# Convergence Results for Entropic Transfer Operators

## Master's Thesis

Faculty of Mathematics and Computer Science
University of Göttingen

September 2022

Supervisor: Prof. Dr. Bernhard Schmitzer
Second Assessor: Prof. Dr. Axel Munk
Author: Dimitrios Oikonomou

# ACKNOWLEDGEMENTS

First and foremost, I would like to thank Prof. Dr. Bernhard Schmitzer for his continuous support and for the detailed and helpful discussions that we had during the last year. His ideas significantly shaped the structure of this thesis and his help in many technical aspects was invaluable.

Furthermore, I want to thank Prof. Dr. Axel Munk for agreeing to be the second assessor for this thesis. Moreover, I am thankful to the Deutscher Akademischer Austauschdienst (DAAD) for its support during the academic year 2019-20.

Last but not least, I want to thank my friends and family for their constant support. Especially my sister, Akrivi.

# CONTENTS

# List of Figures

# CHAPTER 1

# INTRODUCTION

## 1.1 Problem Formulation and Related Work

Suppose that we are given a measure-preserving dynamical system $(X, F, \mu)$, i.e. we have a compact subspace $X \subseteq \mathbb{R}^d$, a continuous function $F : X \to X$ and a $F$-invariant probability measure $\mu \in \mathcal{P}(X)$, which means that $F_\sharp \mu = \mu$. Examples of such systems can be found in ergodic theory as well as in classical mechanics and thermodynamics. All measure-preserving systems are conservative systems, i.e. they satisfy Poincaré recurrence theorem [Poi90]. One very common and important theme in the study of dynamical systems is the approximation of the behavior of the system. This is typically done by direct simulation and this method can be very useful when a specific orbit has to be approximated for a finite period of time. However, if one is interested in the long term behavior or if the underlying system exhibits complicated dynamics then the information derived from one single trajectory is not always satisfying. In this work we want to explore other ways of approximating dynamical systems. The question that we are going to mainly deal with in this thesis can be vaguely stated as follows:

**Problem.** *Given a (measure preserving) dynamical system $(X, F, \mu)$, how can we find a stochastic map $F^N$ defined on a finite subspace $X^N \subseteq X$ such that $F^N$ extracts the most important dynamical features of the system?*

One approach to the above problem is known as Ulam's method, cf. [Ula60] and [Hsu81]. In this method, we consider a reference measure $m$ and a $m$-essentially disjoint covering $C_1, \dots, C_N$ of the support of $\mu$ (the $F$-invariant measure of the system). Then we can write $\mu$ in the following form

$$\mu(A) = \sum_{i=1}^{N} \pi_i \frac{m(C_i \cap A)}{m(C_i)},$$

for some coefficients $\pi_i \geq 0$. Since $F_\sharp \mu = \mu$ we get

$$\pi_i = \mu(C_i) = F_\sharp \mu(C_i) = \sum_{k=1}^{N} P_{ik} \pi_k,$$

where $P_{ik} = \frac{m(C_k \cap F^{-1}(C_i))}{m(C_k)}$, or equivalently $P\pi = \pi$ with $\pi = (\pi_1, \dots, \pi_N)$ and $P$ is the matrix with entries $P_{ik}$. This means that the coefficients $\pi_i$ are the stationary distribution

that corresponds to the Markov matrix $P$. Hence we can choose the stochastic map $F^N$ to be defined by the Markov matrix $P$. In this way, the map $F^N$ captures the tranfer of mass between the covering sets $C_i$, i.e. the $(i,j)$ entry of $F^N$ is the probability of moving from subset $C_j$ to subset $C_i$. One downside of this method is that it can be quite inefficient numerically, especially on high dimensions, as it requires the computation of the volume $m(C_k \cap F^{-1}(C_i))$.

A new and modern approach was presented in [JMS22] using entropic optimal transport. In this paper, in order to capture the dynamics of the dynamical system one considers the transfer operator $T : L^2(\mu) \to L^2(\mu)$ given by $Th = \frac{\mathrm{d}F_\sharp(h\mu)}{\mathrm{d}\mu}$ which encodes all the dynamic information of the system. Then after a small perturbation of the original deterministic system we can get a regularized tranfer operator $T^\varepsilon : L^2(\mu) \to L^2(\mu)$. According to [DJ99], the spectrum of the operator $T^\varepsilon$ can reveal the most essential dynamical features of the system. For example, if $F$ decomposes the space $X$ into $k$ almost invariant sets, then the spectrum of $T^\varepsilon$ contains $k$ real eigenvalues close to 1. Similarly if $F$ exhibits a $n$-cycle then the spectrum of $T^\varepsilon$ is close to the $n$-th roots of unity.

More specifically, in [JMS22] the "blurring" $T^\varepsilon : L^2(\mu) \to L^2(\mu)$, of the map $T$ is constructed by composing $T$ with a tranfer operator induced by the optimal $\varepsilon$-entropic transport plan between $\mu$ and $\mu$. The smoothed operator $T^\varepsilon$ is called an *entropic transfer operator*. With this regularization, it can be proved that $T^\varepsilon$ is compact. Now we assume a weak approximation $\mu^N$ of the $F$-invariant measure $\mu$ and we repeat the aforementioned construction, i.e. we construct approximating entropic tranfer operators $T^{N,\varepsilon} : L^2(\mu^N) \to L^2(\mu^N)$ from $\mu^N$. With this setup it can be proved that (an extension of) $T^{N,\varepsilon}$ converges to $T^\varepsilon$ in the operator norm (Theorem 4.1.1) and as a result (since $T^\varepsilon$ is compact) we also get convergence of the spectrum of $T^{N,\varepsilon}$ to the spectrum of $T^\varepsilon$ (Corollary 4.1.3). This approach has the benefit that that it can be solved numerically very efficiently, using the Sinkhorn algorithm (Section 2.3.1) and moreover it is more robust since it poses no assumption on the approximation $\mu^N \to \mu$. In addition, by tuning $\varepsilon$ (the regularization parameter) we can dictate the blurring in order to focus on specific dynamical features on certain scale lengths.

Both of the above proposals make the assumption that $F$ is a deterministic map. In this thesis our main goal is to study the case where the map $F$ is not deterministic but it is a stochastic map. If we follow exactly the same route as in [JMS22] then the best we can have is convergence of the operators $T^{N,\varepsilon}$ to $T^\varepsilon$ in the $L^2$ norm (Theorem 4.2.3). But this is not enough to give us convergence of the spectra. We will present an approach (Chapter 5) where we compose the transfer operator $T$ with *two* optimal entropic transport plans. This extra regularization is enough to give us convergence in the operator norm (Theorem 5.5.1) and thus convergence of the spectra (Corollary 5.5.3).

## 1.2 Contributions

The main contributions of this thesis are the following:

1. We formally present the theory of tranfer maps induced by general transport plans and we show how this construction generalizes the classical definition of transfer operators of dynamical systems. This work is built upon [JMS22, Section 4.1].

2. We formalize the stochastic setup of the above problem and we show that the direct generalization of the results in [JMS22] only gives us convergence in the $L^2$ norm.

3. We modify the ideas in [JMS22] by considering a double smoothing via entropic optimal transport. We formally present this theory and in the end we prove that using this construction we get convergence in the operator norm even in the stochastic case. We also numerically compare the spectra of the single smoothed operator with the double smoothed operator.

4. Finally we study the convergence rate of the kernels of the operators $T^{N,\varepsilon}$ and $T^{\varepsilon}$. While this is still work in progress, we give some preliminary results. Moreover, we present numerical expiriments regarding the convergence rate of the eigenvalues of the operators $T^{N,\varepsilon}$ and $T^{\varepsilon}$.

## 1.3 Structure of the thesis

**Chapter 2: Optimal Transport.** In this chapter we give a brief overview of the basic theory of Optimal Transport. We start by giving the definition of optimal transport according to Monge and Kantorovich. Then we state the dual problem and we give the definition of the Wasserstein distance. Moreover, we introduce the Entropic Optimal Transport and we also study the discrete case. We finish the chapter with the Sinkhorn Algorithm.

**Chapter 3: Transfer operators induced by a transport plan.** In this chapter we present a new construction of tranfer operators. We formally define it and we prove some fundamental results. Furthermore, we study this construction in the discrete case and in the end we compare this new construction of transfer operators with the classical one.

**Chapter 4: Entropic Transfer Operators.** In this chapter we start by recalling the constructions and results of the entropic transfer operators from [JMS22] for the deterministic case. Then we formally introduce the stochastic case and we try to prove the corresponding results presented in the deterministic case. In the end we prove that following the ideas in [JMS22] in this new setting can only go as far as proving convergence in the $L^2$ norm.

**Chapter 5: Double entropic regularization.** In this chapter we present the theory of double entropic regularization of the tranfer operator. With this new construction we are able to prove convergence in the operator norm and as a result convergence of the spectra. In the end we give a numerical comparison of the spectrum of the single smoothing versus the spectrum of the double smoothing.

**Chapter 6: Convergence Rates.** In this chapter we present some recent results about the convergence rates of the regularized and unregularized Optimal Transport. Using these results we establish a convergence rate about the kernels $t^{N,\varepsilon}$ and $t^{\varepsilon}$ of the operators $T^{N,\varepsilon}$ and $T^{\varepsilon}$, respectively. We finish this chapter with a numerical experiment about the convergence rates of the spectra of the operators $\hat{T}^{N,\varepsilon}$ and $T^{\varepsilon}$.

**Appendix A: Some Mathematical Background.** In this chapter we give a brief overview of the basic results about the pushforward of measures, the disintegration theorem and the modulus of continuity. We use many of these results repeatedly in the main chapters of the thesis.

## Notation and general assumptions

Throughout this thesis we make the following assumptions and we use the following notation unless it is specified otherwise.

- By $\mathbb{R}_{\geq 0}$ we denote the set of non negative real numbers.
- All metric spaces are *separable* metric spaces.
- All metric spaces are considered as measurable spaces via the Borel $\sigma$-algebra.
- For a metric space $X$, let $C(X)$ denote the space of continuous real valued functions over $X$, let $C_b(X)$ denote the space of bounded continuous real valued functions over $X$ and let $C^s(X)$ denote the space of $s$-continuously differentiable real valued functions over $X$.
- For a metric space $X$, let $\mathcal{M}(X)$ denote the space of *signed* Radon measures over $X$, let $\mathcal{M}_+(X)$ denote the space of non-negative Radon measures over $X$ and let $\mathcal{P}(X)$ denote the space of probability measures over $X$.
- For a metric space $X$, a measure $\mu \in \mathcal{M}(X)$ and a measurable function $h : X \to \mathbb{R}$, define the measure $h\mu \in \mathcal{M}(X)$ by

  $$(h\mu)(A) = \int_A h(x)\,\mathrm{d}\mu(x),$$

  for any $A \in \mathcal{B}(X)$.
- Let $X$ be a metric space and let $x \in X$. The Dirac measure for the point $x$ is defines by $\delta_x \in \mathcal{P}(X)$

  $$\delta_x(A) = \begin{cases} 1, & x \in A \\ 0, & x \notin A, \end{cases}$$

  for all $A \in \mathcal{B}(X)$.
- We use the symbol $|x|$ to denote the euclidean norm of a vector $x \in \mathbb{R}^d$ and the symbol $\|f\|_\infty$ to denote the supremum norm of a function $f : X \to \mathbb{R}$. Moreover we use the symbol $\|\mu\|_{TV}$ to denote the total variation distance of a measure $\mu$.

# CHAPTER 2

# OPTIMAL TRANSPORT

In this chapter we are going to briefly review some basic results of the theory of mathematical and computational optimal transport. We are going to use some of these results many times later. Extensive studies in the mathematical theory of optimal transport can be found in [AGS05], [San15], [Vil09] and [Vil21]. The main reference for the computational aspect of optimal transport is [PC19].

## 2.1 Basic Definitions

The main goal of Optimal Transport (OT) is to find the best possible way of moving a mass distribution into another. The standard example for this is to consider the following: Suppose that we have a pile of sand of some shape and we want to move it into a hole of a different shape. What is the most efficient way to do it? More formally, let $\mu \in \mathcal{P}(X)$ (the sand) and $\nu \in \mathcal{P}(Y)$ (the hole) be probability measures and let $c : X \times Y \to \mathbb{R}$ be a cost function which represent the cost (the effort) $c(x, y)$ of moving the sand from the point $x \in X$ into the point $y \in Y$. The objective is to find an optimal map $\Phi : X \to Y$ of moving the sand to the hole (i.e. we send the sand from point $x$ into the point $\Phi(x)$; this means that $\Phi_\sharp \mu = \nu$) such that the total cost is as low as possible. Mathematically, this problem was first introduced by Gaspard Monge (1746-1818) as presented in the following definition.

**Definition 2.1.1** (Monge formulation of OT, [Mon81])**.** Let $X$, $Y$ be metric spaces and let probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Let $c : X \times Y \to [0, +\infty]$ be a Borel measurable function, called the *cost function*. Then the *optimal transport map* is the map $\Phi : X \to Y$ that realizes the following infimum

$$\inf \left\{ \int_X c(x, \Phi(x)) \, \mathrm{d}\mu(x) \middle| \Phi : X \to Y, \Phi_\sharp \mu = \nu \right\}. \tag{2.1}$$

**Remark 2.1.2.** Although the above formulation is very natural there is an issue with it. An admissible map $\Phi$ might not exist in general. For example, suppose that $\mu$ is a Dirac measure while $\nu$ is not.

In order to overcome this issue, Leonid Kantorovich (1912-1986) proposed a slight relaxation of Monge's formulation. We first start with some basic definitions.

**Definition 2.1.3** (General transport plans)**.** Let $X$, $Y$ be metric spaces and let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. A *transport plan* between the measures $\mu$ and $\nu$ is a probability measure $\gamma \in \mathcal{P}(X \times Y)$ such that its marginals are $\mu$ and $\nu$. More formally we define the set of all transport plans from $\mu$ to $\nu$ as

$$\Gamma(\mu,\nu) = \left\{\gamma \in \mathcal{P}(X \times Y) \middle| \pi^1_\sharp \gamma = \mu, \pi^2_\sharp \gamma = \nu\right\}, \tag{2.2}$$

where $\pi^1 : X \times Y \to X$ and $\pi^2 : X \times Y \to Y$ are the projections to the first and second coordinate respectively.

Using the disintegration theorem (Theorem A.2.1) we can define the composition operation between transport plans.

**Proposition 2.1.4** (Composition of transport plans, [AGS05, Remark 5.3.3])**.** *Let $X_1$, $X_2$, $X_3$ be metric spaces and let $\mu_1 \in \mathcal{P}(X_1)$, $\mu_2 \in \mathcal{P}(X_2)$, $\mu_3 \in \mathcal{P}(X_3)$. Also let $\gamma_1 \in \Gamma(\mu_1,\mu_2)$ and $\gamma_2 \in \Gamma(\mu_2,\mu_3)$. Then there is a transport plan $\gamma_2 \circ \gamma_1 \in \Gamma(\mu_1,\mu_3)$, called the composition plan of $\gamma_1$ and $\gamma_2$, defined by*

$$\int_{X_1\times X_3} f(x_1,x_3)\,\mathrm{d}(\gamma_2\circ\gamma_1)(x_1,x_3) = \int_{X_2}\left(\int_{X_1\times X_3} f(x_1,x_3)\,\mathrm{d}(\gamma^1_{x_2}\times\gamma^2_{x_2})(x_1,x_3)\right)\mathrm{d}\mu_2(x_2)$$

*for any Borel function $f : X_1 \times X_3 \to \mathbb{R}$.*

**Remark 2.1.5.** For a probability measure $\mu \in \mathcal{P}(X)$, the set $\Gamma(\mu,\mu)$ equipped with the composition operation is a *monoid*.

Now we are ready for the Kantorovich formulation.

**Definition 2.1.6** (Kantorovich formulation of OT, [Kan42])**.** Let $X$, $Y$ be metric spaces and let probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Let $c : X \times Y \to [0,+\infty]$ be a Borel measurable cost function. Then the *optimal transport plan* is the plan $\gamma \in \Gamma(\mu,\nu)$ that realizes the following infimum

$$C(\mu,\nu) := \inf\left\{\int_{X\times Y} c(x,y)\,\mathrm{d}\gamma(x,y) \middle| \gamma \in \Gamma(\mu,\nu)\right\}. \tag{2.3}$$

**Remark 2.1.7.** It is not hard to prove that the set $\Gamma(\mu,\nu)$ is non empty (e.g. $\mu\times\nu \in \Gamma(\mu,\nu)$), so there always exist an admissible plan. Moreover, if $X$ and $Y$ are compact metric spaces and the cost function $c$ is continuous it also can be proved that there exists an optimal transport plan $\gamma_{opt}$, see [San15, Theorem 1.4].

### 2.1.1 Dual problem

The Kantorovich problem is a convex optimization problem under convex constraints. Hence, an important tool is the duality theory, which is typically used for convex problems.

**Proposition 2.1.8** (Dual Kantorovich problem, [San15, Section 1.2])**.** *Let $X$, $Y$ be metric spaces, let probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and let $c : X \times Y \to [0,+\infty]$ be a cost function. Then the dual problem of Equation* (2.3) *is given by*

$$\sup\left\{\int_X a(x)\,\mathrm{d}\mu(x) + \int_Y b(y)\,\mathrm{d}\nu(y) \middle| a \in C_b(X), b \in C_b(Y), a\oplus b \le c\right\}. \tag{2.4}$$

*Here $a \oplus b : X \times Y \to \mathbb{R}$ is defined by $(a\oplus b)(x,y) = a(x) + b(y)$.*

When the cost function $c$ is reasonably good, the dual problem admits an optimal solution and also the value of the primal objective function is equal to the value of the dual objective function as shown by the following theorem.

**Theorem 2.1.9** ([San15, Theorem 1.39]). *Let $X$, $Y$ be* complete *metric spaces (i.e. Polish metric spaces) and let probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Now let $c : X \times Y \to [0, +\infty]$ be a* uniformly continuous and bounded *cost function. Then the dual problem in Equation* (2.4) *admits an optimal solution $(a, b)$ and moreover we have that*

$$C(\mu,\nu) = \max\left\{\int_X a(x)\,\mathrm{d}\mu(x) + \int_Y b(y)\,\mathrm{d}\nu(y) \middle| a \in C_b(X), b \in C_b(Y), a \oplus b \leq c\right\}. \quad (2.5)$$

### 2.1.2 Wasserstein spaces

Using the Kantorovich formulation of OT we can define a new metric in a space of measures, called the *Wasserstein metric*.

**Definition 2.1.10.** Let $(X, d)$ be a metric space. For $x_0 \in X$ and $p \geq 1$, define the space

$$\mathcal{P}_p(X) = \left\{\mu \in \mathcal{P}(X) \middle| \int_X d^p(x, x_0)\,\mathrm{d}\mu(x) < +\infty\right\}.$$

Note that the finiteness of this integral does not depend on the choice of $x_0$. Now for $\mu, \nu \in \mathcal{P}_p(X)$ define

$$W_p(\mu,\nu) = \min\left\{\int_{X\times X} d^p(x,y)\,\mathrm{d}\gamma(x,y) \middle| \gamma \in \Gamma(\mu,\nu)\right\}^{\frac{1}{p}}.$$

Note that $W_p^p(\mu,\nu) = C(\mu,\nu)$ with the cost function $c(x,y) = d^p(x,y)$.

**Proposition 2.1.11** ([San15, Proposition 5.1]). *The quantity $W_p$ as defined in Definition 2.1.10 is a metric in the space $\mathcal{P}_p(X)$. It is called the* Wasserstein distance.

Now we state one of the most important properties of the Wasserstein distance.

**Theorem 2.1.12** ([San15, Theorem 5.9]). *Suppose that $X \subseteq \mathbb{R}^d$ is a compact metric space and let $(\mu_n)$ be a sequence of measures in $\mathcal{P}_p(X)$ and $\mu \in \mathcal{P}_p(X)$. Then the sequence $(\mu_n)$ converges weakly to $\mu$ if and only if $W_p(\mu_n, \mu) \to 0$.*

## 2.2 Entropic regularization

While the Kantorovich problem is a convex problem, it is not a strictly convex problem. For the applications, especially for the numerics, it desirable to have a strictly convex objective function. In order to make the Kantorovich problem strictly convex we are going to add regularization. The regularization function in this case is the probabilistic entropy. With this regularization we can get fast numerical algorithms as well as better convergence rates as we will see in Chapter 6.

**Definition 2.2.1** (Probabilistic entropy). Let $X$, $Y$ be metric spaces and consider $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Now for any $\gamma \in \Gamma(\mu,\nu)$ define

$$\mathbf{H}(\gamma) = \mathbf{H}(\gamma \mid \mu,\nu) = \begin{cases} \int_{X\times Y}\left(\log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}(\mu\times\nu)}\right) - 1\right)\mathrm{d}\gamma(x,y), & \text{if } \gamma \ll \mu\times\nu \\ \infty, & \text{otherwise.} \end{cases} \quad (2.6)$$

**Definition 2.2.2** (Entropic Optimal Transport)**.** Let $X$, $Y$ be metric spaces and consider $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Given a measurable cost function $c : X \times X \to \mathbb{R}$ the Entropic regularization of the Optimal Transport (EOT) with regularization parameter $\varepsilon > 0$ is the minimization problem

$$C^\varepsilon(\mu,\nu) := \inf\left\{\int_{X\times Y} c(x,y)\,\mathrm{d}\gamma(x,y) + \varepsilon \mathbf{H}(\gamma) \middle| \gamma \in \Gamma(\mu,\nu)\right\}, \tag{2.7}$$

Note that if $\varepsilon = 0$, we have $C^0(\mu,\nu) = C(\mu,\nu)$, i.e. we get the unregularized version.

A very useful property of EOT is that it always have an optimal solution and this solution has a closed formula.

**Proposition 2.2.3** ([Nut21, Theorem 4.2])**.** *The entropic regularization problem in Equation* (2.7) *has a unique minimizer $\gamma^\varepsilon_{opt}$ that can be written in the form*

$$\gamma^\varepsilon_{opt} = g^\varepsilon(\mu \times \nu), \tag{2.8}$$

*where $g^\varepsilon : X \times Y \to [0,1]$ is defined by*

$$g^\varepsilon(x,y) = \exp\left(\frac{-c(x,y) + a(x) + b(y)}{\varepsilon}\right), \tag{2.9}$$

*for some measurable functions $a : X \to \mathbb{R}$ and $b : Y \to \mathbb{R}$. We call the functions $a$, $b$ the* entropic potentials*.*

**Remark 2.2.4.** By Remark A.2.3 we get that the measure $(\gamma^\varepsilon_{opt})_y = [g^\varepsilon(\mu\times\nu)]_y = g^\varepsilon(-,y)\mu$ are probability measures, thus $\int_X \mathrm{d}g^\varepsilon(-,y)\mu = 1$ for all $y \in Y$. Hence we have

$$\int_X g^\varepsilon(x,y)\,\mathrm{d}\mu(x) = 1, \text{ for almost all } y \in Y \tag{2.10}$$

and similarly

$$\int_Y g^\varepsilon(x,y)\,\mathrm{d}\nu(y) = 1, \text{ for almost all } x \in X. \tag{2.11}$$

These equations give the following formulas for the optimal entropic potentials $a$ and $b$

$$a(x) = -\varepsilon \cdot \log\left(\int_X \exp\left(\frac{-c(x,y) + b(y)}{\varepsilon}\right)\mathrm{d}\nu(y)\right), \quad \mu\text{-a.e.} \tag{2.12}$$

and

$$b(y) = -\varepsilon \cdot \log\left(\int_Y \exp\left(\frac{-c(x,y) + a(x)}{\varepsilon}\right)\mathrm{d}\mu(x)\right), \quad \nu\text{-a.e.} \tag{2.13}$$

Now, as we did in the unregularized version, there is a dual problem for Equation (2.7).

**Proposition 2.2.5** ([Nut21, Theorem 4.7])**.** *Let $c \in L^1(\mu \times \nu)$. Then*

$$\begin{aligned} C^\varepsilon(\mu,\nu) = \sup_{u\in L^1(\mu), v\in L^1(\nu)} &\int_X u(x)\,\mathrm{d}\mu(x) + \int_Y v(y)\,\mathrm{d}\nu(y) \\ &- \varepsilon\left(\int_{X\times Y} \exp\left(\frac{-c(x,y) + u(x) + v(y)}{\varepsilon}\right)\mathrm{d}(\mu\times\nu)(x,y) - 1\right) \end{aligned}$$

*The supremum is attained by the entropic potentials $a \in L^1(\mu)$, $b \in L^1(\nu)$ from Proposition 2.2.3 and in this case we have*

$$C^\varepsilon(\mu,\nu) = \int_X a(x)\,\mathrm{d}\mu(x) + \int_Y b(y)\,\mathrm{d}\nu(y)$$

*The maximizers are almost surely unique up to an additive constant.*

Now we are going to examine how EOT behaves with respect to weak convergence.

**Proposition 2.2.6.** *Let $(X,d)$ be a* compact *metric space. Consider $\varepsilon > 0$ and a* continuous *cost function $c : X \times X \to \mathbb{R}$. Let $(\mu_n)_n$ and $(\nu_n)_n$ be sequences of probability measures such that $\mu_n \to \mu$ and $\nu_n \to \nu$ weakly in $\mathcal{P}(X)$. Suppose that $\gamma_n^\varepsilon = g_n^\varepsilon(\mu_n \times \nu_n)$ is the optimal entropic plan between $\mu_n$ and $\nu_n$ with respect to c, and analogously suppose that $\gamma^\varepsilon = g^\varepsilon(\mu \times \nu)$ to is the optimal entropic plan between $\mu$ and $\nu$ with respect to c. Then the family of functions $(g_n^\varepsilon)_n$ is uniformly equicontinuous and $g_n^\varepsilon \to g^\varepsilon$ uniformly as $n \to \infty$.*

*Proof.* This proof can be essentially found as part of the proof of [JMS22, Proposition 1]. Let $a_n^\varepsilon$ and $b_n^\varepsilon$ be the entropic potentials of the EOT problem between $\mu_n$ and $\nu_n$. By Remark 2.2.4 we have $\mu_n$-almost everywhere that

$$a_n^\varepsilon(x) = -\varepsilon \cdot \log\left(\int_X \exp\left(\frac{-c(x,y) + b_n^\varepsilon(y)}{\varepsilon}\right) \mathrm{d}\nu_n(y)\right).$$

Obviously we can extend this definition for *all* $x \in X$. Since $X$ is compact and $c$ is continuous we get that $c$ is uniformly continuous hence there exists a modulus of continuity $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ (cf. Definition A.3.1), i.e. a continuous, increasing and concave function with $\omega(0) = 0$ such that

$$|c(x,y) - c(x',y')| \leq \omega\left(\sqrt{d(x,x')^2 + d(y,y')^2}\right).$$

Now it is easy to see that $a_n^\varepsilon$ has the same modulus of continuity with $c$. Indeed,

$$\begin{aligned} a_n^\varepsilon(x') &= -\varepsilon \cdot \log\left(\int_X \exp\left(\frac{-c(x',y) + b_n^\varepsilon(y)}{\varepsilon}\right) \mathrm{d}\nu_n(y)\right) \\ &\leq -\varepsilon \cdot \log\left(\int_X \exp\left(\frac{-c(x,y) - \omega(d(x,x')) + b_n^\varepsilon(y)}{\varepsilon}\right) \mathrm{d}\nu_n(y)\right) \\ &= a_n^\varepsilon(x) + \omega(d(x,x')). \end{aligned}$$

Hence the family $(a_n^\varepsilon)_n$ is uniformly equicontinuous since all of the functions $a_n^\varepsilon$ have a common modulus of continuity. By fixing the additive shift invariance, e.g. $a_n^\varepsilon(x_0) = 0$ for some $x_0 \in X$, and the fact that $X$ is compact we get that $(a_n^\varepsilon)_n$ is also uniformly bounded. Hence by the Arzela-Ascoli theorem (cf. Theorem A.3.4), there exists a uniformly convergent subsequence of $(a_n^\varepsilon)_n$ into some function $a^\varepsilon : X \to \mathbb{R}$. Same for the family $(b_n^\varepsilon)_n$. By going into the limit (uniform convergence of $a_n^\varepsilon$, $b_n^\varepsilon$ and the weak limit $\mu_n \to \mu$) we see that $a^\varepsilon$ and $b^\varepsilon$ are the entropic potentials for the regularized problem between $\mu$ and $\nu$. Hence $a^\varepsilon$ and $b^\varepsilon$ are unique up to additive constant shift. But we already have $a^\varepsilon(x_0) = \lim_{n\to\infty} a_n^\varepsilon(x_0)$. So we get that the whole sequence $(a_n^\varepsilon)_n$ converges uniformly to $a^\varepsilon$. Same for $(b_n^\varepsilon)_n$. The extension of $(a_n^\varepsilon, b_n^\varepsilon)$ carries over to $g_n^\varepsilon(x,y) = \exp\left(\frac{-c(x,y)+a^\varepsilon(x)+b^\varepsilon(y)}{\varepsilon}\right)$. As a result we get that $g_n^\varepsilon$ converges uniformly to $g$ and that all $(g_n^\varepsilon)_n$ have a common modulus of continuity (obtained by combining the moduli of $a_n^\varepsilon$, $b_n^\varepsilon$ and $c$) which means that the family $(g_n^\varepsilon)_n$ is uniformly equicontinuous. □

## 2.3 Discretization

The study of computational optimal transport begins when we assume that we are working with discrete spaces or measures. For this section suppose that $X = \{x_1, \dots, x_M\}$ and $Y = \{y_1, \dots, y_N\}$. Then

$$\mu = \sum_{i=1}^M \mu_i \delta_{x_i}, \text{ with } \mu_i \geq 0 \text{ and } \sum_{i=1}^M \mu_i = 1$$

and

$$\nu = \sum_{j=1}^{N} \nu_j \delta_{y_j}, \text{ with } \nu_j \geq 0 \text{ and } \sum_{j=1}^{N} \nu_j = 1.$$

Equivalently, we can identify $\mu$ and $\nu$ with the vectors $(\mu_1, \dots, \mu_M) \in \mathbb{R}^M$ and $(\nu_1, \dots, \nu_N) \in \mathbb{R}^N$. In this setting a plan $\gamma$ can be identified with a matrix $(\gamma_{ij})_{i,j}\mathbb{R}^{M\times N}$ with $\gamma_{ij} \geq 0$ and Equation (2.2) becomes

$$\Gamma(\mu,\nu) = \left\{ \gamma = \sum_{i=1}^{M}\sum_{j=1}^{N} \gamma_{ij}\delta_{(x_i,y_j)} \in \mathcal{P}(X\times Y) \Big| \sum_{j=1}^{N} \gamma_{ij} = \mu_i, \sum_{i=1}^{M} \gamma_{ij} = \nu_j \forall i,j \right\}. \tag{2.14}$$

**Remark 2.3.1** (Discrete disintegration)**.** Suppose that $\gamma = \sum_{i,j} \gamma_{ij}\delta_{(x_i,y_j)}$ and so $\mu = \sum_i \mu_i \delta_{x_i}$ with $\mu_i = \sum_j \gamma_{ij}$ and $\nu = \sum_j \nu_j \delta_{y_j}$ with $\nu_j = \sum_i \gamma_{ij}$. Then in this setting, the disintegration takes the following form

$$\sum_{i,j} f(x_i,y_j)\gamma_{ij} = \sum_i \left( \sum_j f(x_i,y_j)\gamma_{x_i}(y_j) \right) \mu_i,$$

$$\sum_{i,j} f(x_i,y_j)\gamma_{ij} = \sum_j \left( \sum_i f(x_i,y_j)\gamma_{y_j}(x_i) \right) \nu_j.$$

Hence $\gamma_{x_i}(y_j) = \frac{\gamma_{ij}}{\mu_i}$ and $\gamma_{y_j}(x_i) = \frac{\gamma_{ij}}{\nu_j}$.

In this discrete setting, the Kantorovich problem in Equation (2.3) becomes

$$C(\mu,\nu) := \inf \left\{ \langle c,\gamma\rangle \Big| \gamma \in \Gamma(\mu,\nu) \right\}, \tag{2.15}$$

where we have identified $c : X \times Y \to \mathbb{R}$ to the matrix $c \in \mathbb{R}^{M\times N}$ with $c_{ij} = c(x_i, y_j)$ and $\langle \cdot,\cdot \rangle$ is the inner product. The dual Kantorovich problem in Equation (2.4) becomes

$$C(\mu,\nu) = \max \left\{ \langle a,\mu\rangle + \langle b,\nu\rangle \Big| a \in \mathbb{R}^M, b \in \mathbb{R}^N, a_i + b_j \leq c_{ij} \right\}, \tag{2.16}$$

and the entropic Kantorovich problem in Equation (2.7) becomes

$$C^\varepsilon(\mu,\nu) := \inf \left\{ \sum_{ij} \gamma_{ij} \left( c_{ij} + \varepsilon \log \left( \frac{\gamma_{ij}}{\mu_i\nu_j} - 1 \right) \right) \Big| \gamma \in \Gamma(\mu,\nu) \right\}. \tag{2.17}$$

The optimal plan for the discrete entropic OT is given by

$$\left(\gamma^\varepsilon_{opt}\right)_{ij} = \exp\left( \frac{-c_{ij} + a_i + b_j}{\varepsilon} \right) \mu_i\nu_j, \tag{2.18}$$

where the vectors $a \in \mathbb{R}^M$ and $b \in \mathbb{R}^N$ represent the optimal entropic potentials and for these potentials we have the following formulas according to the Equations (2.12) to (2.13)

$$a_i = -\varepsilon \cdot \log \left( \sum_{j=1}^{N} \exp\left( \frac{-c_{ij} + b_j}{\varepsilon} \right) \nu_j \right), \tag{2.19}$$

$$b_j = -\varepsilon \cdot \log \left( \sum_{i=1}^{M} \exp\left( \frac{-c_{ij} + a_i}{\varepsilon} \right) \mu_i \right). \tag{2.20}$$

### 2.3.1 Sinkhorn algorithm

The Sinkhorn algorithm is a fast and efficient algorithm for calculating the optimal entropic plan in the discrete case. It was firstly introduced in [Yul12] and the convergence was proved in [Sin64], hence the name. More recently, [Cut13] showed that Sinkhorn's algorithm work very well numerically, especially on GPUs. The last paper can be considered as the starting point of computational Optimal Transport, as it attracted the interest of applied data sciences in OT.

The Sinkhorn algorithm is an iterative algorithm that approximates the optimal entropic duals $a$ and $b$. Once we have an approximation of $a$ and $b$, we plug them in Equation (2.18) to get an approximation of the optimal plan. The algorithm works as follows: Start with an initial random vector $b^{(0)} = \left(b_1^{(0)}, \dots, b_N^{(0)}\right)^\top \in \mathbb{R}^N$ and for $k = 1, 2, \dots$ we set

$$a_i^{(k)} = -\varepsilon \cdot \log\left(\sum_{j=1}^{N} \exp\left(\frac{-c_{ij} + b_j^{(k-1)}}{\varepsilon}\right) \nu_j\right),$$

$$b_j^{(k)} = -\varepsilon \cdot \log\left(\sum_{i=1}^{M} \exp\left(\frac{-c_{ij} + a_i^{(k)}}{\varepsilon}\right) \mu_i\right),$$

until a stopping criterion terminates the algorithm.

# CHAPTER 3

# TRANSFER OPERATORS INDUCED BY A TRANSPORT PLAN

In this chapter we are going to present a general construction of transfer operators induced by general transport plans. To the best of our knowledge, this was first introduced in Section 4.1 of [JMS22]. We further extend these ideas by proving that the construction is functorial in a certain sense and we also study the discrete case. Moreover, we show that the classical definition of a transfer operator of a dynamical system is just a special case of this construction. We are going to use the results of this chapter many times later.

## 3.1 Basic definitions and results

**Definition 3.1.1.** Let $X$, $Y$ be metric spaces. Consider probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and let $\gamma \in \Gamma(\mu, \nu)$. The transport plan $\gamma$ induces a linear operator $G = G(\gamma) : L^p(X, \mu) \to L^p(Y, \nu)$ with $p \geq 1$, defined by

$$Gh = \frac{\mathrm{d}\pi^2_\sharp(H\gamma)}{\mathrm{d}\pi^2_\sharp\gamma} = \frac{\mathrm{d}\pi^2_\sharp(H\gamma)}{\mathrm{d}\nu},$$

where $\pi^2 : X \times Y \to Y$ is the projection to the second coordinate and $H(x, y) = h(x)$ for $h \in L^p(X, \mu)$. Note that $H\gamma \ll \gamma$, so $\pi^2_\sharp(H\gamma) \ll \pi^2_\sharp\gamma$ hence we indeed have a Radon-Nikodym derivative.

First we show that this map is well defined.

**Lemma 3.1.2.** *The map $G : L^p(X, \mu) \to L^p(Y, \nu)$ with $p \geq 1$, is well defined.*

*Proof.* Let $h \in L^p(X, \mu)$. We need to show that $Gh \in L^p(Y, \nu)$. For this, apply Lemma A.2.4 with $A = X \times Y$, $B = Y$, $\alpha = \gamma$, $\beta = \nu$, $F = \pi^2$ and $f(x) = |x|^p$. Then we get

$$\begin{aligned}\int_Y |Gh|^p \,\mathrm{d}\nu &= \int_Y \left|\frac{\mathrm{d}\pi^2_\sharp(H\gamma)}{\mathrm{d}\pi^2_\sharp\gamma}\right|^p \mathrm{d}\pi^2_\sharp\gamma \\ &\leq \int_{X\times Y} \left|\frac{\mathrm{d}(H\gamma)}{\mathrm{d}\gamma}\right|^p \mathrm{d}\gamma\end{aligned}$$

$$
\begin{aligned}
&= \int_{X\times Y} |H(x,y)|^p \,\mathrm{d}\gamma(x,y) \\
&= \int_{X\times Y} \left|h(\pi^1(x,y))\right|^p \,\mathrm{d}\gamma(x,y) \\
&= \int_X |h(x)|^p \,\mathrm{d}\pi^1_\sharp\gamma(x) \\
&= \int_X |h|^p \,\mathrm{d}\mu < +\infty,
\end{aligned}
$$

as wanted. □

**Remark 3.1.3** (Universal property of $G$)**.** Using the Radon-Nikodym theorem we can see that the map $G$ satisfies the following *universal property*, i.e. the map $G$ can be characterized by the following equation:

$$\int_Y \varphi(y)(Gh)(y)\,\mathrm{d}\nu(y) = \int_{X\times Y} \varphi(y)h(x)\,\mathrm{d}\gamma(x,y), \tag{3.1}$$

for all $\varphi \in C(Y)$ and $h \in L^p(X,\mu)$.

**Lemma 3.1.4.** *The map $G : L^p(X,\mu) \to L^p(Y,\nu)$ is linear.*

*Proof.* This is immediate by the universal property of $G$ and the linearity of the integral. □

**Remark 3.1.5** (Explicit formula for $G$)**.** Using the universal property of $G$ and disintegration we get

$$
\begin{aligned}
\int_X \varphi(y)(Gh)(y)\,\mathrm{d}\nu(y) &= \int_{X\times Y} \varphi(y)h(x)\,\mathrm{d}\gamma(x,y) \\
&= \int_Y \left(\int_X \varphi(y)h(x)\,\mathrm{d}\gamma_y(x)\right) \mathrm{d}(\pi^2_\sharp\gamma)(y) \\
&= \int_X \varphi(y)\left(\int_X h(x)\,\mathrm{d}\gamma_y(x)\right)\mathrm{d}\nu(y),
\end{aligned}
$$

for all $\varphi \in C(Y)$, hence we have that $\nu$-a.e.

$$(Gh)(y) = \int_X h(x)\,\mathrm{d}\gamma_y(x), \tag{3.2}$$

where $(\gamma_y)_y$ is the disintegration of $\gamma \in \mathcal{P}(X\times Y)$ with respect to the second component. In the case where $\gamma = g(\mu\times\nu) \in \Gamma(\mu,\nu)$ for some function $g : X\times Y \to [0,1]$, then it is easy to see that Equation (3.2) reduces to

$$(Gh)(y) = \int_X h(x)g(x,y)\,\mathrm{d}\mu(x). \tag{3.3}$$

With this remark in mind we have the following proposition.

**Proposition 3.1.6.** *Let $\gamma_1 = g_1(\mu\times\nu)$ and $\gamma_2 = g_2(\mu\times\nu)$ in $\Gamma(\mu,\nu)$. Then*

$$\|G(\gamma_1) - G(\gamma_2)\|_{L^2(X,\mu)\to L^2(Y,\nu)} \le \|g_1 - g_2\|_{L^2(X\times Y,\mu\times\nu)} \tag{3.4}$$

*In particular, if $\gamma_n = g_n(\mu\times\nu)$ and $\gamma = g(\mu\times\nu)$ with $g_n \to g$ in the $L^2(X\times Y,\mu\times\nu)$ norm, then $G(\gamma_n) \to G(\gamma)$ in the $L^2(X,\mu) \to L^2(Y,\nu)$ operator norm.*

*Proof.* Let $h \in L^2(X,\mu)$. Then we have

$$\begin{aligned}
\|G(\gamma_1)h - G(\gamma_2)h\|^2_{L^2(Y,\nu)} &= \int_X \left[G(\gamma_1)h(y) - G(\gamma_2)h(y)\right]^2 \mathrm{d}\nu(y) \\
&= \int_X \left|\int_X (g_1(x,y) - g_2(x,y))h(x)\,\mathrm{d}\mu(x)\right|^2 \mathrm{d}\nu(y), \text{ by (3.3)} \\
&\leq \int_X \left[\int_X (g_1(x,y) - g_2(x,y))^2\,\mathrm{d}\mu(x) \cdot \int_X h(x)^2\,\mathrm{d}\mu(x)\right] \mathrm{d}\nu(y) \\
&= \int_X \int_X (g_1(x,y) - g_2(x,y))^2\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y) \cdot \int_X h(x)^2\,\mathrm{d}\mu(x) \\
&= \|g_1 - g_2\|^2_{L^2(X\times Y,\mu\times\nu)} \cdot \|h\|^2_{L^2(X,\mu)},
\end{aligned}$$

which implies Equation (3.4). □

Now we prove that $G$ can be viewed as a functor.

**Proposition 3.1.7** (Functoriality of G)**.** *Let $(X,\mu)$, $(Y,\nu)$ and $(Z,\kappa)$ be Borel probability spaces. Consider transport plans $\gamma^1 \in \Gamma(\mu,\nu)$, $\gamma^2 \in \Gamma(\nu,\kappa)$ and the composition plan $\gamma^2 \circ \gamma^1 \in \Gamma(\mu,\kappa)$, as defined in Proposition 2.1.4. Then $G(\gamma^2 \circ \gamma^1) = G(\gamma^2) \circ G(\gamma^1)$ as functions $L^p(X,\mu) \to L^p(Z,\kappa)$.*

*Proof.* Let $h \in L^p(X,\mu)$ and $\varphi \in C(Z)$. Then

$$\begin{aligned}
\int_Z \varphi(z)(G(\gamma^2 \circ \gamma^1)h)(z)\,\mathrm{d}\kappa(z) &= \int_{X\times Z} \varphi(z)h(x)\,\mathrm{d}(\gamma^2 \circ \gamma^1)(x,z) \\
&= \int_Y \left(\int_{X\times Z} \varphi(z)h(x)\,\mathrm{d}(\gamma^1_y \times \gamma^2_y)(x,z)\right)\mathrm{d}\nu(y) \\
&= \int_Y \int_Z \int_X \varphi(z)h(x)\,\mathrm{d}\gamma^1_y(x)\,\mathrm{d}\gamma^2_y(z)\,\mathrm{d}\nu(y)
\end{aligned}$$

and

$$\begin{aligned}
\int_Z \varphi(z)\left(G(\gamma^2) \circ G(\gamma^1)h\right)(z)\,\mathrm{d}\kappa(z) &= \int_{Y\times Z} \varphi(z)\left(G(\gamma_1)h\right)(y)\,\mathrm{d}\gamma^2(y,z) \\
&= \int_Y \left(\int_Z \varphi(z)\left(G(\gamma_1)h\right)(y)\,\mathrm{d}\gamma^2_y(z)\right)\mathrm{d}\nu(y) \\
&= \int_Y v(y)G(\gamma_1)h(y)\,\mathrm{d}\nu(y), \quad v(y) = \int_Z \varphi(z)\,\mathrm{d}\gamma^2_y(z) \\
&= \int_{X\times Y} v(y)h(x)\,\mathrm{d}\gamma^1(x,y) \\
&= \int_Y \left(\int_X v(y)h(x)\,\mathrm{d}\gamma^1_y(x)\right)\mathrm{d}\nu(y) \\
&= \int_Y \int_Z \int_X \varphi(z)h(x)\,\mathrm{d}\gamma^1_y(x)\,\mathrm{d}\gamma^2_y(z)\,\mathrm{d}\nu(y).
\end{aligned}$$

Hence

$$\int_Z \varphi(z)(G(\gamma^2 \circ \gamma^1)h)(z)\,\mathrm{d}\kappa(z) = \int_Z \varphi(z)\left(G(\gamma^2) \circ G(\gamma^1)h\right)(z)\,\mathrm{d}\kappa(z),$$

for all $h \in L^p(X,\mu)$ and $\varphi \in C(X)$, which proves what we wanted. □

**Remark 3.1.8.** Let $(X,d)$ be a metric space. We define the category $\mathcal{C} = \mathcal{C}(X)$ as follows:

- $\mathrm{Obj}(\mathcal{C}) = \mathcal{P}(X)$,
- $\mathcal{C}(\mu,\nu) = \Gamma(\mu,\nu)$.
- Composition: As defined in Proposition 2.1.4.

Then Proposition 3.1.7 shows that $G$ is a functor from the category $\mathcal{C}$ to the category of real vector spaces, with $G(\mu) = L^p(\mu)$ (for a fixed $p$) and $G(\gamma)$ as defined above.

## 3.2 The operator $G$ in the discrete case

In this section we are going to study the behaviour of the operator $G$ in the case where the measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $\gamma \in \Gamma(\mu,\nu)$ are discrete, i.e. convex combinations of Dirac measures. For this section suppose that $X = \{x_1,\dots,x_M\}$ and $Y = \{y_1,\dots,y_N\}$. Then

$$\gamma = \sum_{i=1}^{M}\sum_{j=1}^{N} \gamma_{ij}\delta_{(x_i,y_j)}, \quad \gamma_{i,j} \geq 0, \sum_{i=1}^{M}\sum_{j=1}^{N}\gamma_{ij} = 1,$$

with $x_1,\dots,x_M \in X$ and $y_1,\dots,y_N \in Y$. Moreover we have

$$\mu = \sum_{i=1}^{M} \mu_i \delta_{x_i}, \quad \mu_i = \sum_{j=1}^{N}\gamma_{ij}$$

and

$$\nu = \sum_{j=1}^{N} \nu_j \delta_{y_j}, \quad \nu_j = \sum_{i=1}^{M}\gamma_{ij}.$$

In this setting, Equation (3.2) becomes

$$(Gh)(y_j) = \sum_{i=1}^{M} h(x_i)(\gamma_{y_j})(x_i),$$

hence $Gh = G\cdot h$, where $h \in L^p(X,\mu)$ can be represented as $h = (h(x_1),\dots,h(x_M))^T \in \mathbb{R}^M$ and $G \in \mathbb{R}^{N\times M}$ is a matrix with $G_{ji} = (\gamma_{y_j})(x_i) = \frac{\gamma_{ij}}{\nu_j}$, by Remark 2.3.1. This means that the linear operator $G$ is just left multiplication with the matrix $G$.

If moreover we have that $\gamma = g(\mu\times\nu)$ (like in the optimal entropic plan), then we get

$$\begin{aligned} G_{ji} &= (\gamma_{y_j})(x_i) \\ &= [g(\mu\times\nu)]_{y_j}(x_i) \\ &= (g(-,y_j)\mu)(x_i), \text{ by Remark A.2.3} \\ &= g(x_i,y_j)\mu_i. \end{aligned}$$

Hence we have

$$(Gh)(y_j) = \sum_{i=1}^{M} h(x_i)g(x_i,y_j)\mu_i.$$

## 3.3 Comparison with the classical transfer operator

In the theory of dynamical systems, a very useful tool for study of these systems is the transfer operator, see for example [Klu+18], [DJ99] or [Sar12]. Here we will compare the definition found in the dynamical systems area with our previous construction.

**Definition 3.3.1** (Classical definition of transfer operator)**.** Let $X$ be a metric space, let $\mu \in \mathcal{P}(X)$ be a probability measure and let $F : X \to X$ be a continuous function. Then define the linear operator $T : L^p(\mu) \to L^p(F_\sharp\mu)$ by

$$Th = \frac{\mathrm{d}F_\sharp(h\mu)}{\mathrm{d}F_\sharp\mu}. \tag{3.5}$$

Using Radon-Nikodym's theorem and the property of pushforwards, it is easy to prove that the operator $T$ satisfies the following universal property

$$\int_X \varphi(y)(Th)(y)\,\mathrm{d}F_\sharp\mu(y) = \int_X \varphi(F(x))h(x)\,\mathrm{d}\mu(x), \tag{3.6}$$

for all continuous test functions $\varphi : X \to \mathbb{R}$. We can prove, as we did previously, that the map $T$ is well defined and linear.

Now we show that Definition 3.3.1 is a special case of the Definition 3.1.1. In order to do that we first need a few more definitions. We start with the definition of a Markov kernel. A Markov kernel can be considered as a stochastic generalization of a deterministic map.

**Definition 3.3.2** (Markov kernel, [Kle13, Definition 8.25])**.** Let $X$ and $Y$ be metric spaces. A *Markov kernel* from $X$ to $Y$ is a map $\kappa : \mathcal{B}(Y) \times X \to [0,1]$ with the following properties:

1. For every (fixed) $B \in \mathcal{B}(Y)$, the map $\kappa_B : X \to [0,1]$ defined by $x \mapsto \kappa(B,x)$ is Borel measurable.
2. For every (fixed) $x \in X$, the map $\kappa_x : \mathcal{B}(Y) \to [0,1]$ defined by $B \mapsto \kappa(B,x)$ is a probability measure on $Y$.

We will denote a Markov kernel $\kappa$ by its probability measures $(\kappa_x)_{x\in X}$.

**Example 3.3.3** (Deterministic case)**.** *Let $(\kappa_x)$ be a Markov kernel such that $\kappa_x$ is a Dirac measure for all $x \in X$. Then the Markov kernel $\kappa$ induces a mapping $F : X \to Y$.*

**Remark 3.3.4.** Intuitively, the quantity $\kappa_x(B)$ describes the probability of the image of $x \in X$ to land on the set $B \subseteq Y$. This means that the image of $x$ is "determined" by the measure $\kappa_x$.

**Definition 3.3.5** (Markov operator)**.** Let $X$ be a metric space and let $(\kappa_x)_{x\in X}$ be a Markov kernel from $X$ to $X$. Consider the map $K : \mathcal{P}(X) \to \mathcal{P}(X)$, defined by

$$\int_X \varphi(x)\,\mathrm{d}K\mu(x) = \int_X \left(\int_X \varphi(y)\,\mathrm{d}\kappa_x(y)\right)\mathrm{d}\mu(x), \tag{3.7}$$

for all test functions $\varphi \in C(X)$ and all probability measures $\mu \in \mathcal{P}(X)$. In particular, for a fixed $\mu \in \mathcal{P}(X)$ we have

$$(K\mu)(A) = \int_X \kappa_x(A)\,\mathrm{d}\mu(x). \tag{3.8}$$

We will refer to the map $K$, as the *Markov operator induced by the kernel* $(\kappa_x)$.

**Example 3.3.6.** *Let $\mu = \sum_j \mu_j \delta_{x_j}$ with $x_j \in X$, $\mu_j \geq 0$ and $\sum_j \mu_j = 1$. Then*

$$K\mu = \sum_j \mu_j \kappa_{x_j}.$$

*Indeed*

$$\begin{aligned} K\mu(A) &= \int_X \kappa_x(A)\,\mathrm{d}\mu(x) \\ &= \int_X \kappa_x(A)\,\mathrm{d}\left(\sum_j \mu_j \delta_{x_j}\right)(x) \\ &= \sum_j \mu_j \int_X \kappa_x(A)\,\mathrm{d}\delta_{x_j}(x) \\ &= \sum_j \mu_j \kappa_{x_j}. \end{aligned}$$

**Definition 3.3.7** (Markov plan)**.** Let $X$, $Y$ be metric spaces. Suppose we are given a Markov kernel $(\kappa_x)_{x\in X}$ from $X$ to $Y$. Then by disintegration, any probability measure $\mu \in \mathcal{P}(X)$ defines a unique probability measure $\rho = \rho(\mu, (k_x)) \in \mathcal{P}(X \times Y)$ that satisfies

$$\int_{X\times X} \varphi(x,y)\,\mathrm{d}\rho(x,y) = \int_X \left(\int_X \varphi(x,y)\,\mathrm{d}\kappa_x(y)\right)\mathrm{d}\mu(x), \tag{3.9}$$

for all test functions $\varphi \in C(X \times X)$. In particular,

$$\rho(A \times B) = \int_A \kappa_x(B)\,\mathrm{d}\mu(x), \tag{3.10}$$

for all $A \in \mathcal{B}(X)$ and $B \in \mathcal{B}(Y)$. We will refer to the measure $\rho$, as the *Markov plan induced by the kernel* $(\kappa_x)$ *and the measure* $\mu$.

**Remark 3.3.8.** The measure $\rho$ is indeed a plan, as defined in Definition 2.1.3. Note that the family $(k_x)$ is the (unique) disintegration of the measure $\rho$ with respect to the first marginal. Hence $\pi^1_\sharp \rho = \mu$. Moreover, we have that $\pi^2_\sharp \rho = K\mu$. Indeed,

$$\begin{aligned}\int_X \varphi(x)\,\mathrm{d}\pi^2_\sharp\rho(x) &= \int_{X\times X} \varphi(\pi^2(x,y))\,\mathrm{d}\rho(x,y)\\ &= \int_X \left(\int_X \varphi(y)\,\mathrm{d}\kappa_x(y)\right)\mathrm{d}\mu(x)\\ &= \int_X \varphi(x)\,\mathrm{d}K\mu(x), \text{ by Equation (3.7).}\end{aligned}$$

Thus $\rho \in \Gamma(\mu, K\mu)$.

Now we can compare the classical definition of the transfer operator with Definition 3.1.1.

**Proposition 3.3.9.** *Let $X$ be a metric space and let $F : X \to X$ be a continuous function. Fix a probability measure $\mu \in \mathcal{P}(X)$. Let $T : L^p(\mu) \to L^p(F_\sharp\mu)$, as defined in Definition 3.3.1. Let $(\kappa_x)_{x\in X}$ be a Markov kernel defined by $\kappa_x = \delta_{F(x)}$ and let $K$ and $\rho$ be the corresponding Morkov operator and plan. Then $K\mu = F_\sharp\mu$ and the operators $T : L^p(\mu) \to L^p(F_\sharp\mu)$ and $\widetilde{T} = G(\rho) : L^p(\mu) \to L^p(K\mu)$ coincide.*

*Proof.* Firstly, we show that $K\mu = F_\sharp\mu$. For any $A \in \mathcal{B}(X)$ we have $F_\sharp\mu(A) = \mu(F^{-1}(A))$. Now by Equation (3.8) we have

$$\begin{aligned}(K\mu)(A) &= \int_X \kappa_x(A)\,\mathrm{d}\mu(x)\\ &= \int_X \delta_{F(x)}(A)\,\mathrm{d}\mu(x)\\ &= \int_X \mathbf{1}_{\{F(x)\in A\}}\,\mathrm{d}\mu(x)\\ &= \int_X \mathbf{1}_{\{x\in F^{-1}(A)\}}\,\mathrm{d}\mu(x)\\ &= \mu(F^{-1}(A)),\end{aligned}$$

i.e. $K\mu(A) = F_\sharp\mu(A)$ for all $A \in \mathcal{B}(X)$. Now let $\varphi \in C(X)$ and $h \in L^p(\mu)$. We have

$$\begin{aligned}\int_X \varphi(y)(\widetilde{T}h)(y)\,\mathrm{d}F_\sharp\mu(y) &= \int_X \varphi(y)(\widetilde{T}h)(y)\,\mathrm{d}K\mu(y)\\ &= \int_{X\times X} \varphi(y)h(x)\,\mathrm{d}\rho(x,y), \text{ by Equation (3.1)}\end{aligned}$$

$$
\begin{aligned}
&= \int_X \left( \int_X \varphi(y) h(x) \, \mathrm{d}\delta_{F(x)}(y) \right) \mathrm{d}\mu(x), \text{ by Equation (3.9)} \\
&= \int_X h(x) \varphi(F(x)) \, \mathrm{d}\mu(x) \\
&= \int_X \varphi(y) (Th)(y) \, \mathrm{d}(F_\sharp \mu)(y), \text{ by Equation (3.6)},
\end{aligned}
$$

which proves that $\widetilde{T}h = Th$ almost everywhere, as wanted. □

# CHAPTER 4

# ENTROPIC TRANSFER OPERATORS

In this chapter we are going to present the theory of entropic transfer operators. We start with the deterministic case where it was first introduced in [JMS22]. Next we are going to try to generalize these results the stochastic case.

Throughout this chapter all metric spaces $X$ are assumed compact subspaces of $\mathbb{R}^d$ and we also fix the cost function $c : X \times X \to \mathbb{R}$ to be given by $c(x,y) = d^2(x,y) = \|x-y\|_2^2$. The results also hold for a general compact space $X$ but since in all of the applications we have $X \subseteq \mathbb{R}^d$ we will work with this assumption.

## 4.1 Deterministic case

All the constructions and results in this section are from [JMS22].

As mentioned in the Introduction, the main goal of the authors in [JMS22] is to solve the following problem: Given a measure-preserving dynamical system $(X,F,\mu)$ we want to find a discrete space $X^N \subseteq X$ and a stochastic (i.e. measure-valued) map $F^N : X^N \to X^N$ that "captures the most relevant features of $F$". They propose the following approach: The dynamics of the system $(X,F,\mu)$ can be captured by the transfer operator $T : L^2(\mu) \to L^2(\mu)$. The idea now is to define $F^N$ via a regularized transfer operator $T^{N,\varepsilon} : L^2(\mu^N) \to L^2(\mu^N)$, where $\varepsilon$ denotes the magnitude of the regularization and $(\mu^N)_N$ is a sequence of invariant probability measures that converges weakly to $\mu$. The stochasticity of the function $F^N$ is obtained by an optimal entropic transport plan.

A similar procedure can be applied to the transfer operator $T$, to get a regularized version $T^\varepsilon : L^2(\mu) \to L^2(\mu)$. Intuitively, $T^\varepsilon$ can be considered as a blurring of $T$ below the length scale $\varepsilon$. Finally, the authors have adopted the point of view that the "most relevant features" are given by the spectrum of $T^\varepsilon$.

Now we are going to formaly present these constructions and we are also going to state the basic theorems. Here we note that in the construction of the entropic transfer operators the fact that $\mu$ is $F$-invariant is not essential. Hence from now on we assume that $\mu$ is just a probability measure in $X$ (i.e. not necessarily $F$-invariant).

**Entropic regularization of the tranfer operator.** Recall the tranfer operator $T : L^p(\mu) \to L^p(F_\sharp\mu)$ as defined in Definition 3.3.1. Recall also that the operator $T$ satisfies the following property: For any $\varphi \in C(X)$ we have

$$\int_X \varphi(y)(Th)(y)\,\mathrm{d}F_\sharp\mu(y) = \int_X \varphi(F(x))h(x)\,\mathrm{d}\mu(x).$$

Let $\gamma^\varepsilon \in \Gamma(F_\sharp\mu, \mu)$ be the optimal entropic plan between $F_\sharp\mu$ and $\mu$, i.e. $\gamma^\varepsilon = g^\varepsilon(F_\sharp\mu \times \mu)$, with $g^\varepsilon(x,y) = \exp\left(\frac{-c(x,y)+a(x)+b(y)}{\varepsilon}\right)$, according to Proposition 2.2.3. By Remark 2.2.4 we have

$$\int_X g^\varepsilon(x,y)\,\mathrm{d}(F_\sharp\mu)(x) = 1 \Leftrightarrow \int_X g^\varepsilon(F(x),y)\,\mathrm{d}\mu(x) = 1 \text{ for } F_\sharp\mu\text{-almost all } y \in X$$

and

$$\int_X g^\varepsilon(x,y)\,\mathrm{d}\mu(y) = 1 \text{ for } \mu\text{-almost all } x \in X.$$

The optimal entropic plan $\gamma^\varepsilon$ induces the linear operator $G^\varepsilon = G(\gamma^\varepsilon) : L^p(F_\sharp\mu) \to L^p(\mu)$ which is given by

$$(G^\varepsilon h)(y) = \int_X h(x)g^\varepsilon(x,y)\,\mathrm{d}F_\sharp\mu(x),$$

by Remark 3.1.5 and it satisfies

$$\int_X \varphi(y)(G^\varepsilon h)(y)\,\mathrm{d}\mu(y) = \int_{X\times X} \varphi(y)h(x)g^\varepsilon(x,y)\,\mathrm{d}(F_\sharp\mu \times \mu)(x,y),$$

for all $h \in L^p(F_\sharp)$ and $\varphi \in C(X)$, by Remark 3.1.3.

Now define the operator $T^\varepsilon = G^\varepsilon \circ T : L^p(\mu) \to L^p(\mu)$. Then by [JMS22, Equation 21] we have

$$(T^\varepsilon h)(y) = \int_X g^\varepsilon(F(x),y)h(x)\,\mathrm{d}\mu(x).$$

Moreover, $g^\varepsilon(F(\cdot),\cdot)(\mu\times\mu) \in \Gamma(\mu,\mu)$. Now let $t^\varepsilon(x,y) = g^\varepsilon(F(x),y)$. Then

$$(T^\varepsilon h)(x,y) = \int_X h(x)t^\varepsilon(x,y)\,\mathrm{d}\mu(x)$$

and $t^\varepsilon(\mu\times\mu) \in \Gamma(\mu,\mu)$.

**Discrete approximation.** Let $(\mu^N)_N$ be a sequence of probability measures in $\mathcal{P}(X)$ such that $\mu^N \to \mu$ weakly. In the applications, we usually have $\mu^N = \sum_{k=1}^N m_k^N \delta_{x_k^N}$ with $m_k^N \geq 0$, $\sum_{k=1}^N m_k^N = 1$ and $x_1^N, \dots, x_N^N \in X$.

As described before we can construct the following linear maps: For a fixed $N \in \mathbb{N}$, let $T^N : L^p(\mu^N) \to L^p(F_\sharp\mu^N)$ induced by $F$ and $\mu^N$ and also let $G^{N,\varepsilon} = G(\gamma^{N,\varepsilon}) : L^p(F_\sharp\mu^N) \to L^p(\mu^N)$ induced by the optimal entropic plan $\gamma^{N,\varepsilon}$ between $F_\sharp\mu^N$ and $\mu^N$. Thus we can get the linear map $T^{N,\varepsilon} = G^{N,\varepsilon} \circ T^N : L^p(\mu^N) \to L^p(\mu^N)$ with

$$(T^{N,\varepsilon}h)(y) = \int_X h(x)t^{N,\varepsilon}(x,y)\,\mathrm{d}\mu^N(x),$$

where $t^{N,\varepsilon}(x,y) = g^{N,\varepsilon}(F(x),y)$.

In the case where $\mu^N = \sum_{k=1}^N m_k^N \delta_{x_k^N}$ we get

$$(T^{N,\varepsilon}h)(x_k^N) = \sum_{j=1}^N h(x_j^N) t^{N,\varepsilon}(x_j^N, x_k^N) m_j^N.$$

Note that in this setting $t^{N,\varepsilon}$ is equivalent with a $N \times N$ matrix $(t^{N,\varepsilon}(x_k^N, x_j^N))_{j,k=1}^N$ that satisfies the following relations

$$\sum_{j=1}^N m_j^N t^{N,\varepsilon}(x_k^N, x_j^N) m_j^N = 1, \quad \sum_{k=1}^N m_k^N t^{N,\varepsilon}(x_k^N, x_j^N) m_j^N = 1.$$

The matrix $(t^{N,\varepsilon}(x_k^N, x_j^N))_{j,k=1}^N$ can be efficiently computed by the Sinkhorn algorithm, cf. Section 2.3.1.

**From the discrete to continuous.** Our intuition tells us that we should have a convergence result of some type between $T^{N,\varepsilon}$ and $T^\varepsilon$ as $N \to \infty$. But $T^{N,\varepsilon}$ is an operator in $L^p(\mu^N) \to L^p(\mu^N)$ while $T^\varepsilon$ is an operator in $L^p(\mu) \to L^p(\mu)$. Hence we would like to extend the operator $T^{N,\varepsilon}$ to $L^p(\mu) \to L^p(\mu)$ in order to directly compare them.

Let $\gamma^N$ be the optimal transport plan between $\mu$ and $\mu^N$. This induces an operator $P^N = G(\gamma^N) : L^p(\mu) \to L^p(\mu^N)$. Also let $\tilde{\gamma}^N$ be the reverse (optimal) plan between $\mu^N$ and $\mu$, i.e. $\tilde{\gamma}^N = (\pi^2, \pi^1)_\sharp \gamma^N$. This also induces an operator $P^{N*} = G(\tilde{\gamma}^N) : L^p(\mu^N) \to L^p(\mu)$. Hence we can define the operator $\hat{T}^{N,\varepsilon} = P^{N*} \circ T^{N,\varepsilon} \circ P^N : L^p(\mu) \to L^p(\mu)$. Then according to [JMS22, Equation 26] we have:

$$(\hat{T}^{N,\varepsilon}h)(y) = \int_X h(x) \hat{t}^{N,\varepsilon}(x,y) \, \mathrm{d}\mu(x),$$

where

$$\hat{t}^{N,\varepsilon}(x,y) = \int_X \int_X t^{N,\varepsilon}(v,w) \, \mathrm{d}\gamma_x^N(v) \, \mathrm{d}\tilde{\gamma}_y^N(w),$$

with $(\gamma_x^N)$ being the disintegration of $\gamma^N$ with respect to the first marginal and $(\tilde{\gamma}_x^N)$ being the disintegration of $\tilde{\gamma}^N$ with respect to the second marginal.

**Convergence.** With the above extension we have the following convergence result.

**Theorem 4.1.1** ([JMS22, Proposition 1]). *Let $\mu^N \to \mu$ weakly. Then $\hat{t}^{N,\varepsilon} \to t^\varepsilon$ in the $L^2(\mu \times \mu)$-norm and $\hat{T}^{N,\varepsilon} \to T^\varepsilon$ in the $L^2(\mu) \to L^2(\mu)$ operator norm.*

**Proposition 4.1.2** ([JMS22, Proposition 2]). *The operator $T^\varepsilon : L^2(\mu) \to L^2(\mu)$ is compact.*

**Corollary 4.1.3** ([JMS22, Section 4.5], [DS88]). *Since $T^\varepsilon$ is a compact operator and $\hat{T}^{N,\varepsilon} \to T^\varepsilon$ in the operator norm, we have that the eigenvalues of $\hat{T}^{N,\varepsilon}$ converge to the eigenvalues of $T^\varepsilon$: Let $\hat{\lambda}_1^{N,\varepsilon}, \hat{\lambda}_2^{N,\varepsilon}, \ldots$ be the eigenvalues of $\hat{T}^{N,\varepsilon}$. Then there is an ordering of the eigenvalues of $T^\varepsilon$, $\lambda_1^\varepsilon, \lambda_2^\varepsilon, \ldots$ such that $\hat{\lambda}_k^{N,\varepsilon} \to \lambda_k^\varepsilon$ for all $k$. Similar result holds for the eigenfunctions.*

In [JMS22] the authors confirm the spectrum convergence by various numerical expiriments.

## 4.2 Stochastic Case

In this section we are going to try generalize the above results starting from a stochastic setting. Remember that previously we started from a deterministic $F : X \to X$.

**Setup** Let $X$ be a metric space. Consider a fixed Markov kernel $(k_x)_{x\in X}$ (cf. Definition 3.3.2) from $X$ to $X$ and a fixed probability measure $\mu \in \mathcal{P}(X)$. Let $K : \mathcal{P}(X) \to \mathcal{P}(X)$ be the corresponding Markov operator (cf. Definition 3.3.5) which satisfies

$$\int_X \varphi(x)\,\mathrm{d}K\mu(x) = \int_X \left(\int_X \varphi(y)\,\mathrm{d}\kappa_x(y)\right)\mathrm{d}\mu(x),$$

and let $\rho \in \Gamma(\mu, K\mu)$ be the corresponding Markov plan (cf. Definition 3.3.7) which satisfies

$$\int_{X\times X} \varphi(x,y)\,\mathrm{d}\rho(x,y) = \int_X \left(\int_X \varphi(x,y)\,\mathrm{d}\kappa_x(y)\right)\mathrm{d}\mu(x).$$

Finally, let $\nu := K\mu \in \mathcal{P}(X)$. Note that the Markov kernel $(\kappa_x)_x$ can be identified with $(\rho_x)_x$, the disintegration of $\rho$ with respect to the first marginal.

The Markov kernel $(\kappa_x)$ can be considered as a stochastic generalization of a deterministic map. Indeed, let $F : X \to X$ be a measurable function. Let $(\kappa_x^\delta)$ be the Markov kernel defined by $\kappa_x^\delta = \delta_{F(x)}$. Then there is an one to one correspondance between the function $F$ and the kernel $(\kappa_x^\delta)$. Moreover, in this case we have $K^\delta\mu = F_\sharp\mu$ by Proposition 3.3.9.

In this stochastic case we can again define the stochastic transfer operator $T = G(\rho) : L^p(\mu) \to L^p(K\mu)$ which is a generalization of the classical definition of the transfer operator, as we saw in Proposition 3.3.9. By Remark 3.1.5, for any $h \in L^p(\mu)$ we have

$$(Th)(y) = \int_X h(x)\,\mathrm{d}\rho_y(x), \tag{4.1}$$

where $(\rho_y)_y$ is the disintegration of $\rho$ with respect to the second marginal and by Remark 3.1.3 we have

$$\int_X \varphi(y)(Th)(y)\,\mathrm{d}\nu(y) = \int_{X\times X} \varphi(y)h(x)\,\mathrm{d}\rho(x,y), \tag{4.2}$$

for all $h \in L^p(X)$ and $\varphi \in C(X)$.

We follow the ideas for the constructions given in the previous section, i.e. we will entropically smooth the stochastic tranfer operator $T$ and from this we will try to prove a result similar to Theorem 4.1.1. However, this is not possible in this framework, so we will settle with a slightly weaker result. In order to overcome this shortcoming, we will introduced a modified smoothing construction in the next chapter.

### 4.2.1 Entropic regularization of stochastic tranfer operator

Let $\gamma^\varepsilon \in \Gamma(K\mu,\mu) = \Gamma(\nu,\mu)$ be the optimal entropic transport plan between $\nu$ and $\mu$, ie $\gamma^\varepsilon = g^\varepsilon(\nu\times\mu)$ with $g^\varepsilon(x,y) = \exp\left(\frac{-c(x,y)+a(x)+b(y)}{\varepsilon}\right)$, by Proposition 2.2.3. By the marginal conditions (Remark 2.2.4) we have

$$\int_X g^\varepsilon(x,y)\,\mathrm{d}\nu(x) = 1, \quad \int_X g^\varepsilon(x,y)\,\mathrm{d}\mu(y) = 1.$$

The optimal entropic plan $\gamma^\varepsilon$ induces the linear operator $G^\varepsilon = G(\gamma^\varepsilon) : L^p(\nu) \to L^p(\mu)$ with

$$(G^\varepsilon h)(y) = \int_X h(x)g^\varepsilon(x,y)\,\mathrm{d}\nu(x),$$

by Remark 3.1.5.

Now define $T^\varepsilon = G^\varepsilon \circ T = G(\gamma^\varepsilon \circ \rho) : L^p(\mu) \to L^p(\mu)$. Then we have

$$\begin{aligned}(T^\varepsilon h)(y) &= (G^\varepsilon T h)(y)\\ &= \int_X (Th)(x) g^\varepsilon(x,y)\,\mathrm{d}K\mu(x)\\ &= \int_{X\times X} g^\varepsilon(z,y)h(x)\,\mathrm{d}\rho(x,z), \text{ by Equation (4.1)}\\ &= \int_X \left(\int_X g^\varepsilon(z,y)h(x)\,\mathrm{d}\rho_x(z)\right)\mathrm{d}\mu(x)\\ &= \int_X h(x)t^\varepsilon(x,y)\,\mathrm{d}\mu(x),\end{aligned}$$

where

$$t^\varepsilon(x,y) = \int_X g^\varepsilon(z,y)\,\mathrm{d}\rho_x(z) = \int_X g^\varepsilon(z,y)\,\mathrm{d}\kappa_x(z). \tag{4.3}$$

**Remark 4.2.1** (Comparison with the deterministic case)**.** In the deterministic case, i.e. when $\kappa_x = \delta_{F(x)}$ for some $F : X \to X$ we have

$$\begin{aligned}t^\varepsilon(x,y) &= \int_X g^\varepsilon(z,y)\,\mathrm{d}\delta_{F(x)}(z)\\ &= g^\varepsilon(F(x),y).\end{aligned}$$

This means that all the constructions so far are indeed stochastic generalizations of the deterministic case.

**Proposition 4.2.2.** *Using the function $t^\varepsilon$ defined in Equation* (4.3)*, we have*

$$t^\varepsilon(\mu\times\mu) \in \Gamma(\mu,\mu).$$

*Hence*

$$T^\varepsilon = G(t^\varepsilon(\mu\times\mu)). \tag{4.4}$$

*Proof.* Observe that

$$\begin{aligned}\int_X t^\varepsilon(x,y)\,\mathrm{d}\mu(x) &= \int_X \left(\int_X g^\varepsilon(z,y)\,\mathrm{d}\kappa_x(z)\right)\mathrm{d}\mu(x)\\ &= \int_X g^\varepsilon(x,y)\,\mathrm{d}K\mu(x)\\ &= 1\end{aligned}$$

and

$$\begin{aligned}\int_X t^\varepsilon(x,y)\,\mathrm{d}\mu(y) &= \int_X \left(\int_X g^\varepsilon(z,y)\,\mathrm{d}\kappa_x(z)\right)\mathrm{d}\mu(y)\\ &= \int_{X\times X} g^\varepsilon(z,y)\,\mathrm{d}(\kappa_x\times\mu)(z,y)\\ &= \int_X \left(\int_X g^\varepsilon(z,y)\,\mathrm{d}\mu(y)\right)\mathrm{d}\kappa_x(z)\\ &= \int_X \mathrm{d}\kappa_x(z)\\ &= 1.\end{aligned}$$

□

### 4.2.2 Stochastic discrete approximation

Now suppose that we have an (discrete) approximation of the Markov plan $\rho$, i.e. let $(\rho^N)_N$ be a sequence of measures in $\mathcal{P}(X)$ such that $\rho^N \to \rho$ weakly. Let $\mu^N = \pi^1_\sharp \rho^N$ and $\nu^N = K\mu^N = \pi^2_\sharp \rho^N$. Note that $\mu^N \to \mu$ and $\nu^N \to \nu$ weakly.

As we did before we can construct the following linear maps:

- $T^N = G(\rho^N) : L^p(\mu^N) \to L^p(\nu^N)$
- $G^{N,\varepsilon} = G(\gamma^{N,\varepsilon}) : L^p(\nu^N) \to L^p(\mu^N)$, where $\gamma^{N,\varepsilon}$ is the optimal entropic plan between $\nu^N$ and $\mu^N$. For the plan $\gamma^{N,\varepsilon} \in \Gamma(\nu^N, \mu^N)$ we have:

$$\gamma^{N,\varepsilon} = g^{N,\varepsilon}(\nu^N \times \mu^N), \quad g^{N,\varepsilon}(x,y) = \exp\left(\frac{-c(x,y) + a^N(x) + b^N(y)}{\varepsilon}\right),$$

  and by the marginal conditions we have

$$\int_X g^{N,\varepsilon}(x,y)\,\mathrm{d}\nu^N(x) = 1, \quad \int_X g^{N,\varepsilon}(x,y)\,\mathrm{d}\mu^N(y) = 1.$$

Thus we get the linear map $T^{N,\varepsilon} = G^{N,\varepsilon} \circ T^N = G(\gamma^{N,\varepsilon} \circ \rho^N) : L^p(\mu^N) \to L^p(\mu^N)$ with

$$(T^{N,\varepsilon}h)(y) = \int_X h(x) t^{N,\varepsilon}(x,y)\,\mathrm{d}\mu^N(x), \tag{4.5}$$

where $t^{N,\varepsilon}(x,y) = \int_X g^{N,\varepsilon}(z,y)\,\mathrm{d}\rho^N_x(z)$. As before, we get

$$T^{N,\varepsilon} = G(t^{N,\varepsilon}(\mu^N \times \mu^N)).$$

### 4.2.3 From the discrete to continuous

As in the previous section, let $\gamma^N$, $\tilde{\gamma}^N$ be the optimal transport plans between $\mu$, $\mu^N$ and $\mu^N$, $\mu$. These plans induce the linear maps $P^N = G(\gamma^N) : L^p(\mu) \to L^p(\mu^N)$ and $\widetilde{P}^N = G(\tilde{\gamma}^N) : L^p(\mu^N) \to L^p(\mu)$.

Hence we can define the operator $\hat{T}^{N,\varepsilon} = \widetilde{P}^N \circ T^{N,\varepsilon} \circ P^N = G(\tilde{\gamma}^N \circ \gamma^{N,\varepsilon} \circ \rho^N \circ \gamma^N) : L^p(\mu) \to L^p(\mu)$. Then we have

$$\begin{aligned}
(\hat{T}^{N,\varepsilon}h)(y) &= \widetilde{P}^N(T^{N,\varepsilon}(P^N h))(y) \\
&= \int_X T^{N,\varepsilon}(P^N h)(w)\,\mathrm{d}\tilde{\gamma}^N_y(w), \text{ by Equation (3.2)} \\
&= \int_X \int_X (P^N h)(v) t^{N,\varepsilon}(v,w)\,\mathrm{d}\mu^N(v)\,\mathrm{d}\tilde{\gamma}^N_y(w), \text{ by Equation (4.5)} \\
&= \int_X \int_{X^2} h(x) t^{N,\varepsilon}(v,w)\,\mathrm{d}\gamma^N(x,v)\,\mathrm{d}\tilde{\gamma}^N_y(w), \text{ by Equation (3.1)} \\
&= \int_X \int_X \int_X h(x) t^{N,\varepsilon}(v,w)\,\mathrm{d}\gamma^N_x(v)\,\mathrm{d}\mu(x)\,\mathrm{d}\tilde{\gamma}^N_y(w), \text{ by disintegration} \\
&= \int_X h(x)\hat{t}^{N,\varepsilon}(x,y)\,\mathrm{d}\mu(x),
\end{aligned}$$

where

$$\hat{t}^{N,\varepsilon}(x,y) = \int_X \int_X t^{N,\varepsilon}(v,w)\,\mathrm{d}\gamma^N_x(v)\,\mathrm{d}\tilde{\gamma}^N_y(w).$$

Hence

$$\hat{T}^{N,\varepsilon} = G(\hat{t}^{N,\varepsilon}(\mu \times \mu)). \tag{4.6}$$

Moreover, we have

$$\begin{aligned}\int_X (\hat{T}^{N,\varepsilon}h)(y)\varphi(y)\,\mathrm{d}\mu(y) &= \int_X (P^{N*}(T^{N,\varepsilon}(P^N h)))(y)\varphi(y)\,\mathrm{d}\mu(y)\\ &= \int_{X^2} \varphi(y)T^{N,\varepsilon}(P^N h)(w)\,\mathrm{d}\tilde{\gamma}^N(w,y)\\ &= \int_{X^2}\left(\int_X (P^N h)(x)t^{N,\varepsilon}(x,w)\varphi(y)\,\mathrm{d}\mu^N(x)\right)\mathrm{d}\gamma^N(y,w)\\ &= \int_{X^2}\int_{X^2} h(x)\varphi(y)t^{N,\varepsilon}(v,w)\,\mathrm{d}\gamma^N(x,v)\,\mathrm{d}\gamma^N(y,w)\\ &= \int_{X^4} h(x)\varphi(y)t^{N,\varepsilon}(v,w)\,\mathrm{d}\gamma^N(x,v)\,\mathrm{d}\gamma^N(y,w).\end{aligned}$$

### 4.2.4 Convergence

Here we would like to directly generalize Theorem 4.1.1. Unfortunately, we were not able to prove such a result, so we have to settle with something slightly weaker.

**Theorem 4.2.3.** *Let $\varphi \in C(X)$. Then $\hat{T}^{N,\varepsilon}\varphi \to T^{\varepsilon}\varphi$ in the $L^2(\mu)$ norm as $N \to \infty$.*

In order to prove the above theorem we need some lemmas.

**Lemma 4.2.4.** *Let $\varphi \in C(X)$. Consider the measure $T\varphi\cdot\nu$, defined by*

$$\int_X f(x)\,\mathrm{d}T\varphi\cdot\nu(x) = \int_X (T\varphi)(x)f(x)\,\mathrm{d}\nu(x), \forall f \in C(X)$$

*and similarly define the measures $T^N\varphi\cdot\nu^N$ and $T^NP^N\varphi\cdot\nu^N$. Then we have that $T^N\varphi\cdot\nu^N \to T\varphi\cdot\nu$ and $T^NP^N\varphi\cdot\nu^N \to T\varphi\cdot\nu$ weakly. Moreover, note that $(T\varphi\cdot\nu)(X) \le \|\varphi\|_\infty$ and similarly for the other measures.*

*Proof.* Let $f \in C(X)$. Then

$$\begin{aligned}\int_X f\mathrm{d}T^N\varphi\cdot\nu^N &= \int_X (T^N\varphi)(y)f(y)\,\mathrm{d}\nu^N(y)\\ &= \int_{X^2} f(y)\varphi(x)\,\mathrm{d}\rho^N(x,y)\\ &\to \int_{X^2} f(y)\varphi(x)\,\mathrm{d}\rho(x,y)\\ &= \int_X (T\varphi)(y)f(y)\,\mathrm{d}\nu(y)\\ &= \int_X f\mathrm{d}T\varphi\cdot\nu,\end{aligned}$$

which shows that $T^N\varphi\cdot\nu^N \to T\varphi\cdot\nu$ weakly. In order to prove the second convergence it is enough to show that

$$\left|\int_X (T^N\varphi)(y)f(y)\,\mathrm{d}\nu^N(y) - \int_X (T^NP^N\varphi)(y)f(y)\,\mathrm{d}\nu^N(y)\right| \to 0,$$

because this means that the measures $T^N\varphi\cdot\nu^N$ and $T^NP^N\varphi\cdot\nu^N$ have the same weak limit. We have

$$\int_X (T^N\varphi)(y)f(y)\,\mathrm{d}\nu^N(y) = \int_{X^2} f(y)\varphi(x)\,\mathrm{d}\rho^N(x,y)$$

and

$$\int_X (T^NP^N\varphi)(y)f(y)\,\mathrm{d}\nu^N(y) = \int_{X^3} f(y)\varphi(w)\,\mathrm{d}\gamma_x^N(w)\,\mathrm{d}\rho^N(x,y).$$

Hence

$$\left|\int_X (T^N\varphi)(y)f(y)\,\mathrm{d}\nu^N(y) - \int_X (T^N P^N\varphi)(y)f(y)\,\mathrm{d}\nu^N(y)\right| \\ \leq \|f\|_\infty \int_{X^2} |\varphi(x)-\varphi(w)|\,\mathrm{d}\gamma^N(x,w).$$

Now it is enough to show that

$$\int_{X^2} |\varphi(x)-\varphi(w)|\,\mathrm{d}\gamma^N(x,w) \to 0.$$

Since $\varphi$ is uniformly continuous (continuous function in a compact space), it admits a modulus of continuity $\omega$, i.e. there is a continuous, increasing and concave function $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\omega(0) = 0$ such that

$$|\varphi(x)-\varphi(x')| \leq \omega(|x-x'|), \quad \forall x, x' \in X.$$

Hence we have

$$\int_{X^2} |\varphi(x)-\varphi(w)|\,\mathrm{d}\gamma^N(x,w) \leq \int_{X^2} \omega(|x-w|)\,\mathrm{d}\gamma^N(x,w) \\ \leq \omega\left(\int_{X^2} |x-w|\,\mathrm{d}\gamma^N(x,w)\right).$$

By Jensen inequality we have

$$\left(\int_{X^2} |x-w|\,\mathrm{d}\gamma^N(x,w)\right)^2 \leq \int_{X^2} |x-w|^2\,\mathrm{d}\gamma^N(x,w) =: W_2^2(\mu,\mu^N) \\ \Rightarrow \int_{X^2} |x-w|\,\mathrm{d}\gamma^N(x,w) \leq W_2(\mu,\mu^N),$$

so

$$\left|\int_X (T^N\varphi)(y)f(y)\,\mathrm{d}\nu^N(y) - \int_X (T^N P^N\varphi)(y)f(y)\,\mathrm{d}\nu^N(y)\right| \\ \leq \omega\left(\int_{X^2} |x-w|\,\mathrm{d}\gamma^N(x,w)\right) \\ \leq \omega(W_2(\mu,\mu^N)) \\ \to 0,$$

since $\mu^N \to \mu$ weakly, cf. Theorem 2.1.12. Finally, by the definition of the measure $T\varphi\cdot\nu$ and Equation (4.2) we have

$$(T\varphi\cdot\nu)(X) = \int_X \mathrm{d}T\varphi\cdot\nu(x) \\ = \int_X (T\varphi)(x)\,\mathrm{d}\nu(x) \\ = \int_{X\times X} \varphi(x)\,\mathrm{d}\rho(x,y) \\ \leq \|\varphi\|_\infty.$$

This finishes the proof. □

**Lemma 4.2.5.** *Let $\eta^N$ be a sequence in $\mathcal{M}(X)$ with $\eta^N \to \eta \in \mathcal{M}(X)$ weakly. Moreover suppose that $\eta^N(X) \leq M$ for some $M \geq 0$. Define the functions $f^N, f : X \to \mathbb{R}$ by*

$$f^N(y) = \int_X g^{N,\varepsilon}(x,y)\,\mathrm{d}\eta^N(x), \quad f(y) = \int_X g^\varepsilon(x,y)\,\mathrm{d}\eta(x).$$

*Then the family $(f^N)$ is uniformly equicontinuous and $f^N \to f$ uniformly.*

*Proof.* First we show that $(f^N)$ is uniformly equicontinuous. We already know that the family $(g^{N,\varepsilon})_N$ is uniformly equicontinuous, by Proposition 2.2.6, which means that there exists a continuous, increasing and concave function $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\omega(0) = 0$ such that

$$|g^{N,\varepsilon}(x,y) - g^{N,\varepsilon}(x',y')| \leq \omega\left(\sqrt{|x-x'|^2 + |y-y'|^2}\right),$$

for all $N$ (note that we have the same modulus of continuity $\omega$). Then we have

$$\begin{aligned} |f^N(y) - f^N(y')| &\leq \int_X |g^{N,\varepsilon}(x,y) - g^{N,\varepsilon}(x,y')| \,\mathrm{d}\eta^N(x) \\ &\leq M\omega(|y-y'|). \end{aligned}$$

Hence $(f^N)$ is uniformly equicontinuous. We will abuse the notation and we will use the same symbol $\omega$ for the common modulus of continuity of the family $(f^N)$. Now for the uniform convergence, we have for any $y \in X$

$$\begin{aligned} |f^N(y) - f(y)| &= \left|\int_X g^{N,\varepsilon}(x,y)\,\mathrm{d}\eta^N(x) - \int_X g^{\varepsilon}(x,y)\,\mathrm{d}\eta(x)\right| \\ &\leq \left|\int_X g^{N,\varepsilon}(x,y)\,\mathrm{d}\eta^N(x) - \int_X g^{\varepsilon}(x,y)\,\mathrm{d}\eta^N(x)\right| \\ &\quad + \left|\int_X g^{\varepsilon}(x,y)\,\mathrm{d}\eta^N(x) - \int_X g^{\varepsilon}(x,y)\,\mathrm{d}\eta(x)\right| \\ &\leq M\|g^{N,\varepsilon} - g^{\varepsilon}\|_\infty + \left|\int_X g^{\varepsilon}(x,y)\,\mathrm{d}\eta^N(x) - \int_X g^{\varepsilon}(x,y)\,\mathrm{d}\eta(x)\right|. \end{aligned}$$

Now let $\pi^N \in \Gamma(\eta, \eta^N)$ be the optimal transport plan between the measures $\eta$ and $\eta^N$. Then we have

$$\begin{aligned} \left|\int_X g^{\varepsilon}(x,y)\,\mathrm{d}\eta^N(x) - \int_X g^{\varepsilon}(x,y)\,\mathrm{d}\eta(x)\right| &= \left|\int_{X^2} g^{\varepsilon}(v,y)\,\mathrm{d}\pi^N(u,v) - \int_{X^2} g^{\varepsilon}(u,y)\,\mathrm{d}\pi^N(u,v)\right| \\ &\leq \int_{X^2} |g^{\varepsilon}(v,y) - g^{\varepsilon}(u,y)|\,\mathrm{d}\pi^N(u,v) \\ &\leq \int_{X^2} \omega(|u-v|)\,\mathrm{d}\pi^N(u,v) \\ &\leq \omega\left(\int_{X^2} |u-v|\,\mathrm{d}\pi^N(u,v)\right) \\ &\leq \omega\left(W_2(\eta,\eta^N)\right). \end{aligned}$$

Thus

$$|f^N(y) - f(y)| \leq M\|g^{N,\varepsilon} - g^{\varepsilon}\|_\infty + \omega\left(W_2(\eta,\eta^N)\right) \ \forall y \in X,$$

which means that

$$\begin{aligned} \|f^N - f\|_\infty &\leq M\|g^{N,\varepsilon} - g^{\varepsilon}\|_\infty + \omega\left(W_2(\eta,\eta^N)\right) \\ &\to 0, \end{aligned}$$

because $g^{N,\varepsilon} \to g^{\varepsilon}$ uniformly and $\eta^N \to \eta$ weakly. □

**Lemma 4.2.6.** *Let $f^N$ be a uniformly equicontinuous family in $C(X)$ and suppose that $f^N \to f$ uniformly. Then $P^{N*}f^N \to f$ in the $L^2(\mu)$ norm.*

*Proof.* We have

$$\|P^{N*}f^N - f\|_{L^2(\mu)} \leq \|f^N - f\|_{L^2(\mu)} + \|P^{N*}f^N - f^N\|_{L^2(\mu)}.$$

The first term goes to 0 because $f^N \to f$ uniformly. For the second term, since $(f^N)$ is uniformly equicontinuous there exists a continuous, increasing and concave function $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\omega(0) = 0$ such that

$$\left|f^N(y) - f^N(y')\right| \leq \omega\left(|y - y'|\right).$$

Since $X$ is compact, $\omega$ can be assumed bounded, and thus equivalently we may demand that (by considering the concave hull of $(\omega(\sqrt{\cdot}))^2$)

$$\left|f^N(y) - f^N(y')\right|^2 \leq \omega\left(|y - y'|^2\right).$$

Now we have

$$\begin{aligned}
\|P^{N*}f^N - f^N\|^2_{L^2(\mu)} &= \int_X \left|P^{N*}f^N(x) - f^N(x)\right|^2 \mathrm{d}\mu(x) \\
&= \int_X \left|\int_X (f^N(y) - f^N(x))\, \mathrm{d}\tilde{\gamma}^N_y(x)\right|^2 \mathrm{d}\mu(x), \text{ where } \tilde{\gamma}^N \in \Gamma(\mu^N, \mu) \\
&\leq \int_{X^2} \left|f^N(y) - f^N(x)\right|^2 \mathrm{d}\tilde{\gamma}^N_y(x)\, \mathrm{d}\mu(x), \text{ by Jensen} \\
&= \int_{X^2} \left|f^N(y) - f^N(x)\right|^2 \mathrm{d}\gamma^N(x,y) \\
&\leq \int_{X^2} \omega\left(|y - x|^2\right) \mathrm{d}\gamma^N(x,y) \\
&\leq \omega\left(\int_{X^2} |y - x|^2 \,\mathrm{d}\gamma^N(x,y)\right) \\
&= \omega\left(W_2^2(\mu, \mu^N)\right) \\
&\to 0,
\end{aligned}$$

as wanted. □

Now we are ready for the proof.

*Proof of Theorem 4.2.3.* Let $\eta^N = T^N P^N \varphi \cdot \nu^N$ and $\eta = T\varphi \cdot \nu$. By Lemma 4.2.4 we get that $\eta^N \to \eta$ weakly. Then by Lemma 4.2.5 for these $\eta^N$ and $\eta$, we get that the family $(G^{N,\varepsilon} T^N P^N \varphi)_N$ is uniformly equicontinuous and $G^{N,\varepsilon} T^N P^N \varphi \to G^\varepsilon T\varphi$ uniformly. Hence by Lemma 4.2.6 we get that $p^{N*} G^{N,\varepsilon} T^N P^N \varphi = \hat{T}^{N,\varepsilon}\varphi$ converges to $G^\varepsilon T\varphi = T^\varepsilon \varphi$ in the $L^2(\mu)$ norm, as wanted. □

Even though Theorem 4.2.3 has its own interest, we would like to have a convergence in the operator norm, in order to get a convergence in the spectrum as in Corollary 4.1.3. We propose a different approach that gives the desired results in the next chapter.

# CHAPTER 5

# DOUBLE ENTROPIC REGULARIZATION

In this chapter we are going to introduce a new construction of the entropic transfer operator, based on the previous work. The aim of the new construction is to get good convergence results in the operator norm even in the stochastic setting, which we failed to do in the previous chapter.

Throughout this chapter all metric spaces $X$ are assumed compact subspaces of $\mathbb{R}^d$ and we also fix the cost function $c : X \times X \to \mathbb{R}$ to be given by $c(x,y) = d^2(x,y) = \|x-y\|_2^2$. The results also hold for a general compact space $X$ but since in all of the applications we have $X \subseteq \mathbb{R}^d$ we will work with this assumption.

**Initial Setup.** Let $X$ be a metric space. Consider a fixed Markov kernel $(k_x)_{x\in X}$ (cf. Definition 3.3.2) from $X$ to $X$ and a fixed probability measure $\mu \in \mathcal{P}(X)$. Let $K : \mathcal{P}(X) \to \mathcal{P}(X)$ be the corresponding Markov operator (cf. Definition 3.3.5) which satisfies

$$\int_X \varphi(x)\,\mathrm{d}K\mu(x) = \int_X \left(\int_X \varphi(y)\,\mathrm{d}\kappa_x(y)\right)\mathrm{d}\mu(x),$$

and let $\rho \in \Gamma(\mu, K\mu)$ be the corresponding Markov plan (cf. Definition 3.3.7) which satisfies

$$\int_{X\times X} \varphi(x,y)\,\mathrm{d}\rho(x,y) = \int_X \left(\int_X \varphi(x,y)\,\mathrm{d}\kappa_x(y)\right)\mathrm{d}\mu(x).$$

Finally, let $\nu := K\mu \in \mathcal{P}(X)$.

In this setting we define the stochastic transfer operator $T = G(\rho) : L^p(\mu) \to L^p(K\mu)$, by

$$(Th)(y) = \int_X h(x)\,\mathrm{d}\rho_y(x), \tag{5.1}$$

for $h \in L^p(\mu)$.

The main objective of this part is to create new entropic tranfer operators $T^\varepsilon : L^p(\mu) \to L^p(\mu)$ and $T^{N,\varepsilon} : L^p(\mu^N) \to L^p(\mu^N)$ such that $T^{N,\varepsilon} \to T^\varepsilon$ in the operator norm. In the previous chapter, the idea was to compose the tranfer operator $T$ with one entropic transport plan. Now we will compose $T$ with *two* entropic transport plans.

## 5.1 Double entropic regularization of tranfer operator

For this construction we are going to work in a more general context, where we will have four probability measures and then we will restrict these results into our initial setup.

**New Setup.** Let $X$ be a metric space. Consider probability measures $\kappa, \lambda \in \mathcal{P}(X)$ and $\rho \in \mathcal{P}(X \times X)$. Let $\mu := \pi^1_\sharp \rho \in \mathcal{P}(X)$ and $\nu := \pi^2_\sharp \rho \in \mathcal{P}(X)$, so $\rho \in \Gamma(\mu, \nu)$.

Now we consider the following transport plans:

- Let $\gamma^\varepsilon \in \Gamma(\nu, \lambda)$ be the optimal entropic transport plan between $\nu$ and $\lambda$, ie $\gamma^\varepsilon = g^\varepsilon(\nu \times \lambda)$, with $g^\varepsilon(x,y) = \exp\left(\frac{-c(x,y)+a_\gamma(x)+b_\gamma(y)}{\varepsilon}\right)$.
- Let $\rho \in \Gamma(\mu, \nu)$.
- Let $\zeta^\varepsilon \in \Gamma(\kappa, \mu)$ be the optimal entropic transport plan between $\kappa$ and $\mu$, ie $\zeta^\varepsilon = f^\varepsilon(\kappa \times \mu)$, with $f^\varepsilon(x,y) = \exp\left(\frac{-c(x,y)+a_\zeta(x)+b_\zeta(y)}{\varepsilon}\right)$.

By the marginal conditions we have almost everywhere

$$\int_X g^\varepsilon(x,x')\,\mathrm{d}\nu(x) = 1, \quad \int_X g^\varepsilon(x,x')\,\mathrm{d}\lambda(x') = 1$$

and

$$\int_X f^\varepsilon(x,x')\,\mathrm{d}\kappa(x) = 1, \quad \int_X f^\varepsilon(x,x')\,\mathrm{d}\mu(x') = 1.$$

Now these transport plans induce the following linear operators:

- Let $G^\varepsilon = G(\gamma^\varepsilon) : L^p(\nu) \to L^p(\lambda)$ which satisfies
$$\int_X \varphi(y)(G^\varepsilon h)(y)\,\mathrm{d}\lambda(y) = \int_{X\times X} \varphi(y)h(x)g^\varepsilon(x,y)\,\mathrm{d}(\nu\times\lambda)(x,y),$$
for all $h \in L^p(\nu)$ and $\varphi \in C(X)$.
- Let $T = G(\rho) : L^p(\mu) \to L^p(\nu)$ which satisfies
$$\int_Y \varphi(y)(Th)(y)\,\mathrm{d}\nu(y) = \int_{X\times X} \varphi(y)h(x)\,\mathrm{d}\rho(x,y),$$
for all $h \in L^p(\mu)$ and $\varphi \in C(X)$.
- Let $F^\varepsilon = G(\varphi^\varepsilon) : L^p(\kappa) \to L^p(\mu)$ which satisfies
$$\int_X \varphi(y)(F^\varepsilon h)(y)\,\mathrm{d}\mu(y) = \int_{X\times X} \varphi(y)h(x)f^\varepsilon(x,y)\,\mathrm{d}(\kappa\times\mu)(x,y),$$
for all $h \in L^p(\kappa)$ and $\varphi \in C(X)$.

Now define the double smoothed entropic transfer operator $T^\varepsilon = G^\varepsilon \circ T \circ F^\varepsilon = G(\gamma^\varepsilon \circ \rho \circ \zeta^\varepsilon) : L^p(\kappa) \to L^p(\lambda)$. Then for $h \in L^p(\kappa)$, we have

$$\begin{aligned}(T^\varepsilon h)(y) &= \int_X (TF^\varepsilon h)(z) g^\varepsilon(z,y)\,\mathrm{d}\nu(z) \\ &= \int_{X\times X} (F^\varepsilon h)(w) g^\varepsilon(z,y)\,\mathrm{d}\rho(w,z)\end{aligned}$$

$$= \int_{X^3} h(x) f^\varepsilon(x,w) g^\varepsilon(z,y) \,\mathrm{d}\rho(w,z) \,\mathrm{d}\kappa(x)$$
$$= \int_X h(x) t^\varepsilon(x,y) \,\mathrm{d}\kappa(x),$$

where

$$t^\varepsilon(x,y) = \int_{X\times X} f^\varepsilon(x,w) g^\varepsilon(z,y) \,\mathrm{d}\rho(w,z).$$

**Proposition 5.1.1.** *Using the function $t^\varepsilon$ defined in Section 5.1, we have*

$$t^\varepsilon(\kappa\times\lambda) \in \Gamma(\kappa,\lambda).$$

*Hence*

$$T^\varepsilon = G(t^\varepsilon(\kappa\times\lambda)). \tag{5.2}$$

*Proof.* Observe that

$$\begin{aligned}
\int_X t^\varepsilon(x,y) \,\mathrm{d}\kappa(x) &= \int_X \left( \int_X \left( \int_X f^\varepsilon(x,w) g^\varepsilon(z,y) \,\mathrm{d}\rho_w(z) \right) \mathrm{d}\mu(w) \right) \mathrm{d}\kappa(x) \\
&= \int_X \left( \int_X g^\varepsilon(z,y) \left( \int_X f^\varepsilon(x,w) \,\mathrm{d}\kappa(x) \right) \mathrm{d}\rho_w(z) \right) \mathrm{d}\mu(w) \\
&= \int_X \left( \int_X g^\varepsilon(z,y) \,\mathrm{d}\rho_w(z) \right) \mathrm{d}\mu(w) \\
&= \int_{X\times X} g^\varepsilon(z,y) \,\mathrm{d}\rho(w,z) \\
&= \int_X \left( \int_X g^\varepsilon(z,y) \,\mathrm{d}\rho_z(w) \right) \mathrm{d}\nu(z) \\
&= \int_X g^\varepsilon(z,y) \,\mathrm{d}\nu(z) \\
&= 1
\end{aligned}$$

and

$$\begin{aligned}
\int_X t^\varepsilon(x,y) \,\mathrm{d}\lambda(y) &= \int_X \left( \int_X \left( \int_X f^\varepsilon(x,w) g^\varepsilon(z,y) \,\mathrm{d}\rho_w(z) \right) \mathrm{d}\mu(w) \right) \mathrm{d}\lambda(y) \\
&= \int_X \left( \int_X f^\varepsilon(x,w) \left( \int_X g^\varepsilon(z,y) \,\mathrm{d}\lambda(y) \right) \mathrm{d}\rho_w(z) \right) \mathrm{d}\mu(w) \\
&= \int_X \left( \int_X f^\varepsilon(x,w) \,\mathrm{d}\rho_w(z) \right) \mathrm{d}\mu(w) \\
&= \int_X f^\varepsilon(x,w) \,\mathrm{d}\mu(w) \\
&= 1.
\end{aligned}$$

□

## 5.2 Approximation

Now suppose that we have an approximation of the plan $\rho$ (in the applications, we are going to have discrete measures), and the measures $\kappa$ and $\lambda$, i.e. let $(\rho^N)_N$ be a sequence of measures in $\mathcal{P}(X\times X)$ such that $\rho^N \to \rho$ weakly and also let $(\kappa^N)$ and $(\lambda^N)$ be sequences of probability measures in $\mathcal{P}(X)$ such that $\kappa^N \to \kappa$ and $\lambda^N \to \lambda$ weakly. Let $\mu^N := \pi^1_\sharp \rho^N$ and $\nu^N := K\mu^N = \pi^2_\sharp \rho^N$. Note that also $\mu^N \to \mu$ and $\nu^N \to \nu$ weakly.

Similarly to the previous section, consider the following linear maps:

- $G^{N,\varepsilon} = G(\gamma^{N,\varepsilon}) : L^p(\nu^N) \to L^p(\lambda^N)$, where $\gamma^{N,\varepsilon}$ is the optimal entropic plan between $\nu^N$ and $\lambda^N$.

- $T^N = G(\rho^N) : L^p(\mu^N) \to L^p(\nu^N)$
- $F^{N,\varepsilon} = G(\zeta^{N,\varepsilon}) : L^p(\kappa^N) \to L^p(\mu^N)$, where $\zeta^{N,\varepsilon}$ is the optimal entropic plan between $\kappa^N$ and $\mu^N$.

Thus we get the linear map $T^{N,\varepsilon} = G^{N,\varepsilon} \circ T^N \circ F^{N,\varepsilon} = G(\gamma^{N,\varepsilon} \circ \rho^N \circ \zeta^{N,\varepsilon}) : L^p(\kappa^N) \to L^p(\lambda^N)$ with

$$(T^{N,\varepsilon}h)(y) = \int_X h(x)t^{N,\varepsilon}(x,y)\,\mathrm{d}\kappa^N(x), \tag{5.3}$$

where

$$t^{N,\varepsilon}(x,y) = \int_{X\times X} f^{N,\varepsilon}(x,w)g^{N,\varepsilon}(z,y)\,\mathrm{d}\rho^N(w,z).$$

## 5.3 Convergence

Let $\gamma_\kappa^N$ be the optimal transport plan between $\kappa$ and $\kappa^N$ and similarly, let $\gamma_\lambda^N$ be the optimal transport plan between $\lambda^N$ and $\lambda$. These plans induce the linear maps $K^N = G(\gamma_\kappa^N) : L^p(\kappa) \to L^p(\kappa^N)$ and $L^N = G(\gamma_\lambda^N) : L^p(\lambda^N) \to L^p(\lambda)$.

Hence we can define the extension operator $\hat{T}^{N,\varepsilon} = L^N \circ T^{N,\varepsilon} \circ K^N = G(\gamma_\lambda^N \circ \gamma^{N,\varepsilon} \circ \rho^N \circ \zeta^{N,\varepsilon} \circ \gamma_\kappa^N) : L^p(\kappa) \to L^p(\lambda)$. Then we have

$$\begin{aligned}
(\hat{T}^{N,\varepsilon}h)(y) &= L^N(T^{N,\varepsilon}(K^N h))(y) \\
&= \int_X T^{N,\varepsilon}(K^N h)(w)\,\mathrm{d}(\gamma_\lambda^N)_y(w) \\
&= \int_X \int_X (L^N h)(v)t^{N,\varepsilon}(v,w)\,\mathrm{d}\kappa^N(v)\,\mathrm{d}(\gamma_\lambda^N)_y(w) \\
&= \int_X \int_{X\times X} h(x)t^{N,\varepsilon}(v,w)\,\mathrm{d}\gamma_\kappa^N(x,v)\,\mathrm{d}(\gamma_\lambda^N)_y(w) \\
&= \int_X \int_X \int_X h(x)t^{N,\varepsilon}(v,w)\,\mathrm{d}(\gamma_\kappa^N)_x(v)\,\mathrm{d}\kappa(x)\,\mathrm{d}(\gamma_\lambda^N)_y(w) \\
&= \int_X h(x)\hat{t}^{N,\varepsilon}(x,y)\,\mathrm{d}\kappa(x),
\end{aligned}$$

where

$$\hat{t}^{N,\varepsilon}(x,y) = \int_X \int_X t^{N,\varepsilon}(v,w)\,\mathrm{d}(\gamma_\kappa^N)_x(v)\,\mathrm{d}(\gamma_\lambda^N)_y(w),$$

with $(\gamma_\kappa^N)_x$ being the disintegration of $\gamma_\kappa^N$ with respect to the first marginal and $(\gamma_\lambda^N)_y$ being the disintegration of $\gamma_\lambda^N$ with respect to the second marginal. Hence

$$\hat{T}^{N,\varepsilon} = G(\hat{t}^{N,\varepsilon}(\kappa \times \lambda)). \tag{5.4}$$

In this double smoothing setup we have convergence in the operator norm as shown in the following theorem.

**Theorem 5.3.1.** *Suppose that $\kappa^N \to \kappa$, $\lambda^N \to \lambda$ and $\rho^N \to \rho$ weakly. Then $\hat{t}^{N,\varepsilon} \to t^\varepsilon$ in the $L^2(\kappa \times \lambda)$ norm and $\hat{T}^{N,\varepsilon} \to T^\varepsilon$ in the $L^2(\kappa) \to L^2(\lambda)$ operator norm.*

*Proof.* By Equation (5.4), Equation (5.2) and Proposition 3.1.6 we have that

$$\left\|\hat{t}^{N,\varepsilon} - t^\varepsilon\right\|_{L^2(\kappa\times\lambda)} \to 0 \Rightarrow \left\|\hat{T}^{N,\varepsilon} - T^\varepsilon\right\|_{L^2(\kappa)\to L^2(\lambda)} \to 0.$$

So it is enough to show that $\left\|\hat{t}^{N,\varepsilon} - t^\varepsilon\right\|_{L^2(\kappa\times\lambda)} \to 0$. By the triangle inequality we get

$$\left\|\hat{t}^{N,\varepsilon} - t^\varepsilon\right\|_{L^2(\kappa\times\lambda)} \le \left\|\hat{t}^{N,\varepsilon} - t^{N,\varepsilon}\right\|_{L^2(\kappa\times\lambda)} + \left\|t^{N,\varepsilon} - t^\varepsilon\right\|_{L^2(\kappa\times\lambda)},$$

which means that it is enough to show that $\left\|\hat{t}^{N,\varepsilon} - t^{N,\varepsilon}\right\|_{L^2(\kappa\times\lambda)} \to 0$ and $\left\|t^{N,\varepsilon} - t^\varepsilon\right\|_{L^2(\kappa\times\lambda)} \to 0$ seperately.

**The first term.** By Proposition 2.2.6 we have that the family $(g^{N,\varepsilon})$ is equicontinuous, i.e. all $g^{N,\varepsilon}$ admit the same modulus of continuity $\omega_g$. Similarly, all $f^{N,\varepsilon}$ admit the same modulus of continuity $\omega_f$. Since each family has its own modulus of continuity it is easy to see that also the products $\left(f^{N,\varepsilon}g^{N,\varepsilon}\right)_N$ admit the same modulus of continuity, i.e. there is a continuous, increasing and concave function $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\omega(0) = 0$ such that

$$\left|f^{N,\varepsilon}(x,w)g^{N,\varepsilon}(z,y) - f^{N,\varepsilon}(x',w)g^{N,\varepsilon}(z,y')\right| \leq \omega\left(\sqrt{|x-x'|^2 + |y-y'|^2}\right),$$

for $N \in \mathbb{N}$. Since $\rho^N$ is a probability measure, we get

$$\left|t^{N,\varepsilon}(x,y) - t^{N,\varepsilon}(x',y')\right| \leq \omega\left(\sqrt{|x-x'|^2 + |y-y'|^2}\right),$$

for all $x, x', y, y' \in X$. Since $X$ is compact, $\omega$ can be assumed to be bounded, and thus equivalently we may demand that

$$\left|t^{N,\varepsilon}(x,y) - t^{N,\varepsilon}(x',y')\right|^2 \leq \widehat{\omega}\left(|x-x'|^2 + |y-y'|^2\right),$$

where $\widehat{\omega}$ is the concave hull of $[\omega(\sqrt{\cdot})]^2$ and still continuous at $\widehat{\omega}(0) = 0$. As a result, we have

$$\begin{aligned}
\left\|\hat{t}^{N,\varepsilon} - t^{N,\varepsilon}\right\|_2^2 &= \int_{X^2} (\hat{t}^{N,\varepsilon}(x,y) - t^{N,\varepsilon}(x,y))^2 \,\mathrm{d}\kappa(x)\,\mathrm{d}\lambda(y) \\
&= \int_{X^2} \left(\int_{X^2} t^{N,\varepsilon}(v,w) - t^{N,\varepsilon}(x,y)\,\mathrm{d}(\gamma_\kappa^N)_x(v)\,\mathrm{d}(\gamma_\lambda^N)_y(w)\right)^2 \mathrm{d}\kappa(x)\,\mathrm{d}\lambda(y) \\
&\leq \int_{X^4} \left(t^{N,\varepsilon}(v,w) - t^{N,\varepsilon}(x,y)\right)^2 \mathrm{d}(\gamma_\kappa^N)_x(v)\,\mathrm{d}(\gamma_\lambda^N)_y(w)\,\mathrm{d}\kappa(x)\,\mathrm{d}\lambda(y) \\
&= \int_{X^2} \left[\int_{X^2} \left(t^{N,\varepsilon}(v,w) - t^{N,\varepsilon}(x,y)\right)^2 \mathrm{d}(\gamma_\kappa^N)_x(v)\,\mathrm{d}\kappa(x)\right] \mathrm{d}(\gamma_\lambda^N)_y(w)\,\mathrm{d}\lambda(y) \\
&= \int_{X^4} \left(t^{N,\varepsilon}(v,w) - t^{N,\varepsilon}(x,y)\right)^2 \mathrm{d}\gamma_\kappa^N(x,v)\,\mathrm{d}\gamma_\lambda^N(y,w) \\
&\leq \int_{X^4} \widehat{\omega}(|x-v|^2 + |y-w|^2)\,\mathrm{d}\gamma_\kappa^N(x,v)\,\mathrm{d}\gamma_\lambda^N(y,w) \\
&\leq \widehat{\omega}\left(\int_{X^4} |x-v|^2 + |y-w|^2\,\mathrm{d}\gamma_\kappa^N(x,v)\,\mathrm{d}\gamma_\lambda^N(y,w)\right) \\
&= \widehat{\omega}\left(W_2^2(\kappa,\kappa^N) + W_2^2(\lambda,\lambda^N)\right) \\
&\to 0,
\end{aligned}$$

as $\kappa^N \to \kappa$ and $\lambda^N \to \lambda$ weakly.

**The second term.** First of all, by Proposition 2.2.6 we get that $g^{N,\varepsilon} \to g^\varepsilon$ and $f^{N,\varepsilon} \to f^\varepsilon$ uniformly. Now let

$$\tilde{t}^{N,\varepsilon} = \int_{X^2} f^\varepsilon(x,w)g^\varepsilon(z,y)\,\mathrm{d}\rho^N(w,z).$$

We have

$$\left\|t^{N,\varepsilon} - t^\varepsilon\right\|_2 \leq \left\|t^{N,\varepsilon} - \tilde{t}^{N,\varepsilon}\right\|_2 + \left\|\tilde{t}^{N,\varepsilon} - t^\varepsilon\right\|_2.$$

On one hand, we have

$$\begin{aligned}
\left|t^{N,\varepsilon}(x,y) - \tilde{t}^{N,\varepsilon}(x,y)\right| &\leq \int_{X^2} \left|f^{N,\varepsilon}(x,w)g^{N,\varepsilon}(z,y) - f^\varepsilon(x,w)g^\varepsilon(z,y)\right| \mathrm{d}\rho^N(w,z) \\
&\leq \int_{X^2} \left|f^{N,\varepsilon}(x,w)\left(g^{N,\varepsilon}(z,y) - g^\varepsilon(z,y)\right)\right| \mathrm{d}\rho^N(w,z)
\end{aligned}$$

$$+\int_{X^2}\left|\left(f^{N,\varepsilon}(x,w)-f^{\varepsilon}(x,w)\right)g^{\varepsilon}(z,y)\right|\mathrm{d}\rho^N(w,z)$$
$$\leq \left\|f^{N,\varepsilon}\right\|_\infty \left\|g^{N,\varepsilon}-g^{\varepsilon}\right\|_\infty + \left\|f^{N,\varepsilon}-f^{\varepsilon}\right\|_\infty \left\|g^{\varepsilon}\right\|_\infty,$$

hence

$$\left\|t^{N,\varepsilon}-\tilde{t}^{N,\varepsilon}\right\|_\infty \leq \left\|f^{N,\varepsilon}\right\|_\infty \left\|g^{N,\varepsilon}-g^{\varepsilon}\right\|_\infty + \left\|f^{N,\varepsilon}-f^{\varepsilon}\right\|_\infty \left\|g^{\varepsilon}\right\|_\infty \to 0,$$

since $\left\|f^{N,\varepsilon}\right\|_\infty \to \left\|f^{\varepsilon}\right\|_\infty < \infty$ ($X$ is compact). Thus $t^{N,\varepsilon}-\tilde{t}^{N,\varepsilon}\to 0$ uniformly and so $\left\|t^{N,\varepsilon}-\tilde{t}^{N,\varepsilon}\right\|_2 \to 0$.

On the other hand, we have

$$\begin{aligned}\left|\tilde{t}^{N,\varepsilon}(x,y)-t^{\varepsilon}(x,y)\right| &\leq \int_{X^2}\left|f^{\varepsilon}(x,w)g^{\varepsilon}(z,y)\right|\mathrm{d}\left|\rho^N-\rho\right|(w,z)\\ &\leq \left\|f^{\varepsilon}\right\|_\infty \left\|g^{\varepsilon}\right\|_\infty \left\|\rho^N-\rho\right\|_{TV}\\ &\leq 2\left\|f^{\varepsilon}\right\|_\infty \left\|g^{\varepsilon}\right\|_\infty\end{aligned}$$

and $\tilde{t}^{N,\varepsilon}\to t^{\varepsilon}$ pointwise (since $\rho^N\to\rho$ weakly). Hence by Lebesgue's dominated convergence theorem for $L^p$ spaces we get that $\left\|\tilde{t}^{N,\varepsilon}-t^{\varepsilon}\right\|_2\to 0$, as wanted. This finishes the proof. □

## 5.4 Discretization

As we mentioned above, in the applications we only work with discrete approximations of measures. Now suppose that the measure $\rho^N$ has the following form

$$\rho^N=\sum_{i=1}^N\sum_{j=1}^N \rho_{ij}^N\delta_{(x_i^N,x_j^N)},\quad \rho_{ij}^N\geq 0,\ \sum_{i,j=1}^N\rho_{ij}^N=1$$

with $x_1^N,\dots,x_N^N\in X$. Since $\mu^N=\pi^1_\sharp\rho^N$ and $\nu^N=\pi^2_\sharp\rho^N$, we have

$$\mu^N=\sum_{i=1}^N\mu_i^N\delta_{x_i^N},\quad \mu_i^N=\sum_{j=1}^N\rho_{ij}^N, \tag{5.5}$$

and

$$\nu^N=\sum_{j=1}^N\nu_j^N\delta_{x_j^N},\quad \nu_j^N=\sum_{i=1}^N\rho_{ij}^N. \tag{5.6}$$

Hence, by Section 3.2, we have that

- The operator $G^{N,\varepsilon}$ is left multiplication with the matrix $G^N$, where $G^N_{ji}=\frac{\gamma_{ij}^{N,\varepsilon}}{\lambda_j}=g^{N,\varepsilon}(x_i,x_j)\nu_i$.
- The operator $T^N$ is left multiplication with the matrix $P^N$, where $P^N_{ji}=\frac{\rho_{ij}^N}{\nu_j}$. (If $\nu_j=0$, then we set $P^N_{ji}=0$)
- The operator $F^{N,\varepsilon}$ is left multiplication with the matrix $F^N$, where $F^N_{ji}=\frac{\zeta_{ij}^{N,\varepsilon}}{\mu_j}=f^{N,\varepsilon}(x_i,x_j)\kappa_i$.

Now, since $T^{N,\varepsilon}=G^{N,\varepsilon}\circ T^N\circ F^{N,\varepsilon}$, we get that the operator $T^{N,\varepsilon}$ is left multiplication with the matrix $G^NP^NF^N$.

## 5.5 Results about dynamical systems

So far, we have worked with general measures $\rho$, $\kappa$, $\lambda$, $\mu$ and $\nu$. Now we will restrict our attention in the initial setup introduced in the beginning of the chapter. In this case, the measure/plan $\rho$ is the Markov plan to a fixed Markov kernel $(\kappa_x)_x$ from $X$ to $X$, $\mu$ is a fixed probability measure in $X$ and $\nu = K\mu$. Now setting $\kappa = \lambda = \mu$ in the previous work we get a double entropic tranfer operator $T^\varepsilon : L^p(\mu) \to L^p(\mu)$ with

$$(T^\varepsilon h)(y) = \int_X h(x) t^\varepsilon(x,y)\, \mathrm{d}\mu(x), \tag{5.7}$$

where

$$t^\varepsilon(x,y) = \int_{X\times X} f^\varepsilon(x,w) g^\varepsilon(z,y)\, \mathrm{d}\rho(w,z). \tag{5.8}$$

Here $f^\varepsilon$ is the density of the optimal entropic plan between $\mu$ and $\mu$, and $g^\varepsilon$ is the density of the optimal entropic plan between $\nu$ and $\mu$.

Now let $\rho^N \to \rho$ weakly which gives $\mu^N \to \mu$ and $\nu^N \to \nu$ weakly, where $\mu^N = \pi^1_\sharp \rho^N$ and $\nu^N = \pi^2_\sharp \rho^N$. Then we get another linear operator $T^{N,\varepsilon} : L^p(\mu^N) \to L^p(\mu^N)$ with

$$(T^{N,\varepsilon} h)(y) = \int_X h(x) t^{N,\varepsilon}(x,y)\, \mathrm{d}\mu^N(x),$$

where

$$t^{N,\varepsilon}(x,y) = \int_{X\times X} f^{N,\varepsilon}(x,w) g^{N,\varepsilon}(z,y)\, \mathrm{d}\rho^N(w,z).$$

Here $f^{N,\varepsilon}$ is the density of the optimal entropic plan between $\mu^N$ and $\mu^N$, and $g^{N,\varepsilon}$ is the density of the optimal entropic plan between $\nu^N$ and $\mu^N$.

As we did in Chapter 4 we extend the operator $T^{N,\varepsilon} : L^p(\mu^N) \to L^p(\mu^N)$ to an operator $\hat{T}^{N,\varepsilon} : L^p(\mu) \to L^p(\mu)$ by composing with the operators $P^N : L^p(\mu) \to L^p(\mu^N)$ (induced by $\gamma^N$ the optimal plan between $\mu$ and $\mu^N$) and $P^{N*} : L^p(\mu^N) \to L^p(\mu)$ left and right. Then

$$(\hat{T}^{N,\varepsilon} h)(y) = \int_X h(x) \hat{t}^{N,\varepsilon}(x,y)\, \mathrm{d}\mu(x),$$

where

$$\hat{t}^{N,\varepsilon}(x,y) = \int_X \int_X t^{N,\varepsilon}(v,w)\, \mathrm{d}\gamma^N_x(v)\, \mathrm{d}\gamma^N_y(w).$$

Now form Theorem 5.3.1, we get the corresponding results for the stochastic setting with the double smoothing.

**Theorem 5.5.1.** *Suppose that $\rho^N \to \rho$ weakly. Then $\hat{t}^{N,\varepsilon} \to t^\varepsilon$ in the $L^2(\mu\times\mu)$ norm and $\hat{T}^{N,\varepsilon} \to T^\varepsilon$ in the $L^2(\mu) \to L^2(\mu)$ operator norm.*

Since we have convergence in the operator norm we can follow the same steps as we saw in Section 4.1. First we show that the double smoothed operator $\mathrm{T}^\varepsilon$ is compact.

**Proposition 5.5.2.** *The operator $T^\varepsilon : L^2(\mu) \to L^2(\mu)$ is compact.*

*Proof.* By Equation (5.7), it is enough we prove that

$$\int_{X\times X} (t^\varepsilon(x,y))^2\, \mathrm{d}(\mu\times\mu)(x,y) < \infty. \tag{5.9}$$

This would prove that the operator $T^\varepsilon$ is Hilbert-Schmidt integral operator which immediately shows that $T^\varepsilon$ is compact. Equation (5.9) holds since $f^\varepsilon$ and $g^\varepsilon$ are bounded functions (continuous functions in a compact space) so $t^\varepsilon$ is bounded (for fixed $\varepsilon$). □

**Corollary 5.5.3** ([DS88]). *Since $T^\varepsilon$ is a compact operator and $\hat{T}^{N,\varepsilon} \to T^\varepsilon$ in the operator norm, we have that the eigenvalues of $\hat{T}^{N,\varepsilon}$ converge to the eigenvalues of $T^\varepsilon$: Let $\hat{\lambda}_1^{N,\varepsilon}, \hat{\lambda}_2^{N,\varepsilon}, \dots$ be the eigenvalues of $\hat{T}^{N,\varepsilon}$. Then there is an ordering of the eigenvalues of $T^\varepsilon$, $\lambda_1^\varepsilon, \lambda_2^\varepsilon, \dots$ such that $\hat{\lambda}_k^{N,\varepsilon} \to \lambda_k^\varepsilon$ for all $k$. Similar result holds for the eigenfunctions.*

## 5.6 Numerical comparison with single smoothing

In this section we waNT to compare the spectrum of the double smoothed transfer operator against the single smoothed tranfer operator using numerical experiments. We are going to work with the circle example from [JMS22, Section 6.1]. We will use the symbol $T_1^\varepsilon$ for the single smoothed entropic tranfer operator and the symbol $T_2^\varepsilon$ for the double smoothed entropic tranfer operator. We use similar symbols for the discrete approximations. We will present two types of experiments. The first one is the comparison of the two operators in the deterministic case and the second one is the comparison in the stochastic case.

### 5.6.1 Deterministic Setting

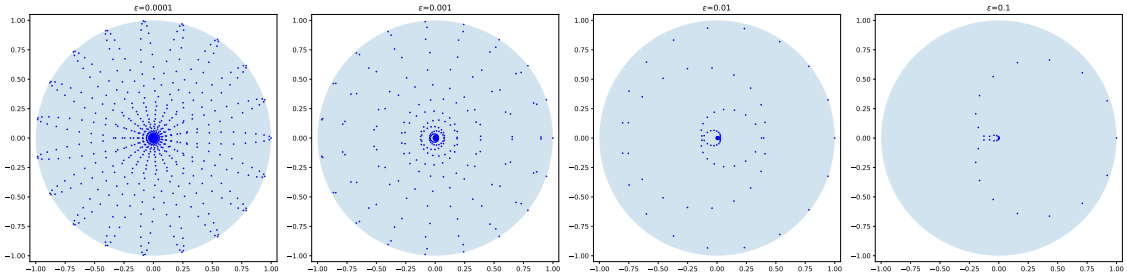
Suppose that $X = S^1 \cong \mathbb{R}/\mathbb{Z} \cong [0,1]/\{0,1\}$ and $F : X \to X$ given by the shift map $F(x) = x + \theta \pmod 1$ for some angle $\theta \in [0,1)$. Let $(\kappa_x)_x$ be the (deterministic) Markov kernel induced by $F$ and let $\mu$ be the uniform probability measure on $S^1$. For the discrete approximation we randomly choose $N = 1000$ points on $S^1$ and we set $\mu^N$ as the uniform probability measure over the $N$ chosen points. At first we set $\theta = \frac{1}{3}$. In this case we expect the spectra of both $T_1^{N,\varepsilon}$ and $T_2^{N,\varepsilon}$ to exhibit the 3rd roots of unity. The results are shown in Figure 5.1. For an irrational $\theta = \frac{1}{\pi} \approx \frac{1}{3}$ the results are shown in Figure 5.2.

Numerically we can see that the qualitive characteristic of both the single and double smoothing are very similar. In the double smoothing we observe that the extra regularization just adds more blurring to the spectrum of the single smoothing.

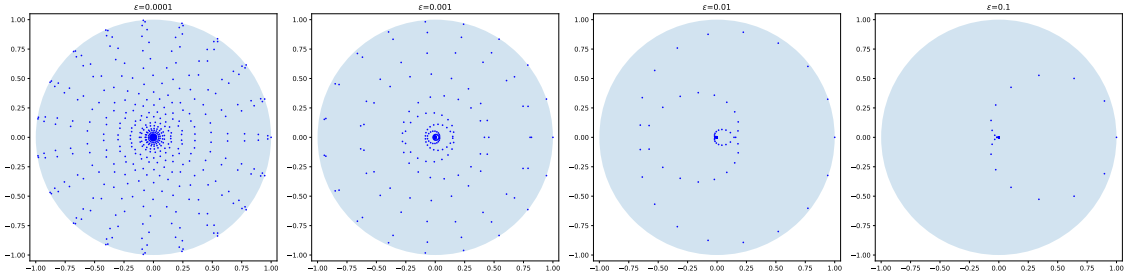
### 5.6.2 Stochastic Setting

We will work again with the unit circle $X = S^1$ and $N = 1000$. In this case we want to create a stochastic "mapping" $F : X \to X$. As mentioned, this can be done via a Markov kernel $\kappa = (\kappa_x)_x$. Essentially we send each $x \in X$ to a point in $X$ according to the measure $\kappa_x \in \mathcal{P}(X)$. In our case the mapping will be the shift map (as before) plus a small random noise. More specifically, we take the angle $\varphi_x$ of a point $x \in X = S^1$ and we send it to an angle $\varphi_x'$ according to $N(\varphi_x + \theta, \sigma^2)$, i.e. we rotate by angle $\theta$ and then we choose a point a bit to the left or a bit to the right according to some normal distribution. Again we choose $\theta = 1/3$ for comparison with the previous case and we let the standard deviation $\sigma$ of the noise and the regularization parameter $\varepsilon$ vary. The results are shown in Figures 5.3 to 5.5.

First of all, we observe that for small values of $\sigma$ the spectra of the stochastic operators $T_1^{N,\varepsilon}$ and $T_2^{N,\varepsilon}$ are similar to their deterministic counterparts. This can be explained by the fact that for small $\sigma$ we have $\varphi_x' \approx \varphi_x + \theta$ so the dataset is close to the deterministic case. On the other hand, what is more surprising is the fact that the spectrum of the single smoothed operator seems to be similar to the spectrum of the double smoothed operator regardless of $\sigma$ and $\varepsilon$. Of course there is more blurring in the double smoothing since we regularized twice.

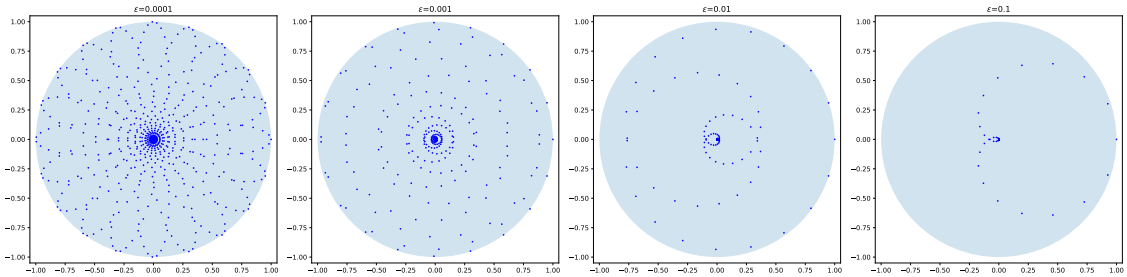



(a) Spectrum of the operator $T_1^{N,\varepsilon}$ (single smoothing, deterministic case)

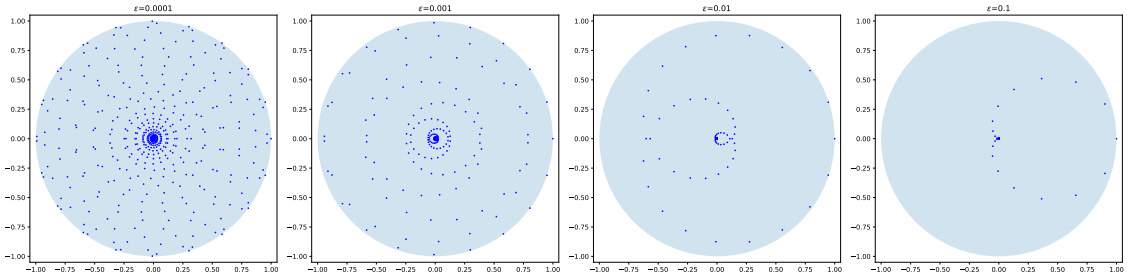



(b) Spectrum of the operator $T_2^{N,\varepsilon}$ (double smoothing, deterministic case)

Figure 5.1: Deterministic circle shift with $\theta = \frac{1}{3}$ using $N = 1000$ points: spectra of $T_1^{N,\varepsilon}$ and $T_2^{N,\varepsilon}$ for $\varepsilon$ from $10^{-4}$ to $10^{-1}$ (left to right, logarithmically).

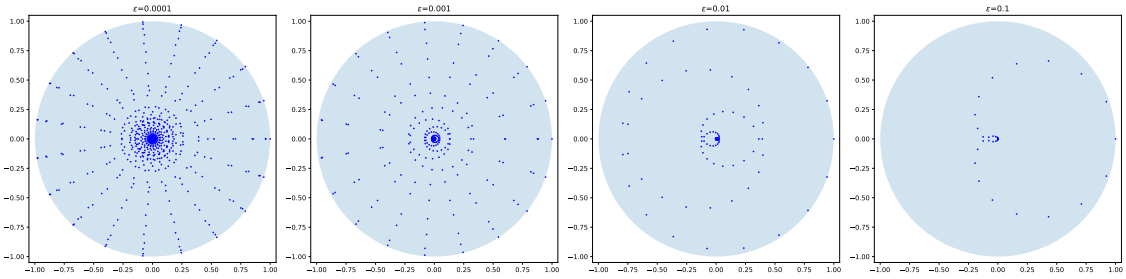



(a) Spectrum of the operator $T_1^{N,\varepsilon}$ (single smoothing, deterministic case)

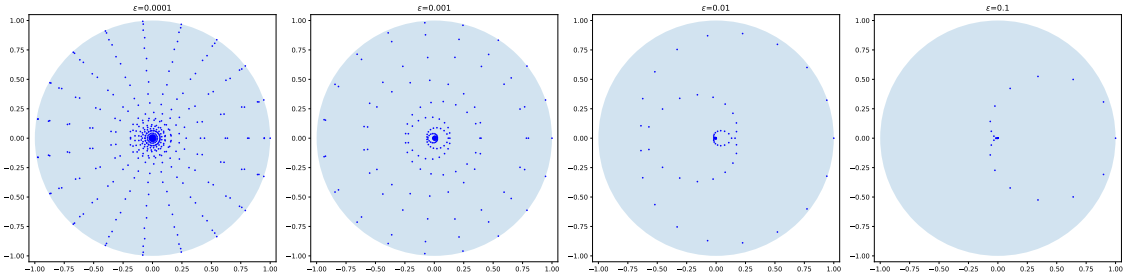



(b) Spectrum of the operator $T_2^{N,\varepsilon}$ (double smoothing, deterministic case)

Figure 5.2: Deterministic circle shift with $\theta = \frac{1}{\pi}$ using $N = 1000$ points: spectra of $T_1^{N,\varepsilon}$ and $T_2^{N,\varepsilon}$ for $\varepsilon$ from $10^{-4}$ to $10^{-1}$ (left to right, logarithmically).



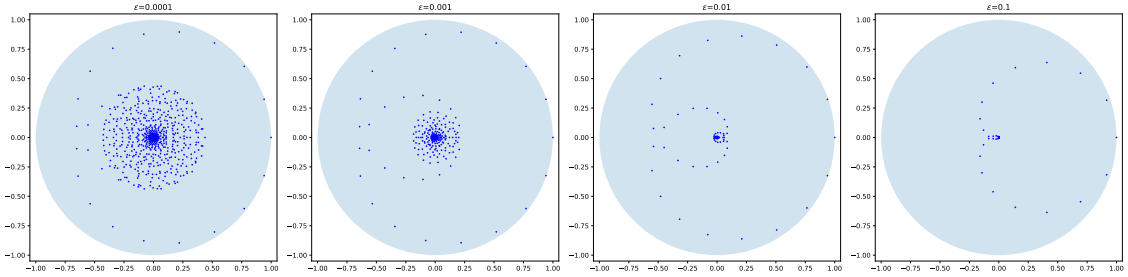

(a) Spectrum of the operator $T_1^{N,\varepsilon}$ with $\sigma = 0.01$ (single smoothing, stochastic case)



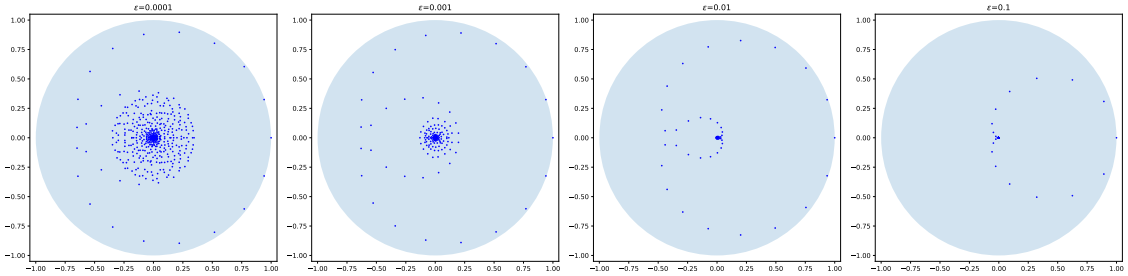

(b) Spectrum of the operator $T_2^{N,\varepsilon}$ with $\sigma = 0.01$ (double smoothing, stochastic case)

Figure 5.3: Stochastic circle shift with $\theta = \frac{1}{3}$ using $N = 1000$ points generated with $\sigma = 0.01$: spectra of $T_1^{N,\varepsilon}$ and $T_2^{N,\varepsilon}$ for $\varepsilon$ from $10^{-4}$ to $10^{-1}$ (left to right, logarithmically).

These plots show that even in the stochastic case the single smoothed operator seems to work well, i.e. we can see that, at least numerically, the spectrum of $T_1^{N,\varepsilon}$ seems to converge to the spectrum of $T_1^{\varepsilon}$ (where for small $\sigma$ it exhibits the third roots of unity as mentioned earlier). Since we could not prove convergence in the operator norm for the single smoothing in the stochastic setting, we did not expect to have convergence of the spectra in this case. We can think two possible explanations about this phenomenon. The first one is that perhaps we do have convergence of the single smoothed entropic operators in the operator norm, and hence we get convergence of the spectra, but we cannot prove this results with the ideas that we explored. The second one is that since we have convergence of the single smoothed entropic operators $T^{N,\varepsilon}\varphi \to T^{\varepsilon}\varphi$ in the $L^2$-norm for all $\varphi \in C(X)$, see Theorem 4.2.3, and since the eigenfunctions of $T_1^{N,\varepsilon}$ are "relatively" regular for large eigenvalues, maybe this is enough to get convergence of the eigenvalues. This is an interesting observation where further research is necessary in order to fully understand and explain this phenomenon.

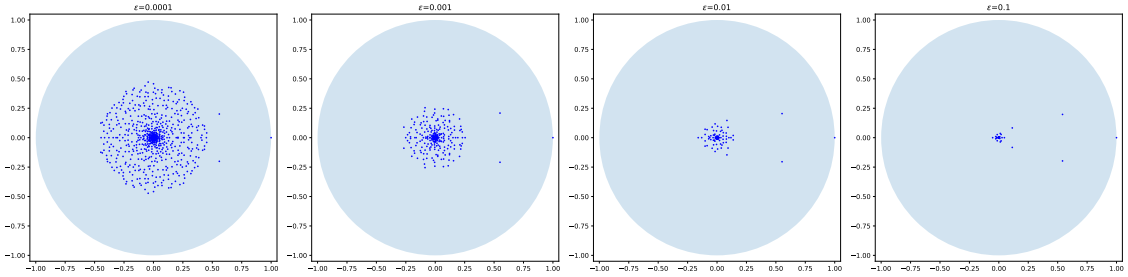



(a) Spectrum of the operator $T_1^{N,\varepsilon}$ with $\sigma = 0.1$ (single smoothing, stochastic case)

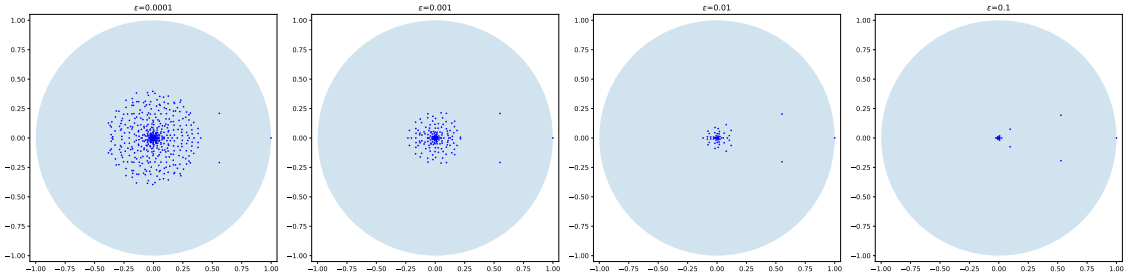



(b) Spectrum of the operator $T_2^{N,\varepsilon}$ with $\sigma = 0.1$ (double smoothing, stochastic case)

Figure 5.4: Stochastic circle shift with $\theta = \frac{1}{3}$ using $N = 1000$ points generated with $\sigma = 0.1$: spectra of $T_1^{N,\varepsilon}$ and $T_2^{N,\varepsilon}$ for $\varepsilon$ from $10^{-4}$ to $10^{-1}$ (left to right, logarithmically).




(a) Spectrum of the operator $T_1^{N,\varepsilon}$ with $\sigma = 1$ (single smoothing, stochastic case)




(b) Spectrum of the operator $T_2^{N,\varepsilon}$ with $\sigma = 1$ (double smoothing, stochastic case)

Figure 5.5: Stochastic circle shift with $\theta = \frac{1}{3}$ using $N = 1000$ points generated with $\sigma = 1$: spectra of $T_1^{N,\varepsilon}$ and $T_2^{N,\varepsilon}$ for $\varepsilon$ from $10^{-4}$ to $10^{-1}$ (left to right, logarithmically).

# CHAPTER 6

# CONVERGENCE RATES

In this chapter we are going to study the convergence rates of the original entropic transfer operator, i.e. the one with the single smoothing in the deterministic case. We start with some recent results about the sample complexity and convergence rates of the regularized and unregularized optimal transport. Next we try to establish similar rates for $t^{N,\varepsilon}$, the kernel of the entropic tranfer operator for an approximation measure $\mu^N$. Finally, we present some numerical experiments regarding the convergence of the spectrum of the operator $T^{N,\varepsilon}$ to the spectrum of the operator $T^{\varepsilon}$.

## 6.1 Sample complexity of Optimal Transport

Throughout this section all metric spaces are assumed to be bounded subsets of $\mathbb{R}^d$ unless specified otherwise. Moreover we assume that the cost function $c \in C^{s+1}(X)$ for some $s > \frac{d}{2}$.

In this section we are going to review the basic results about the sample complexity of the regularized and unregularized optimal transport. The first result in this area came in 1969 by Dudley in [Dud69]. He proved that the sample complexity of the unregularized optimal transport is $O(N^{-1/d})$. More formally, let $\mu \in \mathcal{D}(X)$ and $\nu \in \mathcal{D}(Y)$. Then we can define the *empirical* measures $\hat{\mu}_N$ and $\hat{\nu}_N$ as

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i} \hat{\nu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{Y_i} \tag{6.1}$$

where $(X_1, \dots, X_N)$ and $(Y_1, \dots, Y_N)$ are samples of size $N$ from $\mu$ and $\nu$, respectively. For $d > 2$, Dudley proved that

$$\mathbb{E}\left(|C(\mu,\nu) - C(\hat{\mu}_N, \hat{\nu}_N|\right) = O\left(N^{-\frac{1}{d}}\right). \tag{6.2}$$

In the same paper, Dudley proved that this rate is tight when $X = \mathbb{R}^d$ and if one of the two measures has a density with respect to the Lebesgue measure. Hence this is the best complexity that we can have in the most general case. However, sharper and more refined results have been developed since then, see for example [BGV07], [FG15], [WB19] and [HSM22].

More recently, in 2019, a similar result for the *regularized* optimal transport was presented in [Gen+19].

**Theorem 6.1.1** ([Gen+19, Theorem 3]). *Suppose that $c$ is L-Lipschitz. Then one has*

$$\mathbb{E}\left(|C^\varepsilon(\mu,\nu) - C^\varepsilon(\hat{\mu}_N,\hat{\nu}_N|\right) = O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\sqrt{N}}\left(1+\frac{1}{\varepsilon^{d/2}}\right)\right), \tag{6.3}$$

*where $\kappa = 2L\,\mathrm{diam}(X) + \|c\|_\infty$ and the constants only depend on $\mathrm{diam}(X)$, $\mathrm{diam}(Y)$, $d$ (the dimension of the ambient space) and $\|c^{(k)}\|_\infty$ for $k = 0,\dots,\lfloor\frac{d}{2}\rfloor$.*

This kind of results can also be stated in the following fashion, using concentration inequalities.

**Corollary 6.1.2** ([Gen+19, Corollary 1]). *Assuming the same setup as in Theorem 6.1.1, with probability at least $1-\delta$ one has*

$$|C^\varepsilon(\mu,\nu) - C^\varepsilon(\hat{\mu}_N,\hat{\nu}_N)| \le 6B\frac{\lambda K}{\sqrt{N}} + C\sqrt{\frac{2\log\frac{1}{\delta}}{N}}, \tag{6.4}$$

*for some constants $C$, $\lambda$ and $K$.*

**Remark 6.1.3.** Loosely speaking, we can say that the unregularized OT has sample complexity $O(N^{-1/d})$, while the regularized OT has sample complexity $O(N^{-1/2})$; though in the second case there is a constant that depends on $\varepsilon$ hidden in the big O notation. This shows that in general it might be worth considering to work with EOT rather than OT.

The above theorems give quantitive results for the convergence of $C(\hat{\mu}_N,\hat{\nu}_N)$ and $C^\varepsilon(\hat{\mu}_N,\hat{\nu}_N)$ as $N\to\infty$. Now we present a result regarding the convergence of optimal entropic potentials, $a_N$ and $b_N$.

**Theorem 6.1.4** ([Lui+19, Theorem 6]). *Let $X$ be compact. Then, there exists a constant $\mathbf{r} = \mathbf{r}(X,c,d)$ such that with probability at least $1-\tau$ one has*

$$\|a - a_N\|_\infty \le \frac{8\varepsilon\mathbf{r}e^{3D/\varepsilon}\log\left(\frac{3}{\tau}\right)}{\sqrt{N}}, \tag{6.5}$$

*where $D = \sup_{x,y\in X} c(x,y)$. We have a similar result for $\|b - b_N\|_\infty$.*

Now with this result we can get convergence rates for the density of the optimal entropic plan $g^{N,\varepsilon}$.

**Proposition 6.1.5.** *Let $g^{N,\varepsilon}$ and $g^\varepsilon$ be the densities of the optimal entropic plan for $(\nu^N,\mu^N)$ and $(\nu,\mu)$ respectively, where the measures $\mu^N$, $\nu^N$ have the same form as in Equation (6.1). Then, there exists a constant $\mathbf{r} = \mathbf{r}(X,c,d)$ such that with probability at least $1-\tau$ one has*

$$\|g^{N,\varepsilon} - g^\varepsilon\|_\infty \le \frac{C_d e^{\frac{3D^2}{\varepsilon}}\log\frac{3}{\tau}}{\sqrt{N}}, \tag{6.6}$$

*for some constant $C_d$ which only depends on $d$ (and $X$) and $D = \mathrm{diam}(X)$.*

*Proof.* By Proposition 2.2.3, we have that

$$g^{\varepsilon}(x,y) = \exp\left(\frac{-c(x,y)+a(x)+b(y)}{\varepsilon}\right), \quad g^{N,\varepsilon}(x,y) = \exp\left(\frac{-c(x,y)+a^N(x)+b^N(y)}{\varepsilon}\right).$$

Using the well known inequality $|e^{-x} - e^{-y}| \leq |x-y|$ for all $x, y \geq 0$ we get

$$\begin{aligned}\left|g^{N,\varepsilon}(x,y) - g^{\varepsilon}(x,y)\right| &\leq \frac{1}{\varepsilon}\left|(a^N(x)-a(x)) + (b^N(y)-b(y))\right| \\ &\leq \frac{2C_d e^{\frac{3D^2}{\varepsilon}} \log\frac{3}{\tau}}{\sqrt{N}}, \text{ by Theorem 6.1.4.}\end{aligned}$$

□

**Corollary 6.1.6.** *Let $t^{N,\varepsilon}$ and $t^{\varepsilon}$ be the kernels for the single smoothing operators $T^{N,\varepsilon}$ and $T^{\varepsilon}$, respectively. Then, there exists a constant $\mathbf{r} = \mathbf{r}(X,c,d)$ such that with probability at least $1-\tau$ one has*

$$\|t^{N,\varepsilon} - t^{\varepsilon}\|_{\infty} \leq \frac{C_d e^{\frac{3D^2}{\varepsilon}} \log\frac{3}{\tau}}{\sqrt{N}}, \tag{6.7}$$

*for some constant $C_d$ which only depends on $d$ (and $X$) and $D = \operatorname{diam}(X)$.*

*Proof.* The proof follows immediately from Proposition 6.1.5 and the fact that $t^{\varepsilon}(x,y) = g^{\varepsilon}(F(x),y)$ and $t^{N,\varepsilon}(x,y) = g^{N,\varepsilon}(F(x),y)$. □

**Remark 6.1.7.** Corollary 6.1.6 gives a convergence rate of the kernels of the entropic transfer operators $T^{N,\varepsilon}$ and $T^{\varepsilon}$. It would be very interesting if we had a similar result for the kernels $\hat{t}^{N,\varepsilon}$ and $t^{\varepsilon}$. If we had such a result for the kernels, by Proposition 3.1.6 we would have the same rates for the operator norm $\|\hat{T}^{N,\varepsilon} - T^{\varepsilon}\|$. This would open many possible results about the convergence rates of the eigenvalues and eigenfunctions of the two operators, see for example [JMS22, Theorem 2] for a result about the eigenfunctions.

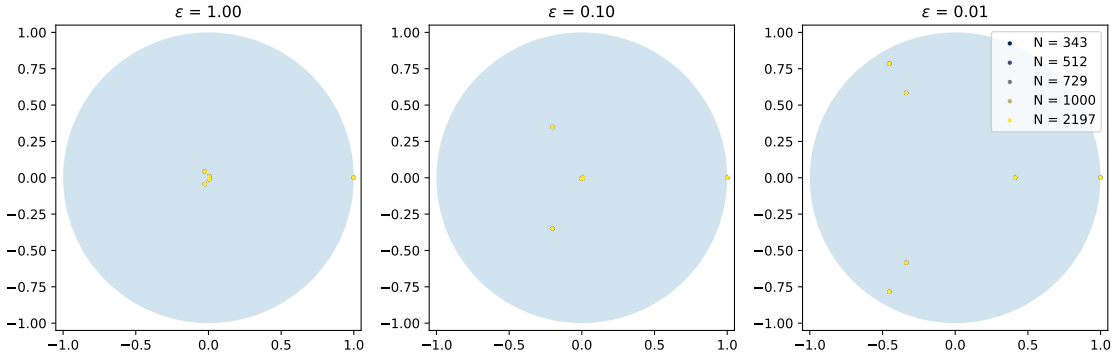
Unfortunately, such a result might not be possible due to the bad sample complexity of the unregularized OT. Recall that in order to extend the operator $T^{N,\varepsilon} : L^2(\mu^N) \to L^2(\mu^N)$ to the operator $\hat{T}^{N,\varepsilon} : L^2(\mu) \to L^2(\mu)$ we used (twice) the optimal transport plan between the measures $\mu$ and $\mu^N$, i.e. we included unregularized OT. Hence the convergence rate of $\|\hat{t}^{N,\varepsilon} - t^{\varepsilon}\|_{\infty}$ should be at least as bad as the sample complexity of OT, which is $O(N^{-1/d})$. However, in [JMS22, Section 4.6] we see that the *spectrum* of $T^{N,\varepsilon}$ and $\hat{T}^{N,\varepsilon}$ is the same, so there might be possible to get convergence rates for the eigenvalues.

## 6.2 Experimental analysis of the convergence of eigenvalues

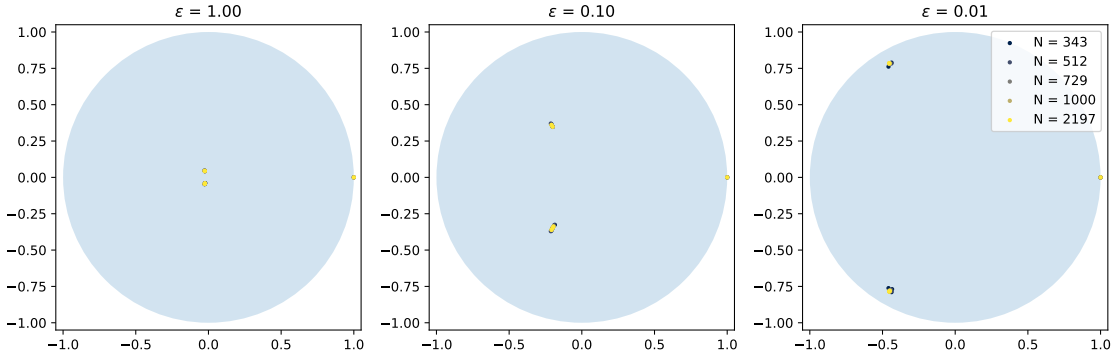
In Corollary 4.1.3 we saw that the spectrum of $T^{N,\varepsilon}$ converges to the spectrum of $T^{\varepsilon}$ as $N \to \infty$. The first goal of this section is to visualize this convergence. The second goal is to see how the speed of this convergence is affected by different regularizations $\varepsilon$ and by the ambient space dimension $d$.

We will work with the $d$-dimensional torus example (we saw a special case of it ($d = 1$) in Section 5.6), cf. [JMS22, Section 5]. Let $X = T^d \cong \mathbb{R}^d/\mathbb{Z}^d$ with the shift function $F : X \to X$, $F(x) = x + \theta \pmod 1$. Now we fix $\theta = (1/3, \ldots, 1/3)$ and $\mu$ is the uniform probability measure. For any $N$, we choose randomly $N$ points on $T^d$ and we let $\mu^N$ to be the uniform probability measure over the $N$ chosen points. In Figure 6.1 we have plotted the

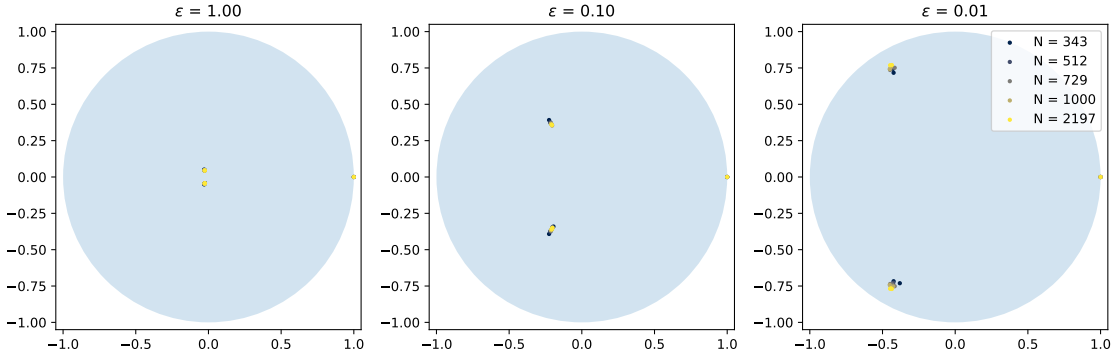
six largest eigenvalues (by absolute value) of each $T^{N,\varepsilon}$ for various dimensions $d$, various regularizations $\varepsilon$ and various sample sizes $N$.




(a) $d = 1$




(b) $d = 3$




(c) $d = 5$

Figure 6.1: The 6 largest eigenvalues (by absolute value) of $T^{N,\varepsilon}$ grouped by the dimension $d$, for $\varepsilon$ from $10^{-2}$ to 1 (logarithmically).

We can see that the eigenvalues for smaller $N$ tend to get closer and closer to the eigenvalues of the larger $N$. This numerically proves the convergence $\lambda_k^{N,\varepsilon} \to \lambda_k^{\varepsilon}$ as $N \to \infty$ for all fixed $\varepsilon$. Moreover, it is noteworty to observe that for bigger $\varepsilon$ all the eigenvalues (except the first eigenvalue which is always equal to 1) are clustered around 0, which hints towards the fact that we have faster convergence for bigger $\varepsilon$. On the other hand, we can see that for $\varepsilon = 0.01$ the convergence of the eigenvalues seems to be slower as the dimension $d$ increases. For example, for $d = 1$ we can only see the yellow points (the limit) while for $d = 5$ there is

a small but clearly visible spread of the colors.

Now we will focus more in the convergence speed of the eigenvalues since it is not very clear how exactly it is affected by the dimension $d$. In order to visualize the speed of the convergence we would like to plot the differences $\left|\lambda_k^{\varepsilon} - \lambda_k^{N,\varepsilon}\right|$ as $N \to \infty$ for various $\varepsilon$ and $d$. Unfortunately we do not have a closed formula for $\lambda_k^{\varepsilon}$ and we cannot calculate them numerically as we do for $\lambda_k^{N,\varepsilon}$. However, there is an approximation of $\lambda_k^{\varepsilon}$ given by the following proposition.

**Proposition 6.2.1** ([JMS22, Proposition 3]). *For any $\varepsilon > 0$, a complete system of eigenfunctions of $T^{\varepsilon}$ is given by the $\varphi_k$ with $\varphi_k(x) = e^{2\pi i k x}$. The respective eigenvalues $\lambda_k^{\varepsilon}$ satisfy*

$$\left|\lambda_k^{\varepsilon} - e^{-\pi^2 \varepsilon |k| \cdot e^{-2\pi i k \theta}}\right| \leq 2^{d+1} e^{-\frac{1}{8\varepsilon}}$$

*uniformly for $0 < \varepsilon < 1/(8(d+2)\ln 2)$.*

Unfortunately, this approximation works only for a specific set of $\varepsilon$ and more importantly the right hand side does not to go to 0 as $N \to \infty$. In order to visualize the errors $\left|\lambda_k^{\varepsilon} - \lambda_k^{N,\varepsilon}\right|$ we propose the following approach: Since $\lambda_k^{N,\varepsilon} \to \lambda_k^{\varepsilon}$ as $N \to \infty$ we get that $\lambda_k^{\varepsilon} \approx \lambda_k^{N_0,\varepsilon}$ for some $N_0$ big enough. Hence $\left|\lambda_k^{\varepsilon} - \lambda_k^{N,\varepsilon}\right| \approx \left|\lambda_k^{N_0,\varepsilon} - \lambda_k^{N,\varepsilon}\right|$. In our experiments we set $N_0 = 2197$. We need to mention that in order to identify the eigenvalues $\lambda_k^{N,\varepsilon}$ for different $N$ (i.e. which eigenvalue corresponds to which "part" of the spectrum), we matched them according to the solution of the assignment problem with respect to the euclidean distance. Again we work with the 6 largest eigenvalues (by absolute value). The results are shown in Figures 6.2 to 6.4, the y axis is logarithmically scaled. Note that the first plot in each figure is just a straight line because all operators have the eigenvalue $\lambda_1^{N,\varepsilon} = 1$.

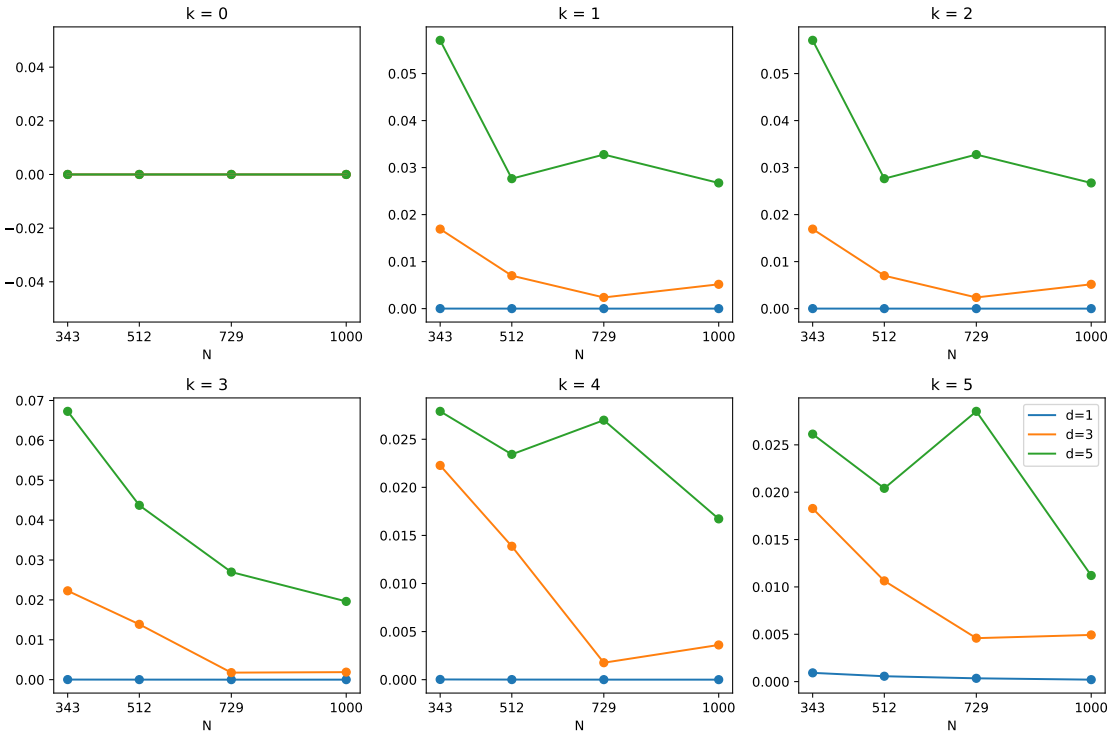



Figure 6.2: The differences $\left|\lambda_k^{2197,\varepsilon} - \lambda_k^{N,\varepsilon}\right|$ with $\varepsilon = 0.01$ for various $k$.

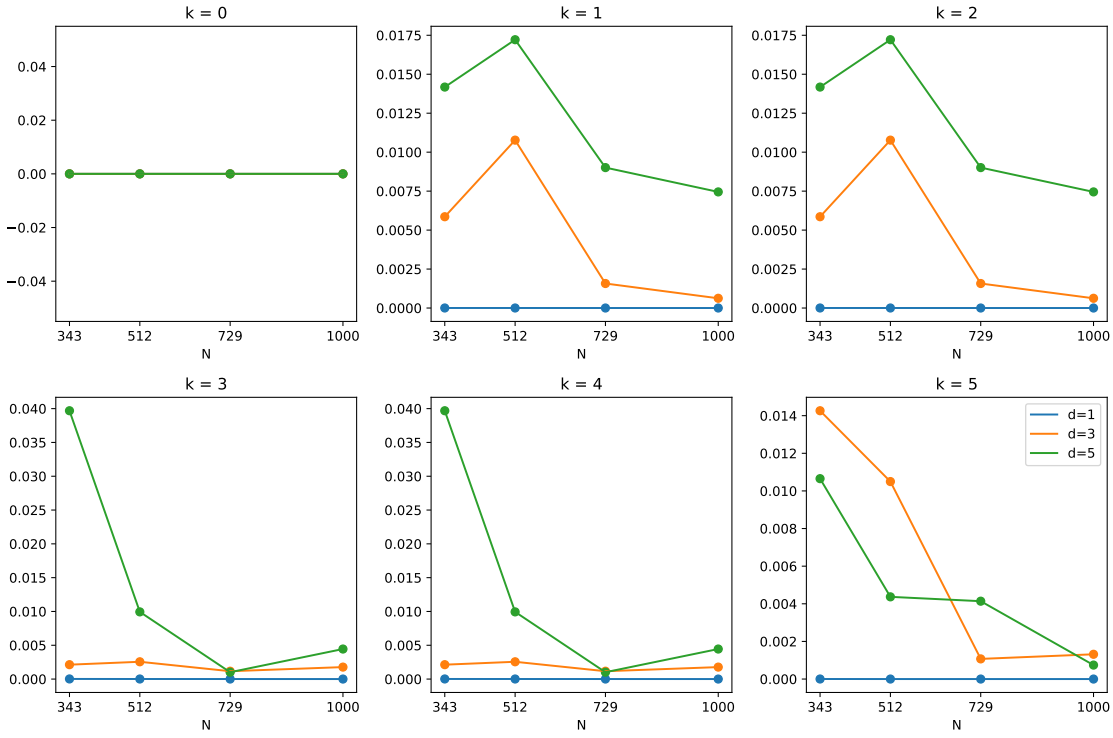



Figure 6.3: The differences $\left|\lambda_k^{2197,\varepsilon} - \lambda_k^{N,\varepsilon}\right|$ with $\varepsilon = 0.1$ for various $k$.

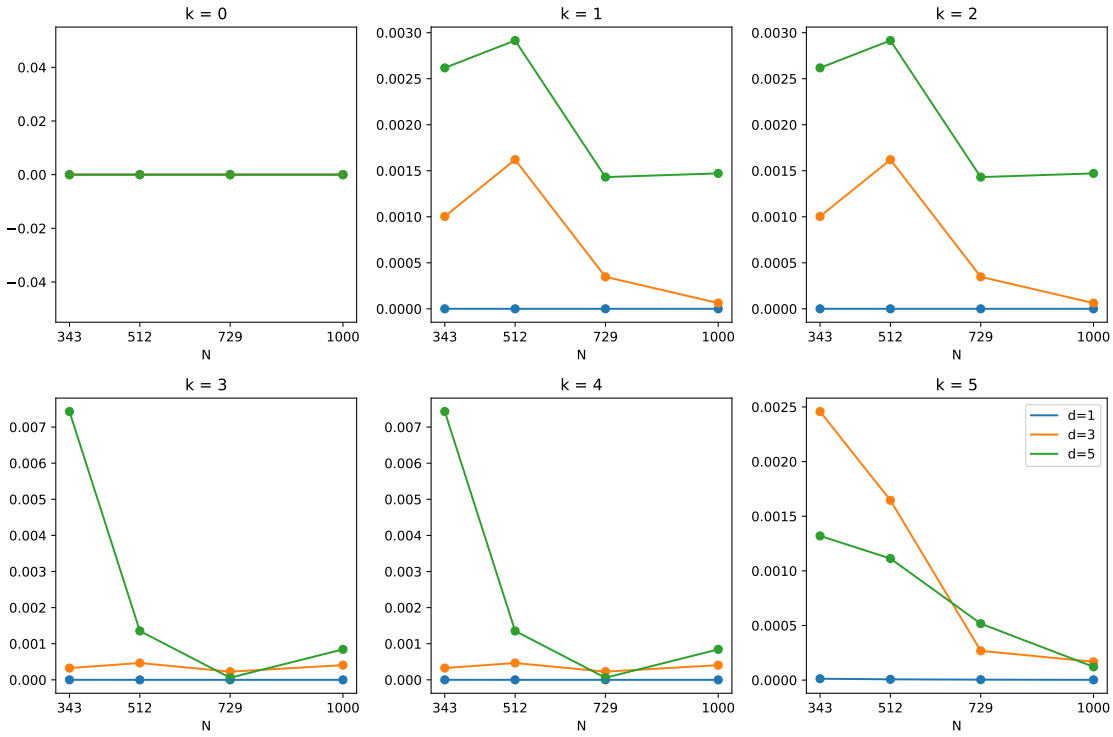



Figure 6.4: The differences $\left|\lambda_k^{2197,\varepsilon} - \lambda_k^{N,\varepsilon}\right|$ with $\varepsilon = 1.0$ for various $k$.

These plots clearly show that the eigenvalues $\lambda_k^{N,\varepsilon}$ indeed converge to $\lambda_k^{\varepsilon}$ as $N \to \infty$.

Moreover, we observe that as the dimension $d$ of the ambient space increases, the convergence rate gets slower. This can be seen from the fact that in almost all of the plots the green line ($d = 5$) tends to be above the orange line ($d = 3$) which itself tends to be above the blue line ($d = 1$). On the other hand, note that when $\varepsilon$ increases, the convergence rate gets faster. This is also supported by Figure 6.5, where we take the largest error of the eigenvalues for all $N$ and d for each $\varepsilon$.

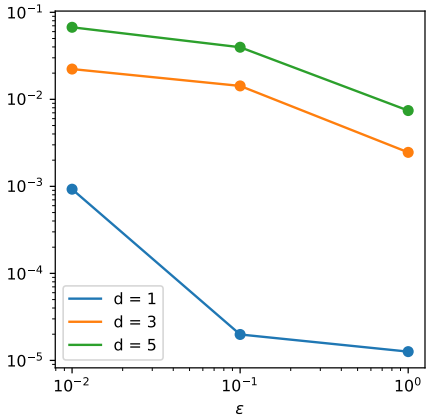



Figure 6.5: The maximum errors for each $\varepsilon$

Finally in Figure 6.6 we give another way to visualize the fact that in higher dimensions we need more points, i.e. bigger $N$, in order to achieve similar convergence results. In this plot we fix $N = 1000$ and we also plot the *full* spectrum of $T^{N,\varepsilon}$. Notice that as $d$ gets bigger and $\varepsilon$ gets smaller (i.e. we have little regularization) while we keep $N$ fixed, the point cloud of the eigenvalues seems to be more spread out.

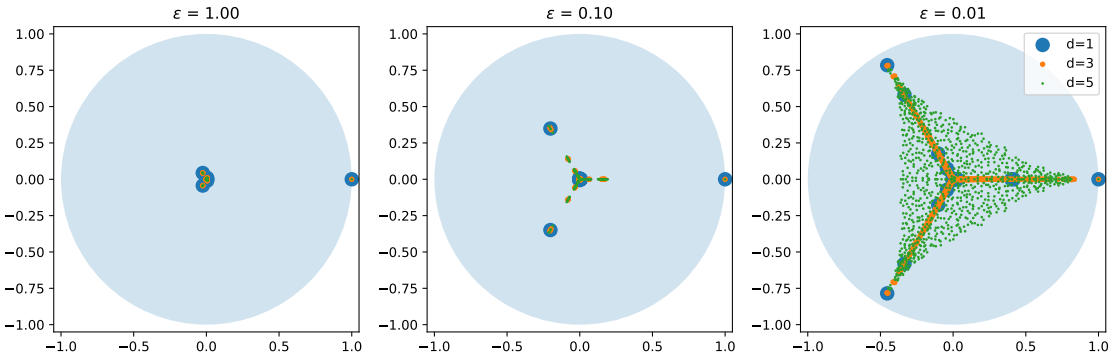



Figure 6.6: Full spectrum of $T^{N,\varepsilon}$ with fixed $N = 1000$ for various $d$ and $\varepsilon$.

We conclude this section with our general thoughts about the above plots. The numerical examples that we presented seem to suggest that the convergence rate of the difference $\left|\lambda_k^{\varepsilon} - \lambda_k^{N,\varepsilon}\right|$ in the torus example is better when $\varepsilon$ increases and worse when $d$ increases. Our intuition is that the rate should be similar as the one established in Corollary 6.1.6 for all systems, not just the $d$-dimensional torus. Further research is required to theoretically explain the above findings.

# CHAPTER 7

# CONCLUSION

## 7.1 Summary

In this thesis we gave an overview of recent results of the theory of entropic transfer operators and we generalized the results appeared in [JMS22] for the stochastic case. In particular, we first proved a result about the single smoothed entropic operator in the stochastic setting which gave us convergence in the $L^2$ norm but it was not enough to give us convergence in the operator norm and hence convergence of the spectra. Then we introduced the double smoothed entropic operator which has much better theoretical guarantees (i.e. we get operator norm and eigenvalue convergence) and we also presented some numerical examples comparing the two kinds of entropic operators. Furthermore, we tried to establish convergence rates for the kernels of the single smoothed entropic operator, $t^{N,\varepsilon}$ and $\hat{t}^{N,\varepsilon}$ based on sample complexities of the potentials of EOT. Finally we explored numerically the convergence rates of the convergence of eigenvalues of $T^{N,\varepsilon}$ to the eigenvalues of $T^{\varepsilon}$.

## 7.2 Future Work

Here are a few thoughts and remarks that have arisen during the writing of this thesis and it would be interesting to further explore.

- One of our main contributions in this thesis is the introduction of double smoothed entropic transfer operators. This happened because we could not prove the operator norm convergence in the stochastic case, even though we had something slightly weaker, cf. Theorem 4.2.3. Hence the question is: Can we somehow push the single smoothed approach to work for the stochastic case? If not, can we pinpoint exactly why and where it fails?
- Another possible area to think more about is the interaction between the double double and single smoothed entropic operators. In particular, can we have a formal proposition about the relation of the spectrum $\hat{T}_1^{N,\varepsilon}$ (single smoothing) and $\hat{T}_2^{N,\varepsilon}$ (double smoothing)?
- Finally an interesting research direction is to try to establish theoretical results regarding the convergence rates of the eigenvalues $\lambda_k^{N,\varepsilon}$. In our work with the $d$-dimensional

torus we observed that the convergence rate of $\lambda_k^{N,\varepsilon}$ of this system seemed to be "close" to $O\left(\frac{1}{\text{poly}(N)}\right)$ (there are constants in the big O notation that depended from $\varepsilon$ and $d$). Hence we pose the following conjecture.

**Conjecture.** Regardless of the system $(X, F, \mu)$, the convergence of the eigenvalues of the (single) entropic tranfer operator $\hat{T}^{N,\varepsilon}$ to the eigenvalues of $T^{\varepsilon}$, have convergence rate given by $O\left(\frac{C_d e^{\frac{3D^2}{\varepsilon}}}{\sqrt{N}}\right)$ where $D = \text{diam}(X)$ and $C_d$ is a constant that depends exponentially on $d$.

# APPENDIX A

# SOME MATHEMATICAL BACKGROUND

In this chapter we are going to review some basic results from measure theory and real analysis that are useful for the proofs in the main chapters.

We assume that the reader is already familiar with basic measure theory such as $\sigma$-algebras, measurable spaces, general measures, measurable functions, Lebesgue integral, $L^p$ spaces, the Fubini-Tonelli theorem, the Radon-Nikodym theorem and the Dominated Convergence theorem.

## A.1 Pushforward

In this section we are giving the definition of the pushforward of a measure via a function.

**Definition A.1.1.** Let $X, Y$ be metric spaces and let $F : X \to Y$ be a measurable function. Fix a measure $\mu \in \mathcal{M}(X)$. We define the *pushforward of $\mu$ via $F$* to be the measure $F_\sharp\mu \in \mathcal{M}(Y)$, given by

$$\left(F_\sharp\mu\right)(B) = \mu\left(F^{-1}(B)\right),$$

for all $B \in \mathcal{B}(Y)$.

**Remark A.1.2.** Intuitively, if a metric space $X$ is distributed according to a probability measure $\mu \in \mathcal{P}(X)$ and $F : X \to Y$ is a function, then the image $F(X)$ is distributed according to the probability measure $F_\sharp\mu \in \mathcal{P}(Y)$.

Now we see the action of the pushforwards with respect to integrals.

**Proposition A.1.3** ([BR07, Theorem 3.6.1])**.** *The pushforward construction satisfies the following property: Let $X, Y$ be metric spaces, let $\mu \in \mathcal{P}(X)$ and let $F : X \to Y$ be a measurable function. Then for any $g : Y \to \mathbb{R}$ measurable function we have*

$$\int_Y g(y)\,\mathrm{d}F_\sharp\mu(y) = \int_X g\left(F(x)\right)\mathrm{d}\mu(x).$$

**Example A.1.4** (Discrete case)**.** *Suppose that $\mu$ is a discrete measure, i.e. $\mu = \sum_{i=1}^n \mu_i\delta_{x_i}$ for some $x_1, \dots, x_n \in X$ and $\mu_i \geq 0$ with $\sum_{i=1}^n \mu_i = 1$. Then its pushforward is given by*

$$F_\sharp\mu(A) = \sum_i \mu_i\delta_{x_i}(F^{-1}(A)) = \sum_i \mu_i\delta_{F(x_i)}(A),$$

*i.e. $F_\sharp\mu = \sum_{i=1}^n \mu_i\delta_{F(x_i)}$*

## A.2 Disintegration

In this section we are going to review the disintegration theorem which intuitively is the "reverse" process of the construction of a product measure.

**Theorem A.2.1** (Disintegration, [AGS05, Theorem 5.3.1]). *Let $X$, $Y$ be metric spaces, let $\pi : Y \to X$ be a Borel measurable function, $\mu \in \mathcal{P}(Y)$ and $\nu = \pi_\sharp \mu \in \mathcal{P}(X)$. Then there exists a $\nu$-a.e. uniquely determined Borel family of probability measures $(\mu_x)_{x\in X} \subseteq \mathcal{P}(Y)$, which provides a* disintegration *of $\mu$ into $\{\mu_x\}_{x\in X}$, such that:*

- *For every (fixed) $B \in \mathcal{B}(Y)$, the map $x \mapsto \mu_x(B)$ is Borel measurable.*
- *The measure $\mu_x$ "lives" on the fiber $\pi^{-1}(x)$:*
$$\mu_x(Y \setminus \pi^{-1}(x)) = 0 \text{ for } \nu\text{-a.e. } x \in X,$$
*so $\mu_x(B) = \mu_x(B \cap \pi^{-1}(x))$.*
- *For every $f : Y \to [0, +\infty]$ Borel measurable function we have*
$$\int_Y f(y)\, \mathrm{d}\mu(y) = \int_X \left( \int_{\pi^{-1}(x)} f(y)\, \mathrm{d}\mu_x(y) \right) \mathrm{d}\nu(x).$$

**Applications.** Now we present some applications of the disintegration theorem. We begin by seeing how we can "restrict" a measure in a product space $X \times Y$ to one of its components $X$.

**Proposition A.2.2** (Disintegration of product measures). *Let $Y = X_1 \times X_2$, $\mu \in \mathcal{P}(X_1 \times X_2)$ and $\pi^1 : X_1 \times X_2 \to X_1$ be the projection to the first coordinate. Then $(\pi^1)^{-1}(x_1) = \{x_1\} \times X_2 \cong X_2$. By disintegration, there exists a Borel family of probability measures $\{\mu_{x_1}\}_{x_1 \in X_1} \subseteq \mathcal{P}(X_1 \times X_2)$ such that:*

- *The measure $\mu_{x_1}$ is a probability measure in $X_2$. Indeed we have*
$$\mu_{x_1}\left((X_1 \times X_2) \setminus (\pi^1)^{-1}(x_1)\right) = 0 \Rightarrow \mu_{x_1}\left((X_1 \setminus \{x_1\}) \times X_2\right) = 0$$
*This means that $\operatorname{supp}(\mu_{x_1}) \subseteq \{x_1\} \times X_2 \cong X_2$. Hence we can assume that $\{\mu_{x_1}\}_{x_1 \in X_1} \subseteq \mathcal{P}(X_2)$.*
- *For every $f : X_1 \times X_2 \to [0, +\infty]$ Borel measurable function we have*
$$\int_{X_1 \times X_2} f(x_1, x_2)\, \mathrm{d}\mu(x_1, x_2) = \int_{X_1} \left( \int_{X_2} f(x_1, x_2)\, \mathrm{d}\mu_{x_1}(x_2) \right) \mathrm{d}\pi^1_\sharp \mu(x_1).$$
*In particular, taking $f$ to be the indicator function of $A_1 \times A_2$ we have that*
$$\begin{aligned}
\mu(A_1 \times A_2) &= \int_{X_1} \left( \int_{X_2} 1_{A_1 \times A_2}(x_1, x_2)\, \mathrm{d}\mu_{x_1}(x_2) \right) \mathrm{d}\pi^1_\sharp \mu(x_1) \\
&= \int_{X_1} \left( \int_{X_2} 1_{A_1}(x_1) 1_{A_2}(x_2)\, \mathrm{d}\mu_{x_1}(x_2) \right) \mathrm{d}\pi^1_\sharp \mu(x_1) \\
&= \int_{X_1} 1_{A_1}(x_1) \left( \int_{X_2} 1_{A_2}(x_2)\, \mathrm{d}\mu_{x_1}(x_2) \right) \mathrm{d}\pi^1_\sharp \mu(x_1) \\
&= \int_{X_1} 1_{A_1}(x_1) \mu_{x_1}(A_2)\, \mathrm{d}\pi^1_\sharp \mu(x_1) \\
&= \int_{A_1} \mu_{x_1}(A_2)\, \mathrm{d}\pi^1_\sharp \mu(x_1),
\end{aligned}$$
*for any $A_1 \in \mathcal{B}(X_1)$ and $A_2 \in \mathcal{B}(X_2)$.*

**Remark A.2.3** (Disintegration of product measure)**.** Let $X$, $Y$ be metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $h : X \times Y \to [0,1]$. Then the disintegration of $h(\mu \times \nu) \in \mathcal{P}(X \times Y)$ is given by $[h(\mu \times \nu)]_y = h(-,y)\mu$ and $[h(\mu \times \nu)]_x = h(x,-)\nu$. Indeed using Fubini's theorem we have

$$\begin{aligned}\int_X \int_Y f(x,y)\, \mathrm{d}[h(\mu \times \nu)]_x(y)\, \mathrm{d}\mu(x) &= \int_{X\times Y} f(x,y)\, \mathrm{d}[h(\mu \times \nu)](x,y) \\ &= \int_{X\times Y} f(x,y)h(x,y)\, \mathrm{d}(\mu \times \nu)(x,y) \\ &= \int_X \int_Y f(x,y)h(x,y)\, \mathrm{d}\nu(y)\, \mathrm{d}\mu(x) \\ &= \int_X \int_Y f(x,y)\, \mathrm{d}[h(x,-)\nu](y)\, \mathrm{d}\mu(x),\end{aligned}$$

for all measurable functions $f : X \times Y \to \mathbb{R}$.

We conclude with another application of the disintegration theorem which will prove useful later. It is essentially Jensen's inequality for Radon-Nikodym derivatives with respect pushforwards.

**Lemma A.2.4** ([Cai+22, Lemma 3.15])**.** *Let $X, Y$ be metric spaces and $F : X \to Y$ measurable. Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ such that $\mu \ll \nu$ (which implies that $F_\sharp \mu \ll F_\sharp \nu$). Finally, let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Then*

$$\int_Y f\left(\frac{\mathrm{d}F_\sharp \mu}{\mathrm{d}F_\sharp \nu}\right) \mathrm{d}F_\sharp \nu \leq \int_X f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu.$$

## A.3 Modulus of continuity

In this section we recall the notion of a modulus of continuity. This formalizes the idea of uniform continuity in metric spaces. We also state the Arzela-Ascoli theorem.

**Definition A.3.1.** Let $(X,\rho)$ and $(Y,\sigma)$ be metric spaces. We say that a function $f : X \to Y$ admits a modulus of continuity if there exists a function $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\omega(0) = 0$ that is continous, increasing and concave which satisfies the following

$$\sigma(f(x), f(x')) \leq \omega(\rho(x,x')), \quad \forall x, x' \in X.$$

**Remark A.3.2.** Obviously a function $f : X \to Y$ is uniformly continous if and only if it admits a modulus of continuity.

Now using moduli of continuity we can define the uniform equicontinuity of a family of functions.

**Definition A.3.3.** Let $(X,\rho)$ and $(Y,\sigma)$ be metric spaces. Let $\mathcal{F}$ be a family of functions from $X$ to $Y$. We call the family $\mathcal{F}$ uniformly equicontinuous if and only if every function $f \in \mathcal{F}$ admits a common modulus of continuity $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.

**Theorem A.3.4** (Arzela-Ascoli, [Rud76])**.** *Let $(X,d)$ be a compact metric space and let $f_n : X \to \mathbb{R}$ be continuous functions. Suppose that $\{f_n\}_n$ is uniformly bounded, i.e. there is a $M \geq 0$ such that $\|f_n\|_\infty \leq M$ for all $n \in \mathbb{N}$, and uniformly equicontinuous. Then there exists a subsequence $\{f_{k_n}\}$ of $\{f_n\}_n$ that converges uniformly to a continous function $f : X \to Y$.*

# BIBLIOGRAPHY


[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[BR07] Vladimir Igorevich Bogachev and Maria Aparecida Soares Ruas. *Measure theory*. Vol. 1. Springer, 2007.

[BGV07] François Bolley, Arnaud Guillin, and Cédric Villani. "Quantitative concentration inequalities for empirical measures on non-compact spaces". In: *Probability Theory and Related Fields* 137.3 (2007), pp. 541–593.

[Cai+22] Tianji Cai, Junyi Cheng, Bernhard Schmitzer, and Matthew Thorpe. "The Linearized Hellinger–Kantorovich Distance". In: *SIAM Journal on Imaging Sciences* 15.1 (2022), pp. 45–83.

[Cut13] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems* 26 (2013).

[DJ99] Michael Dellnitz and Oliver Junge. "On the Approximation of Complicated Dynamical Behavior". In: *SIAM Journal on Numerical Analysis* 36.2 (1999), pp. 491–515.

[Dud69] Richard Mansfield Dudley. "The speed of mean Glivenko-Cantelli convergence". In: *The Annals of Mathematical Statistics* 40.1 (1969), pp. 40–50.

[DS88] Nelson Dunford and Jacob T Schwartz. *Linear operators, part 1: general theory*. Vol. 10. John Wiley & Sons, 1988.

[FG15] Nicolas Fournier and Arnaud Guillin. "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability Theory and Related Fields* 162.3 (2015), pp. 707–738.

[Gen+19] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. "Sample complexity of sinkhorn divergences". In: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, pp. 1574–1583.

[Hsu81] Chieh-Su Hsu. "A generalized theory of cell-to-cell mapping for nonlinear dynamical systems". In: *Trans. ASME Ser. E. J. Appl. Mech.* 48.3 (1981), pp. 634–642.

[HSM22] Shayan Hundrieser, Thomas Staudt, and Axel Munk. "Empirical Optimal Transport between Different Measures Adapts to Lower Complexity". In: *arXiv preprint arXiv:2202.10434* (2022).

[JMS22] Oliver Junge, Daniel Matthes, and Bernhard Schmitzer. "Entropic transfer operators". In: *arXiv preprint arXiv:2204.04901* (2022).

[Kan42] Leonid V Kantorovich. “On the translocation of masses”. In: *Dokl. Akad. Nauk. USSR (NS)*. Vol. 37. 1942, pp. 199–201.

[Kle13] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

[Klu+18] Stefan Klus, Feliks Nüske, Péter Koltai, Hao Wu, Ioannis Kevrekidis, Christof Schütte, and Frank Noé. “Data-driven model reduction and transfer operator approximation”. In: *J Nonlinear Sci* 28.3 (2018), pp. 985–1010.

[Lui+19] Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. “Sinkhorn barycenters with free support via frank-wolfe algorithm”. In: *Advances in neural information processing systems* 32 (2019).

[Mon81] Gaspard Monge. “Mémoire sur la théorie des déblais et des remblais”. In: *Mem. Math. Phys. Acad. Royale Sci.* (1781), pp. 666–704.

[Nut21] Marcel Nutz. *Introduction to Entropic Optimal Transport*. 2021.

[PC19] Gabriel Peyré and Marco Cuturi. “Computational optimal transport: With applications to data science”. In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.

[Poi90] Henri Poincaré. “Sur le problème des trois corps et les équations de la dynamique”. In: *Acta mathematica* 13.1 (1890), A3–A270.

[Rud76] Walter Rudin. *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York, 1976.

[San15] Filippo Santambrogio. “Optimal transport for applied mathematicians”. In: *Birkäuser, NY* 55.58-63 (2015), p. 94.

[Sar12] Omri Sarig. “Introduction to the transfer operator method”. In: *lecture notes, Second Brazilian School on Dynamical Systems* (2012).

[Sin64] Richard Sinkhorn. “A relationship between arbitrary positive matrices and doubly stochastic matrices”. In: *The annals of mathematical statistics* 35.2 (1964), pp. 876–879.

[Ula60] Stanislaw M Ulam. *A collection of mathematical problems*. 8. Interscience Publishers, 1960.

[Vil09] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer, 2009.

[Vil21] Cédric Villani. *Topics in optimal transportation*. Vol. 58. American Mathematical Soc., 2021.

[WB19] Jonathan Weed and Francis Bach. “Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance”. In: *Bernoulli* 25.4A (2019), pp. 2620–2648.

[Yul12] G Udny Yule. “On the methods of measuring association between two attributes”. In: *Journal of the Royal Statistical Society* 75.6 (1912), pp. 579–652.