One can explore a relaxation of those conditions and prove the same or similar result.
- One can further explore the quantitative stability under weak convergence. Similar ideas have been done on the standard entropy regularized optimal transport, see Nutz and Wiesel (2021).
- There are other efficient numerical algorithms, including a greedy version of Sinkhorn algorithm, called Greenkhorn (see Altschuler, Weed, and Stromme (2021)) which allows to select and update columns and rows that most violate the polytope constraints, the Nys-Sink algorithm (see Altschuler et al. (2019)) based on low-rank approximation of the cost matrix using the Nyström method, etc. It would be interesting to compare the results returned by these algorithms.

References

Altschuler, J., F. Bach, A. Rudi, and J. Weed. 2019. "Massively scalable Sinkhorn distances via the Nyström method." *Advances in Neural Information Processing Systems 32.*

Altschuler, J.M., J.N. Weed, and A.J. Stromme. 2021. "Asymptotics for semi-discrete entropic optimal transport," arXiv: 2106.11862v1.

Artzner, P., F. Delbaen, J.M. Eber, and D. Heath. 1999. "Coherent measure of risk." *Mathematical Finance* 9:203–228.

Basel Committee on Banking Supervision. 2019. "Minimum capital requirements for market risk." *Technical report, Bank for International Settlements,* www.bis.org.

Beiglböck, M., P. Henry-Labordère, and F. Penkner. 2013. "Model-independent bounds for option prices: A mass transport approach." *Finance and Stochastics* 17:477–501.

Benamou, J.D., G. Carlier, M. Cuturi, L.Nenna, and G. Peyré. 2015. "Iterative Bregman projections for regularized transportation problems." *SIAM Journal on Scientific Computing* 37 (2): 1111–1138.

Berman, R.J. 2020. "The Sinkhorn algorithm, parabolic optimal transport and geometric Monge–Ampère equations." *Numerische Mathematik* 145:771–836.

Bhattacharya, B. 2006. "An Iterative Procedure for General Probability Measures to Obtain I-Projections onto Intersections of Convex Sets." *Ann. Stat* 34 (32): 878–902.

Bhattacharya, B., and R. Dykstra. 1995. "A general duality approach to I-projections." *Journal of Statistical Planning and Inference* 47 (MR1364994): 203–216.

Billingsley, P. 1999. *Convergence of probability measures, second edition.* SIAM.

Bonnans, J.F. 2019. *Convex and Stochastic Optimization.* Springer.

Borwein, J.M., and Q.J. Zhu. 2005. *Techniques of Variational Analysis.* Springer.

Boyd, S., and L. Vandenberghe. 2004. *Convex Optimization.* Cambridge University Press.

Brezis, H. 2010. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Springer.

Cuturi, M. 2013. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." In *Advances in Neural Information Processing Systems,* vol. 26. Curran Associates, Inc.

Cuturi, M., O. Teboul, and J.P. Vert. 2019. "Differential ranking and sorting using optimal transport." *Advances in Neural Information Processing Systems* 32.

Delbaen, F. 2002. "Coherent Risk Measures on General Probability Spaces." In *K. Sandmann and P. Schönbucher, editors, Advances in Finance and Stochastics. Essays in Honour of Dieter Sondermann,* 1–37.

Dudley, R.M. 2002. *Real Analysis and Probability.* Cambridge University Press.

Dupuis, P., and R.S. Ellis. 1997. *A Weak Convergence Approach to the Theory of Large Deviations.* Wiley Series in Probability / Statistics.

Eckstein, S., and M. Nutz. 2021. "Quantitative Stability of Regularized Optimal Transport," arXiv: 2110.06798 `[math.OC]`.

Ekeland, I., and R. Témam. 1999. *Convex Analysis and Variational Problems.* SIAM.

Föllmer, H., and A. Schied. 2016. *Stochastic Finance: An Introduction in Discrete Time.* de Gruyter, fourth edition.

Galichon, A. 2016. *Optimal Transport Methods in Economics.* Princeton University Press.

Garcia-Cespedes, J.C., J.A. de Juan Herrero, D. Rosen, and D. Saunders. 2010. "Effective modelling of wrong-way risk, counterparty credit risk capital and alpha in Basel II." *Journal of Risk Model Validation* 4 (1): 71–98.

Genevay, A. 2019. "Entropy-Regularized Optimal Transport for Machine Learning." *PhD thesis.*

Genevay, A., M. Cuturi, G. Peyré, and F. Bach. 2016. "Stochastic optimization for large-scale optimal transport." *Advances in Neural Information Processing Systems 29,* 3440–3448.

Ghossoub, M., J. Hall, and D. Saunders. 2020. "Maximum Spectral Measures of Risk with given Risk Factor Marginal Distributions," arXiv: 2010.14673 [q-fin.RM].

Glasserman, P., and L. Yang. 2018. "Bounding wrong-way risk in CVA calculation." *Mathematical Finance* 28 (1): 268–305.

Hansen, L.P., and T.J. Sargent. 2001. "Robust Control and Model Uncertainty." *American Economic Review, Papers and Proceedings* 91:60–66.

Hansen, L.P., T.J. Sargent, G. Turmuhambetova, and N. Williams. 2006. "Robust Control and Model Misspecification." *Journal of Economic Theory* 128:45–90.

Henry-Labordère, P. 2017. *Model-Free Hedging.* CRC Press.

Kallenberg, O. 2021. *Foundations of Modern Probability.* Third Edition. Springer.

Knight, P.A. 2008. "The Sinkhorn-Knopp algorithm: Convergence and applications." *SIAM Journal on Matrix Analysis and Applications* 30 (1): 261–275.

Maccheroni, F., M. Marinacci, and A. Rustichini. 2006. "Ambiguity Aversion, Robustness, and the Variational Representation of Preferences." *Econometrica* 74 (6): 1447–1498.

Malliavin, P. 1995. *Integration and Probability.* Springer.

Manistre, B.J., and G.H. Hancock. 2005. "Variance of the CTE estimator." *North American Actuarial Journal* 9 (2).

Markowitz, H.M. 1952. "Portfolio Selection." *Journal of Finance* 7:77–91.

McNeil, A.J., R. Frey, and P. Embrechts. 2015. *Quantitative Risk Management.* Princeton University Press.

Memartoluie, A. 2017. "Computational Methods in Finance Related to Distributions with Known Marginals." *PhD thesis.*

Memartoluie, A., D. Saunders, and T. Wirjanto. 2012. "Wrong-way risk bounds in counterparty credit risk management." *Journal of Risk Management in Financial Institutions* 10 (2): 150–163.

Mena, G., D. Belanger, G. Munoz, and J. Snoek. 2017. "Sinkhorn Networks: Using Optimal Transport Techniques to Learn Permutations." *NIPS Workshop in Optimal Transport and Machine Learning.*

Mérigot, Q., and B. Thibert. 2021. *Geometric Partial Differential Equations - Part II.* 22:133–212. Handbook of Numerical Analysis. Elsevier.

Nutz, M., and J. Wiesel. 2021. "Entropic Optimal Transport: Convergence of Potentials," arXiv: 2104.11720 [math.AP].

Pele, O., and M. Werman. 2009. "Fast and robust earth mover's distances." In *ICCV'09.*

Pichler, A. 2013. "Evaluations of Risk Measures for Different Probability Measures." *SIAM Journal on Optimization* 23 (1): 530–551.

Rachev, S.T., and L. Rüschendorf. 1998. *Mass Transportation Problems.* Springer.

Riedel, F. 2004. "Dynamic coherent risk measures." *Stochastic Processes and their Applications* 112 (2): 185–200.

Rockafellar, R.T., and S. Uryasev. 2000. "Optimization of conditional value-at-risk." *Journal of Risk* 2 (3): 21–41.

Rosen, D., and D. Saunders. 2012. "CVA the wrong way." *Journal of Risk Management in Financial Institutions* 5 (3): 252–272.

Rüschendorf, L. 2013. *Mathematical Risk Analysis.* Springer.

Simon, B. 2011. *Convexity: An Analytic Viewpoint.* Cambridge University Press.

Sinkhorn, R., and P. Knopp. 1967. "Concerning nonnegative matrices and doubly stochastic matrices." *Pacific J. Math* 27 (2): 343–348.

Tenetov, E., G. Wolansky, and R. Kimmel. 2018. "Fast Entropic Regularized Optimal Transport Using Semidiscrete Cost Approximation." *SIAM Journal on Scientific Computing* 40 (January): A3400–A3422.

Tobin, J. 1958. "Liquidity Preference as Behavior Toward Risk." *Review of Economic Studies* 25:65–86.

Villani, C. 2003. *Topics in Optimal Transportation.* American Mathematical Society.

Villani, C. 2008. *Optimal Transport Old and New.* Springer.

Weed, J. 2018. "An explicit analysis of the entropic penalty in linear programming." *Proceedings of Machine Learning Research* 75:1841–1855.

Zălinescu, C. 2002. *Convex Analysis in General Vector Spaces.* World Scientific.

APPENDICES

Appendix A

Important Theorems

A.1 Some Background Materials on Convex Optimization

In this section, we present some useful background material in (convex) optimization, summarized from Villani (2003), Bonnans (2019), Boyd and Vandenberghe (2004), and Ekeland and Témam (1999).

Definition A.1.1 (Convex optimization)**.** A convex optimization problem is one of the form

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & f_0(x) \\ \text{subject to} & f_i(x) \leq b_i, \quad i = 1, \ldots, m \end{array}$$

where the functions $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are convex, i.e., satisfy

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y)$$

for all $x, y \in \mathbb{R}^n$ and all $\alpha, \beta \in \mathbb{R}$ with $\alpha + \beta = 1, \alpha \geq 0, \beta \geq 0$.

Definition A.1.2 (Conjugate (or Polar) function)**.** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper function (not identically ∞). The function $f^* : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f^*(y) = \sup_{x \in \mathbf{dom} f} (y^T x - f(x))$$

is called the (convex) conjugate function (or Legendre transform) of the function f. The domain of the conjugate function consists of $y \in \mathbb{R}^n$ for which the supremum

is finite, i.e., for which the difference $y^T x - f(x)$ is bounded above on **dom** f. Note that f^* is a proper lower semi-continuous (l.s.c.) convex function, as it is the pointwise supremum of a family of affine functions of y, no matter whether f is convex or not.

Remark 1. Note here we do not assume f to be integrable on any subset of $\mathbb{R}^n$. However the integrability of f does not matter when $\mathbf{dom} f = \mathbb{R}^n$ (Villani (2003) section 2 lemma 2.10).

Before we proceed, we can generalize the concept into more general spaces. Let X be a vector space and X^* be its dual space, equipped with a bilinear mapping $\langle \cdot, \cdot \rangle$. Furthermore, X and X^* are equipped with topologies $\sigma(X, X^*)$ and $\sigma(X^*, X)$ which render them locally convex and Hausdorff.

Definition A.1.3 (Bi-conjugate (or Bi-polar) function)**.** The biconjugate of $f : X \to \mathbb{R} \cup \{\infty\} = \bar{\mathbb{R}}$ is the function $f^{**} : X \to \bar{\mathbb{R}}$ defined by

$$f^{**}(x) = \sup_{x^* \in X^*} \langle x^*, x \rangle - f^*(x^*)$$

Theorem A.1.1. *The biconjugate f^{**} is the supremum of the affine minorants of f. If $f : X \to \bar{\mathbb{R}}$ is proper l.s.c. convex, then $f = f^{**}$.*

Proof. See Bonnans (2019) proposition 1.43 and theorem 1.44. □

After introducing the concept of conjugacy, we are ready to discuss duality theory.

Consider the family of "primal" problems

$$\min_{x \in X} \varphi(x, y) - \langle x^*, x \rangle \tag{A.1.1}$$

where X and Y are Banach spaces, $\varphi : X \times Y \to \bar{\mathbb{R}}, x^* \in X^*, y \in Y$. We denote the associated value function by

$$v(y) = \inf_x (\varphi(x, y) - \langle x^*, x \rangle)$$

Note that

$$\begin{aligned} v^*(y^*) &= \sup_y (\langle y^*, y \rangle - v(y)) \\ &= \sup_y (\langle y^*, y \rangle - \inf_x (\varphi(x, y) - \langle x^*, x \rangle)) \\ &= \sup_y (\langle y^*, y \rangle + \langle x^*, x \rangle - \varphi(x, y)) \\ &= \varphi^*(x^*, y^*) \end{aligned}$$

Hence if follows that

$$v^{**}(y) = \sup_{y^* \in Y^*} (\langle y^*, y \rangle - \varphi^*(x^*, y^*))$$

We define the dual problem as

$$\max_{y^* \in Y^*} (\langle y^*, y \rangle - \varphi^*(x^*, y^*)) \tag{A.1.2}$$

Theorem A.1.2 (Weak Duality)**.**

$$val\ (A.1.1) = v^{**}(y) \leq v(y) = val\ (A.1.2)$$

Theorem A.1.3 (Strong Duality)**.** *If val* $(A.1.1) = val\ (A.1.2)$*, we have the strong duality between the two problems. This means that* $(x, y, y^*) \in X \times Y \times Y^*$ *such that*

$$\varphi(x, y) - < x^*, x > = < y^*, y > - \varphi^*(x^*, y^*)$$

is an optimality condition.

Remark 2. The strong duality has a close connection with a characterization of the subdifferential of value function. Here we omit this part and refer to the books for more details.

A.2 Ulam's Lemma

Theorem A.2.1 (Ulam's Lemma)**.** *Let* $(\mathcal{W}, \tau)$ *be a Polish space and* μ *a positive finite Borel measure on* $\mathcal{W}$*. Then for every* $\epsilon > 0$ *there exists a compact set* $K = K(\epsilon) \subset \mathcal{W}$ *s.t.* $\mu(\mathcal{W} \setminus K) < \epsilon$*.*

A.3 Prokhorov's Theorem

Definition A.3.1. A family $\mathcal{P}(\mathcal{W})$ of probability measures on a topological space $\mathcal{W}$ is said to be **tight** if for any $\epsilon > 0$ there exists a compact set $K_\epsilon \subset \mathcal{W}$ for which

$$\sup_{\mu \in \mathcal{P}} \mu[K_\epsilon^c] \leq \epsilon$$

Theorem A.3.1 (Prokhorov's Theorem)**.** *Let* $\mathcal{W}$ *be a Polish space, then any tight family in* $\mathcal{P}(\mathcal{W})$*, the set of probability measures on* $\mathcal{W}$*, is relatively sequentially compact in* $\mathcal{P}(\mathcal{W})$*: from any* $\{\mu_k\}$ *in* $\mathcal{P}(\mathcal{W})$ *one can extract a subsequence, still denoted* $\{\mu_k\}$*, and a probability measure* μ_* *on* $\mathcal{W}$*, such that for any* $\varphi \in C_b(\mathcal{W})$*,*

$$\lim_{k \to \infty} \int_{\mathcal{W}} \varphi \, d\mu_k = \int_{\mathcal{W}} \varphi \, d\mu_*$$

Proof. See Billingsley (1999) section 5 for a detailed proof. □

A.4 Riesz–Markov–Kakutani Representation Theorem

Theorem A.4.1 (Riesz–Markov–Kakutani Representation Theorem)**.** *Let $\mathcal{W}$ be a locally compact Hausdorff space. For any continuous linear functional ψ on $C_0(\mathcal{W})$, there is a unique regular countably additive complex Borel measure μ on $\mathcal{W}$ such that*

$$\forall f \in C_0(\mathcal{W}): \qquad \psi(f) = \int_{\mathcal{W}} f(w)\, d\mu(w).$$

The norm of ψ as a linear functional is the total variation of μ, that is

$$\|\psi\| = |\mu|(\mathcal{W})$$

Finally, ψ is positive if and only if the measure μ is non-negative.

Proof. See Malliavin (1995) chapter II section 5, or Dudley (2002) theorem 7.4.1, for a detailed proof. □

A.5 Urysohn's Lemma

Definition A.5.1 (Normal space)**.** A topological space $(\mathcal{W}, \mathcal{T})$ is called normal if for every pair of disjoint nonempty closed subsets $C, D \subset \mathcal{W}$ there exist disjoint open sets U, V such that $C \subset U$ and $D \subset V$.

Theorem A.5.1 (Urysohn's Lemma)**.** *A topological space $(\mathcal{W}, \mathcal{T})$ is normal if and only if for every pair of disjoint nonempty closed subsets $C, D \subset \mathcal{W}$ there is a continuous function $f : \mathcal{W} \to [0, 1]$ such that $f(w) = 0$, $\forall w \in C$ and $f(w) = 1$, $\forall w \in D$.*

A.6 Tietze Extension Theorem

Theorem A.6.1 (Tietze Extension Theorem)**.** *If X is a normal space and $f : A \to \mathbb{R}$ is a continuous map from a closed subset A of $\mathcal{W}$ into the real numbers $\mathbb{R}$ carrying the standard topology, then there exists a continuous extension of f to $\mathcal{W}$, which by definition is a continuous map $F : X \to \mathbb{R}$ with $F(a) = f(a)$ for all $a \in A$. Moreover, F may be chosen such that $\sup\{|f(a)| : a \in A\} = \sup\{|F(w)| : w \in \mathcal{W}\}$ that is, if f is bounded then F may be chosen to be bounded (with the same bound as f).*