MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Entropy-Regularised Optimal Transport and Schrödinger's Problem“

verfasst von / submitted by

Jacob Harrison Hand

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2020 / Vienna, 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A UA 066 821
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Mathematik
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Ing. Dr. techn. Mathias Beiglböck

Contents

Abstract

Section 1 introduces the Optimal Transport problem and outlines the advantages of formulating this problem in terms of probability and measure theory; this will be called the Kantovorich formulation.[1] While this is better than the original formulation since it indicates that a solution exists, it is still a complex problem to solve directly (see 1.2.1). In Section 2, entropy of probability measures is introduced and it will be shown in Section 2.2.2 that the solution to the entropy-regularised Optimal Transport problem converges to the solution of the non-regularised problem as $\epsilon \to 0$. Entropy-regularised Optimal Transport is computationally easier to solve using the Sinkhorn Algorithm, discussed in Section 2.3. In Section 3, Schrödinger's problem is introduced and formulated in the language of modern probability theory to show that it is equivalent to the entropy-regularised Optimal Transport problem. In Section 3.1, the dynamic Schrödinger problem is formulated. Schilder's Theorem is presented in Section 3.2, it is the Large Deviation result used to make the link to the static Schrödinger problem in Section 3.3. In Section 3.3, the static Schrödinger problem is defined to establish the connection to the entropy-regularised Optimal Transport problem.

Abstract (German)

Abschnitt 1 stellt das Problem des optimalen Transports vor und erläutert die Vorteile der Formulierung dieses Problems im Sinne der Wahrscheinlichkeits- und Maßtheorie. Im Weiteren wird diese als Kantovorich-Formulierung bezeichnet. Diese ist besser als die ursprüngliche Formulierung , da sie die Existenz einer Lösung zeigen kann. Es ist aber immer noch nicht ideal, weil es schwierig ist das Problem direkt zu lösen (siehe 1.2.1). Im 2. Abschnitt wird die Entropie von Wahrscheinlichkeitsmaßen eingeführt und im Kapitel 2.2.2 wird gezeigt, dass die Lösung des Entropie-regulierten Optimaltransport Problems für $\epsilon \to 0$ zur Lösung des nicht-regulierten Problems konvergiert. Der Entropie-regulierte optimale Transport ist mit dem in Abschnitt 2.3 beschriebenen Sinkhorn-Algorithmus rechnerisch einfacher zu lösen. Im 3. Abschnitt wird Schrödingers Problem vorgestellt und in der Sprache der modernen Wahrscheinlichkeitstheorie formuliert, um zu zeigen, dass das Problem gleichwertig dem Entropie-regulierten Optimalen Transport Problem ist. Im Abschnitt 3.1 wird das dynamische Schrödinger-Problem formuliert. Schilder's Theorem wird im Abschnitt 3.2 vorgestellt, es ist das Large Deviation Resultat, das verwendet wird, um den Zusammenhang des statischen Schrödinger Problems im Abschnitt 3.3 herzustellen. Im Abschnitt 3.3 wird das statische Schrödinger Problem definiert, um die Verbindung zum Entropie-regulierten Optimalen Transport Problems herzustellen.

[1] Named this way because of Leonid Kantovorich: L.V. Kantorovich. On the translocation of masses. C. R. (Dokl.) Acad. Sci. URSS, 37:199–201, 1942.

1 The Optimal Transport Problem

The Optimal Transport (OT) problem can briefly be described as finding the cheapest or most efficient way to move mass from one distribution to another, without losing or gaining mass on the way. The OT problem is relevant in many applications. The main heuristic example is the transportation of sand particles from a random formation to form a sandcastle, where the goal is to minimise the work required. An example from economics is a business with a limited number of products and customers, trying to maximise profits by distributing its products to customers in an optimal way.[2] Optimal Transport is also applicable in image processing, where the similarity of two images can be represented as the closeness of the two distributions of image pixels.[3]

1.1 Original Formulation

Let X and Y be polish spaces and let $\mathcal{P}(X)$ denote the set of probability measures on the set X. From 1.7 onwards, it is supposed that $X = Y = \mathbb{R}^n$. The original formulation of the OT problem can be found in the first two pages of [19] and is as follows.

Given two probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ and a measurable map $T : X \to Y$, set the constraint $\nu(A) = (T_{\#}\mu)(A) := \mu(T^{-1}(A))$ for every measurable set $A \subseteq Y$. This constraint ensures that the requirement that the first distribution is transferred to the second without losing any mass is satisfied. Equivalently, $\int_Y \phi \; d\nu = \int_X (\phi \circ T) \; d\mu$ for all measurable functions ϕ.

Definition 1.1. *The original OT problem is to find the map $T : X \to Y$ satisfying*

$$\text{minimise } \left\{ \int c(x, T(x)) d\mu(x) : (T_{\#}\mu) = \nu \right\},$$

where the function $c : X \times Y \to [0, \infty]$ is the transport cost from X to Y.

Analysing optimal transport in this original formulation is problematic. For example, it is unclear how to define a map T that can split mass and satisfy the constraint $\nu(A) = (T_{\#}\mu)(A)$ for all $A \subseteq Y$. Suppose $\mu := \delta_x$ for some $x \in X$ and ν is anything other than a single Dirac measure. Then if $y = T(x) \in Y$, it must be a point with mass 1, because otherwise $\nu(y) \neq \mu(T^{-1}(y))$.

Another issue is that the constraint on T given by $\nu(A) = (T_{\#}\mu)(A)$ is <u>not closed</u> under weak convergence. Consider the example from [19] in Figure 1, the sequence of continuous bounded functions $T_n(x) = \sin nx$ on $[0, 2\pi]$ under the Lebesgue measure λ on $[0, 2\pi]$. For each $n \in \mathbb{N}$, the measure $\nu_n([0, 2\pi]) := \int\limits_{[0,2\pi]} T_n(x) d\lambda(x) \in \mathcal{P}(\mathbb{R})$ satisfies the constraint, but as $n \to \infty$, the measure ν_n converges weakly to $\frac{1}{4\pi}\lambda_{[0,2\pi]} \times \lambda_{[-1,1]}$ which is in $\mathcal{P}(\mathbb{R} \times \mathbb{R})$. Closure under weak convergence is imperative to prove the existence of a minimum value.

Observe again the requirement that $\int_Y \phi \; d\nu = \int_X (\phi \circ T) \; d\mu$ for all measurable functions ϕ. Consider the example $\phi = id$, the identity function on Y, let g have the law ν and let f have the law μ. By the transformation formula, the constraint condition is that

$$\int_Y g(y) dy = \int_X g(T(x)) \det DT(x) dx = \int_X f(x) dx.$$

[2] From [19], p 41.

[3] For example, see Nicolas Papadakis. Optimal Transport for Image Processing. Signal and Image Processing. Université de Bordeaux; Habilitation thesis, 2015.

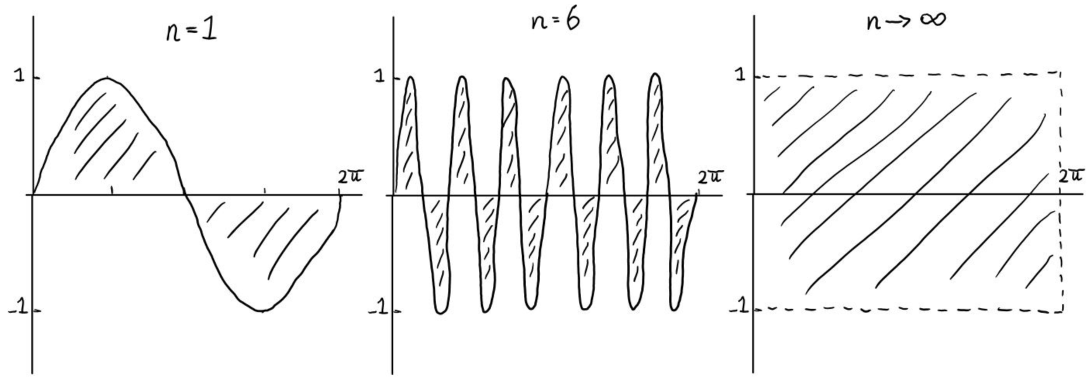

Figure 1: Example of $\nu_n([0,2\pi]) := \int\limits_{[0,2\pi]} T_n(x) d\lambda(x)$ showing that the constraint for the original OT problem is not closed under weak convergence.

Hence, establishing existence of a solution to the OT problem requires proving the existence of a solution to the nonlinear partial differential equation $g(T(x)) \det DT(x) = f(x)$. It is not clear if a map T satisfying this PDE exists.[4]

In the next section, a generalised formulation of the OT problem is introduced to overcome these issues. The Kantovorich formulation presented follows [13], page 7.

1.2 Kantovorich Formulation

Let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ be probability measures representing the distributions of the original mass and the new formation. The movement of particles from μ to ν is described by a transport plan $\gamma \in \mathcal{P}(X \times Y)$ with marginal distributions μ and ν. Instead of specifying the destination $T(x)$ for each particle $x \in X$, the value $\gamma(A, B)$ represents the amount of mass moving from $A \in X$ to $B \in Y$. Denote by $\Pi(\mu, \nu) := \{\gamma \in \mathcal{P}(X \times Y) : \gamma_X = \mu, \gamma_Y = \nu\}$ the set of all transport plans from μ to ν, where γ_X denotes the projection of γ onto the X marginal distribution.

Definition 1.2. [5] *Given a cost function $c : X \times Y \to [0, \infty]$, the OT problem is to find*

$$\inf_{\gamma\in\Pi(\mu,\nu)}(c, \gamma) := \inf \left\{ \int\limits_{X\times Y} c(x, y)\ d\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\}.$$

Note that, in this manuscript, the focus will be on the OT problem with cost function

$$c(x, y) := |x - y|^2.$$

The advantage of the Kantovorich formulation can be seen by addressing the three highlighted issues with the original problem. Firstly, splitting the mass of the Dirac measure becomes manageable. The plan γ has marginal distributions μ and ν, which means that for all measurable sets $A \subseteq X$ and $B \subseteq Y$, $\gamma(X \times B) = \nu(B)$ and $\gamma(A \times Y) = \mu(A)$. Moreover, $\gamma(x, y)$ represents the mass being moved from point $x \in X$ to $y \in Y$. Hence, if μ is a Dirac measure at $x \in X$ and

[4] [19], p 2.
[5] From [4], p4.

this mass needs to be sent to many points in Y, the transport plan γ can be defined accordingly. For example, see Figure 2.

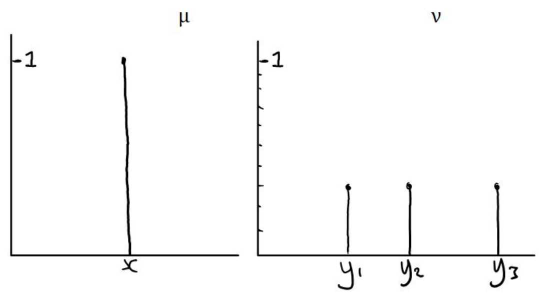

Figure 2: Although there is no map $T : X \to Y$ that can split the mass of δ_x, we can define $\gamma \in \Pi(\mu, \nu)$ as $\gamma(x, y_1) = \gamma(x, y_2) = \gamma(x, y_3) = \frac{1}{3}$ and $\gamma(x, y) = 0$ for all other $y \in Y$.

The second issue encountered with the original OT problem was that the space of solutions is not closed. In the Kantovorich formulation, this is addressed by the following result.

Theorem 1.3. $\Pi(\mu, \nu)$ *is closed.*

Proof: Suppose a sequence of probability measures $(\gamma_k)_{k\in\mathbb{N}} \in \Pi(\mu, \nu)$ converges weakly to some measure γ on $X \times Y$. Then for all continuous bounded functions $f : X \times Y \to \mathbb{R}$, $\int f(x, y)d\gamma_k(x, y) \to \int f(x, y)d\gamma(x, y)$ as $k \to \infty$. Consequently,

$$1 = \int \mathbb{1}_{X\times Y} d\gamma_k \to \int \mathbb{1}_{X\times Y} d\gamma,$$

so $\gamma(X \times Y) = 1$ and γ is a probability measure. To show that γ has the correct marginal distributions, choose $f(x, y) = \hat{f}(x)$, where $\hat{f}$ is continuous and bounded. Then,

$$\int_X \hat{f}(x)d\gamma_k(x) \to \int\limits_{X\times Y} \hat{f}(x)d\gamma(x, y) = \int_X \hat{f}(x)d\gamma_X(x)$$

where γ_X is the X-marginal distribution. Therefore, the X-marginals of the sequence $(\gamma_k)_{k\in\mathbb{N}}$, defined as μ for each k, converge to the X-marginal of γ. Since this holds for any continuous bounded function $f : X\times Y \to \mathbb{R}$, it follows that $\gamma_X = \mu$. Similarly, $\gamma_Y = \nu$. That is, $\gamma \in \Pi(\mu, \nu)$ and so $\Pi(\mu, \nu)$ is closed. $\square$

Recall that the existence of a possible solution to the OT problem in the original formulation could not be guaranteed. For the Kantovorich formulation, there always exists a transport plan $\gamma \in \Pi(\mu, \nu)$. Indeed, a transport plan can be defined as the product measure $\mu\otimes\nu$, which always exists. For $A \subset X$ and $B \subset Y$, $(\mu \otimes \nu)(A \times B) = \mu(A) \cdot \nu(B)$.

In the rest of Section 1, some useful results about the OT problem will be shown. These results follow [19], [1] and [4].

The Kantorovich formulation is indeed a generalisation of the original OT problem. Take the plan $\gamma = (id, T)_{\#}\mu$, which belongs to $\Pi(\mu, \nu)$ if and only if $\gamma(X, A) = \nu(A) = \mu(T^{-1}(A))$ for any $A \in X$. Therefore, the constraint from Definition 1.1 holds and the functional to minimise is $(c, \gamma) = \int_X c(x, T(x))d\mu(x)$, thus generalising the original problem. The original formulation then represents a special case of optimal transport, when the transport plan is given by some deterministic map T.

Theorem 1.4. $\Pi(\mu,\nu)$ *is tight.*

Proof Fix $\epsilon > 0$ and find two compact sets $K_X \subset X$ and $K_Y \subset Y$ such that $\mu(K_X) > 1-\frac{\epsilon}{2}$ and $\nu(K_Y) > 1-\frac{\epsilon}{2}$. These two compact sets exist because any single probability measure is tight, by total probability. Then, the set $K_X \times K_Y$ is compact in $X \times Y$ and for any $\gamma \in \Pi(\mu,\nu)$,

$$\begin{aligned}\gamma((X\times Y)\backslash(K_X\times K_Y)) &\leq \gamma(X\backslash K_X)+\gamma(Y\backslash K_Y)\\ &= \mu(X\backslash K_X)+\nu(Y\backslash K_Y)\\ &\leq \frac{\epsilon}{2}+\frac{\epsilon}{2}\\ &= \epsilon.\end{aligned}$$

□

Definition 1.5. *A function $f : X \to \mathbb{R}$ is called lower semi-continuous* (l.s.c.) *if $\forall\ a \in \mathbb{R}$ and $\forall x \in X$ the set $\{x \in X : f(x) \leq a\}$ is closed. Equivalently, for $x_k \to x$, $f(x) \leq \liminf\limits_{k\to\infty} f(x_k)$.*

Theorem 1.6. *If $f : X \to \mathbb{R} \cup \{\infty\}$ is lower semi-continuous and X is compact, then there exists $\hat{x} \in X$ such that $f(\hat{x}) = \inf\{f(x) : x \in X\}$.*

Proof: Define $l := \inf\{f(x) : x \in X\}$. By definition of the infimum, there exists a sequence $(x_n)_{n\in\mathbb{N}} \in X$ such that $f(x_n) \to l$. By compactness, there is a subsequence $(x_j)_{j\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ and an $\hat{x} \in X$ such that $x_j \to \hat{x}$. The lower semi-continuity of f ensures that $f(\hat{x}) \leq \liminf\limits_{j\to\infty} f(x_j) = l$ and the opposite inequality holds since l is the infimum of f. Hence, $l = f(\hat{x})$ is the minimum value of f, realised at $\hat{x} \in X$. □

Theorem 1.7. [6] *There exists a solution to the OT problem with cost function $c(x,y) = |x-y|^2$. That is, there exists $\hat{\gamma} \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$ such that*

$$\inf\left\{\int\limits_{\mathbb{R}^n\times\mathbb{R}^n} |x-y|^2\ d\gamma(x,y) : \gamma \in \Pi(\mu,\nu)\right\} = \int\limits_{\mathbb{R}^n\times\mathbb{R}^n} |x-y|^2\ d\hat{\gamma}(x,y).$$

Proof: $c(x,y) := |x-y|^2$ is l.s.c. and non-negative. Therefore, there is a non-decreasing sequence of continuous real-valued functions $(c_k)_{k\in\mathbb{N}}$ such that $c = \lim\limits_{k\to\infty} c_k$. [7]

$$\begin{aligned}\int c\ d\gamma &= \lim_{k\to\infty}\int c_k\ d\gamma && \text{(Monotone Convergence Theorem)}\\ &= \lim_{k\to\infty}\lim_{l\to\infty}\int c_k d\gamma_l && \text{(Weak convergence)}\\ &= \lim_{l\to\infty}\lim_{k\to\infty}\int c_k d\gamma_l\\ &= \liminf_{l\to\infty}\int c\ d\gamma_l && \text{(Monotone Convergence Theorem)}\end{aligned}$$

As a result, it has been shown that $\gamma \to \int c\ d\gamma$ is lower semi-continuous. By tightness of $\Pi(\mu,\nu)$ in $\mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$ and by Prokhorov's theorem, $\Pi(\mu,\nu)$ is relatively compact. Since $\Pi(\mu,\nu)$ is also closed (Theorem 1.3), this implies it is compact. By Theorem 1.6, there exists a solution to the OT problem. □

[6] From [1], Theorem 1.5.

[7] Theorem 26.6 from the Lecture Notes of Günther Hörmann & David Langer, Analysis-Reste – Teil 2, 2008.

Definition 1.8. *The quadratic Wasserstein distance is defined as*

$$W^2(\mu, \nu) := \inf \left\{ \int_{\mathbb{R}^n \times \mathbb{R}^n} |x - y|^2 d\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\}.$$

Finding the quadratic Wasserstein distance between two probability measures is equivalent to solving the optimal transport problem with cost function $c(x, y) = |x - y|^2$.

1.2.1 Finding a direct solution

Section 1.2.1 follows page 26 of [11].

In Section 2, entropy-regularised optimal transport is introduced. The reason for introducing this adaptation to the OT problem is due to computational and statistical challenges. A linear program solving the OT problem has a complexity of about $O(n^3)$ where n is the number of points of mass to be moved in the discretised version of the problem.[8] This is too slow to be implemented in practice. The second issue is that as the dimension of the space $\mathbb{R}^d$ increases, the complexity of the problem increases exponentially, so reasonable algorithms will no longer be efficient in higher dimensions; this is called the curse of dimensionality. More precisely, if $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ is estimated empirically by some $\tilde{\gamma}$, then the expected W^2 distance between γ and $\tilde{\gamma}$ grows at $O(n^{-\frac{1}{d}})$.

[8] From [22].

2 Entropy-regularised Optimal Transport

Definition 2.1. *The entropy-regularised OT problem is to find a transport plan γ that minimises the following for some $\epsilon > 0$:*

$$W_\epsilon^2(\mu,\nu) := \left\{ \int_{\mathbb{R}^n \times \mathbb{R}^n} |x-y|^2 d\gamma(x,y) + \epsilon H_2(\gamma) : \gamma \in \Pi(\mu,\nu) \right\}$$

where $H_2(\gamma)$ is the entropy of γ with respect to the Lebesgue measure λ, defined in Definition 2.2.

The main result of Section 2 is Theorem 2.23, stating that entropy-regularised Optimal Transport Γ-converges to the OT problem as $\epsilon \to 0$. This implies that if the plan $\hat{\gamma}_\epsilon$ solves the entropy-regularised OT problem and γ is the minimising plan for the original OT problem, then $\hat{\gamma}_\epsilon \to \gamma$ weakly as $\epsilon \to 0$ and $W_\epsilon^2(\mu,\nu) \to W^2(\mu,\nu)$ (Theorem 2.8). In Sections 2.1 and 2.2, various definitions and preliminary results are presented in order to prove this main result. Section 2.3 shows the advantage of entropy-regularised OT as an approximation of the OT problem, informally outlining an exceptionally efficient method for solving the regularised problem.

2.1 Entropy

The idea behind entropy regularisation is to consider a family of transport plans chosen not from the whole set $\Pi(\mu,\nu)$ but from a restricted set of probability measures in $\Pi(\mu,\nu)$ with finite entropy. This has a significant computational advantage, outlined at Section 2.3. It is also shown at Theorem 2.7 that the addition of an entropy term makes the optimisation problem strictly convex, implying that the problem has a unique solution.

Definition 2.2. [9] *The entropy H of a probability measure π with respect to the Lebesgue measure λ is defined as*

$$H(\pi) = \mathbb{E}_\pi \left(\log \tfrac{d\pi}{d\lambda} \right)$$

if $\pi \ll \lambda$ and $H(\pi) = \infty$ otherwise.

In Section 2, the following notation from [4] is used for the entropy: $H_2(\gamma)$ denotes the entropy of $\gamma \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$ and $H_1(\mu)$ denotes the entropy of $\mu \in \mathcal{P}(\mathbb{R}^n)$. Furthermore, $H(\pi) = \int_X \log \frac{d\pi}{d\lambda} d\pi = \int_X \frac{d\pi}{d\lambda}(x) \log\left(\frac{d\pi}{d\lambda}(x)\right) d\lambda(x)$ will be simply written as $H(\pi) = \int_X \pi(x) \log(\pi(x))\, dx$. In Section 3 the reference measure μ replaces λ, so the entropy of π with respect to $\mu \in \mathcal{P}(X)$ is denoted $H(\pi|\mu)$.

The remaining definitions and results in Section 2.1 follow from pages 3-8 and 25-26 of [4].

Notation: Let $M(\mu) := \int_{\mathbb{R}^n} |x|^2 d\mu(x)$ be the second moment of the measure μ and let

$$\mathcal{P}_2^r(\mathbb{R}^n) := \{\mu \in \mathcal{P}(\mathbb{R}^n) : \mu \ll \lambda \text{ and } M(\mu) < \infty\}.$$

Theorem 2.3. *$H_2(\gamma)$ is lower semicontinuous.*

The proof of this result will follow from Lemma 2.4, Proposition 2.5 and Theorem 2.6 below.

[9] From page 6 of [15].

Lemma 2.4. *Let $\alpha \in \left(\frac{n}{n+2}, 1\right)$. There exists a finite constant $C > 0$ depending only on n and α such that for any $\mu \in P_2^r(\mathbb{R}^n)$ and for all $R \geq 0$:*

$$\int_{\mathbb{R}^n \setminus B_R} (\mu(x))^\alpha dx \leq (1 + M(\mu))^\alpha \cdot C \cdot \left(\tfrac{1}{R^2+1}\right)^{\alpha - \frac{n}{2}(1-\alpha)}$$

Proof:

$$\int_{\mathbb{R}^n \setminus B_R} (\mu(x))^\alpha \, dx = \int_{\mathbb{R}^n \setminus B_R} (\mu(x))^\alpha (1+|x|^2)^\alpha \cdot \frac{1}{(1+|x|^2)^\alpha} dx$$

$$\leq \left(\int_{\mathbb{R}^n \setminus B_R} \big(\mu(x)(1+|x|^2)\big) \, dx \right)^\alpha \cdot \left(\int_{\mathbb{R}^n \setminus B_R} \left(\frac{1}{1+|x|^2} \right)^{\frac{\alpha}{1-\alpha}} dx \right)^{1-\alpha}$$

(Hölder's inequality)

$$= \left(\int_{\mathbb{R}^n \setminus B_R} \mu(dx) + \int_{\mathbb{R}^n \setminus B_R} |x|^2 \mu(dx) \right)^\alpha \cdot \left(\int_{\mathbb{R}^n \setminus B_R} \left(\frac{1}{1+|x|^2} \right)^{\frac{\alpha}{1-\alpha}} dx \right)^{1-\alpha}$$

$$\leq (1 + M(\mu))^\alpha \cdot \left(\int_{\mathbb{R}^n \setminus B_R} \left(\frac{1}{1+|x|^2} \right)^{\frac{\alpha}{1-\alpha}} dx \right)^{1-\alpha}.$$

Recognise that two useful bounds can be applied: the total mass $\lambda(\mathbb{R}^n \setminus B_R) \leq C_0$ for some $C_0 > 0$ and on this domain, $1 + |x|^2 \geq 1 + R^2$, so $\frac{1}{1+|x|^2} \leq \frac{1}{1+R^2}$. Hence, it follows that

$$\int_{\mathbb{R}^n \setminus B_R} \left(\frac{1}{1+|x|^2} \right)^{\frac{\alpha}{1-\alpha}} dx \leq C_0 \cdot \left(\frac{1}{1+R^2} \right)^{\frac{\alpha}{1-\alpha}}$$

$$\leq C_0 \cdot \left(\frac{1}{1+R^2} \right)^{\frac{\alpha}{1-\alpha} - \frac{n}{2}} \qquad \left(\text{Since } \tfrac{\alpha}{1-\alpha} \geq \tfrac{n}{2},\ R \geq 0\right)$$

$$\implies \left(\int_{\mathbb{R}^n \setminus B_R} \left(\frac{1}{1+|x|^2} \right)^{\frac{\alpha}{1-\alpha}} dx \right)^{1-\alpha} \leq {C_0}^{1-\alpha} \cdot \left(\frac{1}{1+R^2} \right)^{\left(\frac{\alpha}{1-\alpha} - \frac{n}{2}\right)(1-\alpha)}$$

$$= C \cdot \left(\frac{1}{R^2+1} \right)^{\alpha - \frac{n}{2}(1-\alpha)}.$$

□

Proposition 2.5. *Let $u : [0, \infty) \to [0, \infty]$ be convex, l.s.c. and suppose $\lim_{s\to\infty} \frac{u(s)}{s} = \infty$. Let $(\nu_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{P}_2^r(\mathbb{R}^n)$. Let $\nu \in \mathcal{P}(\mathbb{R}^n)$ be the weak limit of this sequence. Define $U : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ by*

$$U(\mu) := \begin{cases} \int_{\mathbb{R}^n} u(\mu(x)) dx & \text{if } \mu \ll \lambda \\ \infty & \text{otherwise.} \end{cases}$$

Then, $U(\nu) \leq \liminf_{k\to\infty} U(\nu_k)$.

Proof (See Theorem 2.34 in Luigi Ambrosio, Nicola Fusco, Diego Pallara, Functions of bounded variation and free discontinuity problems, Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.)

Theorem 2.6. *Let $u : [0,\infty) \to [0,\infty]$ be convex and l.s.c. and suppose $\lim_{s\to\infty} \frac{u(s)}{s} = \infty$. Let also $u(0) = 0$ and suppose there exist constants $C > 0$, $\alpha \in \left(\frac{n}{n+2}, 1\right)$ with*

$$u(s) \geq -Cs^\alpha \ \forall s \in [0,\infty).$$

As above, define

$$U(\mu) := \begin{cases} \int_{\mathbb{R}^n} u(\mu(x))dx \ if \ \mu \ll \lambda \\ \infty \ \ otherwise. \end{cases}$$

Let $(\nu_k)_{k\in\mathbb{N}}$ be a sequence in $P_2^r(\mathbb{R}^n)$, weakly converging to some $\nu \in P_2^r(\mathbb{R}^n)$ and assume there is a $C_M < \infty$ such that $M(\nu_k), M(\nu) < C_M$. Then,

$$U(\nu) \leq \liminf_{k\to\infty} U(\nu_k).$$

Proof: First define $u^+ := \max\{u, 0\}$ and $u^- := \max\{-u, 0\}$, so $u = u^+ - u^-$. Notice that $0 \leq u^-(s) \leq Cs^\alpha$ by the lower bound assumption. For some $R > 0$ and $\mu \in P_2^r(\mathbb{R}^n)$, define

$$\begin{aligned} U_{B_R}(\mu) &:= \int_{B_R} u(\mu(x))dx \\ U^+_{\mathbb{R}^n\setminus B_R}(\mu) &:= \int_{\mathbb{R}^n\setminus B_R} u^+(\mu(x))dx \\ U^-_{\mathbb{R}^n\setminus B_R}(\mu) &:= \int_{\mathbb{R}^n\setminus B_R} u^-(\mu(x))dx \end{aligned}$$

and $U(\mu) = U^+(\mu) = \infty$ if μ is not absolutely continuous with respect to the Lebesgue measure. From Lemma 2.4, it follows that

$$0 \leq U^-_{\mathbb{R}^n\setminus B_R}(\mu) \leq (1 + M(\mu))^\alpha \cdot g(R)$$

where $g(R) \to 0$ as $R \to \infty$. We know that $U(\mu) = U_{B_R}(\mu) + U^+_{\mathbb{R}^n\setminus B_R}(\mu) - U^-_{\mathbb{R}^n\setminus B_R}(\mu)$ and by applying 2.5 to U and U^+ it follows that

$$\begin{aligned} \liminf_{k\to\infty} U(\nu_k) &= \liminf_{k\to\infty} U_{B_R}(\nu_k) + \liminf_{k\to\infty} U^+_{\mathbb{R}^n\setminus B_R}(\nu_k) - \liminf_{k\to\infty} U^-_{\mathbb{R}^n\setminus B_R}(\nu_k) \\ &\geq U_{B_R}(\nu) + U^+_{\mathbb{R}^n\setminus B_R}(\nu) - (1 + M(\mu))^\alpha \cdot g(R) \end{aligned}$$

holds for all $R > 0$. So, taking $R \to \infty$, the right hand side converges to $U(\nu)$.

$\square$

Proof of Theorem 2.3: Define $u : [0,\infty] \to [0,\infty]$ as the convex and l.s.c function $u(x) = x \log x$. $u(0) = 0$ and $\lim_{s\to\infty} \frac{u(s)}{s} = \lim_{s\to\infty} \log s = \infty$. Furthermore, for any $\alpha \in (0,1)$, there is a $C > 0$ such that $x \log x \geq -Cx^\alpha$; indeed, the minimum value for $x \log x$ is $-\frac{1}{e}$, so C can always be taken large enough such that the right hand side is lower than this value. Hence, for

$$H_2(\gamma) = \begin{cases} \int\limits_{\mathbb{R}^n\times\mathbb{R}^n} u(\gamma(x,y))d(x,y) \text{ if } \gamma \ll \lambda \\ \infty \ \text{ otherwise} \end{cases}$$

it follows from Theorem 2.6 that if γ_k converges weakly to γ, then $H_2(\gamma) \leq \liminf_{k\to\infty} H_2(\gamma_k)$. $\square$

Theorem 2.7. *There exists a unique solution in $\Pi(\mu,\nu)$ to the entropy-regularised OT problem.*

Proof (existence):
$\Pi(\mu,\nu)$ is tight (Theorem 1.4) and closed (Theorem 1.3) in the topology induced by weak convergence, so it is compact in $\mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$. By Theorem 1.7, $\gamma \mapsto \int |x-y|^2 \, d\gamma(x,y)$ is l.s.c. and by Theorem 2.6, $H_2(\gamma)$ is l.s.c. Hence, $F(\gamma) := \int |x-y|^2 \, d\gamma(x,y) + H_2(\gamma)$ is also lower semi-continuous. By compactness of $\Pi(\mu,\nu)$ and lower semicontinuity of F it follows by Theorem 1.6 that the minimising limit γ is in $\Pi(\mu,\nu)$.
Proof (uniqueness):
The minimiser is unique because $\nu \mapsto W_\epsilon^2(\mu,\nu)$ is strictly convex. Indeed, consider first $H_2(\gamma)$. The function $\gamma \mapsto \gamma \log \gamma$ is strictly convex, so for all $0 < p,q < 1$ such that $p+q=1$ and for all $\gamma_1,\gamma_2 \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$,

$$p\gamma_1 \log \gamma_1 + q\gamma_2 \log \gamma_2 > (p\gamma_1 + q\gamma_2)\log(p\gamma_1 + q\gamma_2).$$

Integrating with respect to the Lebesgue measure, it follows that

$$pH_2(\gamma_1) + qH_2(\gamma_2) > H_2(p\gamma_1 + q\gamma_2),$$

so $H_2(\gamma)$ is strictly convex.
The integral map $\gamma \mapsto \int c \, d\gamma$ is linear since $\int c \, d(\beta_1\gamma_1 + \beta_2\gamma_2) = \beta_1 \int c \, d\gamma_1 + \beta_2 \int c \, d\gamma_2$.
Now, let $\nu_1,\nu_2 \in \mathcal{P}(\mathbb{R}^n)$ have corresponding optimal maps γ_1,γ_2 and set $\nu = p\nu_1 + q\nu_2$, where $p+q=1$. Then $\gamma := p\gamma_1 + q\gamma_2 \in \Pi(\mu,\nu)$ is a plan transporting μ to ν. By the linearity of $\gamma \mapsto \int c \, d\gamma$, the strict convexity of H_2 and the optimality of $W_\epsilon^2(\mu,\nu)$,

$$W_\epsilon^2(\mu,\nu) \leq \int c d\gamma + \epsilon H_2(\gamma) < pW_\epsilon^2(\mu,\nu_1) + qW_\epsilon^2(\mu,\nu_2).$$

$\square$

2.2 Γ-convergence

Section 2.2 follows [4] unless otherwise stated.

Theorem 2.8. *Let $\hat{\gamma}_\epsilon \in \Pi(\mu,\nu)$ be the optimal transport plan for the entropy-regularised OT problem and let $\hat{\gamma} \in \Pi(\mu,\nu)$ be the optimal transport plan for the original OT problem. That is,*

$$W_\epsilon^2(\mu,\nu) = \int_{\mathbb{R}^n \times \mathbb{R}^n} |x-y|^2 d\hat{\gamma}_\epsilon(x,y) + \epsilon H_2(\hat{\gamma}_\epsilon)$$
$$W^2(\mu,\nu) = \int_{\mathbb{R}^n \times \mathbb{R}^n} |x-y|^2 d\hat{\gamma}(x,y).$$

Then the plans $\hat{\gamma}_\epsilon$ converge weakly to $\hat{\gamma}$ and the distances $W_\epsilon^2(\mu,\nu)$ converge to $W^2(\mu,\nu)$.

The proof of Theorem 2.8 follows by applying Theorem 2.12 to the functionals F_k and F defined in Definition 2.9.

Definition 2.9. *Define the functionals $F_k : \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n) \to [0,\infty]$ and $F : \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n) \to [0,\infty]$ as*

$$F_k(\gamma) = (c,\gamma) + \epsilon_k H_2(\gamma) + \iota_{\gamma \in \Pi(\mu,\nu)}$$
$$F(\gamma) = (c,\gamma) + \iota_{\gamma \in \Pi(\mu,\nu)}$$

where $(\epsilon_k)_{k\in\mathbb{N}}$ is a sequence of numbers decreasing to zero and $\iota_{\gamma\in\Pi(\mu,\nu)} = \begin{cases} 0 \text{ if } \gamma \in \Pi(\mu,\nu) \\ \infty \text{ if } \gamma \notin \Pi(\mu,\nu). \end{cases}$

Definition 2.10. *The sequence $(F_k)_{k\in\mathbb{N}}$ is said to* Γ*-converge to F with respect to the topology induced by weak convergence if the following two conditions hold for all $\gamma \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$.*
Liminf Condition – *For any sequence $(\gamma_k)_{k\in\mathbb{N}}$ in $\mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$ such that $\gamma_k \to \gamma$ weakly,*

$$F(\gamma) \leq \liminf_{k\to\infty} F_k(\gamma_k).$$

Limsup Condition – *There exists a sequence $(\gamma_k)_{k\in\mathbb{N}}$ in $\mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$, called a recovery sequence, such that $\gamma_k \to \gamma$ weakly and*

$$F(\gamma) \geq \limsup_{k\to\infty} F_k(\gamma_k).$$

Definition 2.11. *A family of functionals $(F_k)_{k\in\mathbb{N}}$ on X is said to be equi-coercive if for every $\alpha \in \mathbb{R}$, there is a compact set $K_\alpha \subseteq X$ such that the sets $\{x \in X : F_n(x) \leq \alpha\} \subseteq K_\alpha$ for all $n \in \mathbb{N}$.*

The following theorem shows that Γ-convergence of the functionals F_k to F and equi-coerciveness of the sequence $(F_k)_{k\in\mathbb{N}}$ implies that $W^2_\epsilon(\mu,\nu) \to W^2(\mu,\nu)$ as $\epsilon \to 0$ and that the optimal plan solving the entropy-regularised OT problem converges to the optimal plan solving the OT problem.

Theorem 2.12. ***(Fundamental Theorem of*** Γ***-convergence)*** [10] *Let (X,d) be a metric space. Let (F_ϵ) be an equi-coercive sequence of functions on X and let $F = \Gamma - \lim_{\epsilon\to 0} F_\epsilon$. Then there exists a solution*

$$\min_{x\in X} F(x) = \lim_{\epsilon\to 0} \inf_{x\in X} F_\epsilon(x).$$

Moreover, if (x_ϵ) is a precompact sequence such that $\lim_{\epsilon\to 0} F_\epsilon(x_\epsilon) = \lim_{\epsilon\to 0} \inf_{x\in X} F_\epsilon(x)$, then every limit of a subsequence of (x_ϵ) is a minimum point for F.

Sketch of Proof: Equicoercivity of (F_ϵ) and the Liminf condition imply that: if x_ϵ is a minimising sequence for F_ϵ (so $F_\epsilon(x_\epsilon) \leq \inf F_\epsilon + o(1)$) and, up to extraction of a subsequence, $x_\epsilon \to x_0 \in X$, then

$$\inf F \leq F(x_0) \leq \liminf_{\epsilon\to 0} F_\epsilon(x_\epsilon) = \liminf_{\epsilon\to 0} \inf F_\epsilon. \tag{1}$$

The Limsup condition implies that a recovery sequence $\bar{x}_\epsilon \to x_0$ exists such that $F(x_0) \geq \limsup_{\epsilon\to 0} F_\epsilon(\bar{x}_\epsilon) \geq \limsup_{\epsilon\to 0} \inf F_\epsilon$. Since this recovery sequence exists $\forall x \in X$,

$$\inf F \geq \limsup_{\epsilon\to 0} \inf F_\epsilon. \tag{2}$$

From (1) and (2),

$$\lim_{\epsilon\to 0} \inf F_\epsilon = \min_{x\in X} F(x).$$

Moreover, every limit point of a minimising sequence x_ϵ is a minimum point for F. $\square$

[10] From [3] Theorem 2.10, p 14.

2.2.1 Preliminary details

In this subsection, a number of results are presented, all of which will be used in the Γ-convergence proofs in Section 2.2.2.

Proposition 2.13. [11] *The Wasserstein distance W^2 metrizes weak convergence. That is, $W^2(\gamma_k, \gamma) \to 0$ if and only if γ_k converges weakly to γ and $M_2(\gamma_k) \to M_2(\gamma)$. Furthermore, $W^2(\gamma_k, \gamma) \to 0$ if and only if $\int \phi(x) d\gamma_k(x) \to \int \phi(x) d\gamma(x)$ for all continuous functions ϕ with $|\phi(x)| \leq C(1 + |x|^2), C \in \mathbb{R}$.*

Proposition 2.14. *There exist constants $C > 0$ and $\alpha \in (0, 1)$ such that for every $\mu \in \mathcal{P}_2^r(\mathbb{R}^n)$, $H_1(\mu) \geq -C(M(\mu) + 1)^\alpha$.*

Proof: Recall from the proof of Theorem 2.3 that for any $\alpha \in (0, 1)$, there is a $k > 0$ such that $\forall x \in [0, 1]$, $x \log x \geq -kx^\alpha$. Hence,

$$\begin{aligned} H_1(\mu) := \int\limits_{\mathbb{R}^n} \mu(x) \log(\mu(x)) dx &\geq -k \cdot \int\limits_{\mathbb{R}^n} (\mu(x))^\alpha \, dx \\ &\geq -C \cdot (1 + M(\mu))^\alpha \end{aligned}$$

for some new constant $C > 0$, by Lemma 2.4. □

Corollary 2.15. *If $M(\gamma) < \infty$, then there exists a constant $C < \infty$ such that*

$$\int\limits_{\mathbb{R}^n \times \mathbb{R}^n} |\min\{\gamma(x, y) \log(\gamma(x, y)), 0\}| \, d(x, y) > -C.$$

Proof: Suppose that $\gamma(x, y) \log(\gamma(x, y)) < 0$. Then

$$\begin{aligned} \int\limits_{\mathbb{R}^n \times \mathbb{R}^n} |\min\{\gamma(x, y) \log(\gamma(x, y)), 0\}| d(x, y) &= \int\limits_{\mathbb{R}^n \times \mathbb{R}^n} |\gamma(x, y) \log(\gamma(x, y))| d(x, y) \\ &\geq -\hat{C} \cdot (1 + M(\gamma))^\alpha, \end{aligned}$$

by Proposition 2.14. Since it is assumed that $M(\gamma) = A < \infty$, it follows that

$$\int\limits_{\mathbb{R}^n \times \mathbb{R}^n} |\min\{\gamma(x, y) \log(\gamma(x, y)), 0\}| d(x, y) > -C,$$

for some $C < \infty$. □

Definition 2.16. ***(Block Approximation)*** *Let $\mu, \nu \in \mathcal{P}_2^r(\mathbb{R}^n)$ both have finite entropy. Suppose that $\gamma \in \Pi(\mu, \nu)$. Given $k = (k_1, k_2, \ldots, k_n) \in \mathbb{Z}^n$ we define $Q_k := [k_1, k_1 - 1) \times \ldots \times [k_n, k_n - 1) \subset \mathbb{R}^n$ as the unit cube at some point k and for $l > 0$, denote the scaled block $Q_k^l = l \cdot Q_k$. The* block approximation *of γ at scale l is*

$$\gamma_l := \sum_{(j,k) \in (\mathbb{Z}^n)^2} \gamma(Q_j^l \times Q_k^l)(\mu_{j,l} \otimes \nu_{k,l}),$$

an approximation of γ by pieces of the product measure $(\mu_{j,l} \otimes \nu_{k,l})$ such that the mass of γ_l and γ is equal on each block, where for every Borel set $\sigma \subset \mathbb{R}^n$,

$$\mu_{j,l}(\sigma) := \begin{cases} \frac{\mu(\sigma \cap Q_j^l)}{\mu(Q_j^l)} & \text{if } \mu(Q_j^l) > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \nu_{k,l}(\sigma) := \begin{cases} \frac{\nu(\sigma \cap Q_k^l)}{\nu(Q_k^l)} & \text{if } \nu(Q_k^l) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

[11] From [23] Definition 6.7 and Theorem 6.8.

γ_l is a probability measure on $\mathbb{R}^n \times \mathbb{R}^n$ with density

$$\gamma_l(x,y) = \begin{cases} \gamma(Q_j^l \times Q_k^l) \frac{\mu(x)\nu(y)}{\mu(Q_j^l)\nu(Q_k^l)} & \text{if } \mu(Q_j^l) > 0 \text{ and } \nu(Q_k^l) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The choice of $(j,k) \in (\mathbb{Z}^n)^2$ is uniquely determined by $(x,y) \in (Q_j^l \times Q_k^l)$.

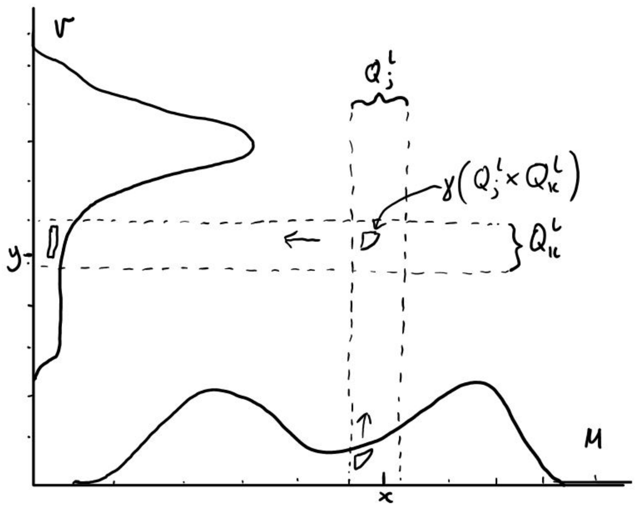

Figure 3: The plan γ_l measures the amount of mass being moved from block $Q_j^l \ni x$ to $Q_k^l \ni y$ by the plan γ. This is multiplied by the probabilities of $x \in Q_j^l$ and $y \in Q_k^l$, making γ_l a probability measure.

Proposition 2.17. *For $\gamma \in \Pi(\mu,\nu)$, the block approximation γ_l is also in $\Pi(\mu,\nu)$, for any $l \in \mathbb{N}$.*

Proof: It needs to be shown that the probability measure γ_l has the marginal distributions μ and ν. From Definition 2.16 of the block approximation, it follows that

$$(\mu_{j,l} \otimes \nu_{k,l})(\mathbb{R}^n \times A) = \nu_{k,l}(A) \text{ if } \mu(Q_j^l) > 0 \text{ and 0 otherwise, for any } A \subset \mathbb{R}^n. \quad (1)$$

From the definition of the transport plan γ and the blocks Q_j^l, we have that

$$\gamma(Q_j^l \times Q_k^l) = 0 \textit{ if } \mu(Q_j^l) = 0 \quad (2)$$

$$\sum_{j \in \mathbb{Z}^n} \gamma(Q_j^l \times Q_k^l) = \gamma(\mathbb{R}^n \times Q_k^l) = \nu(Q_k^l). \quad (3)$$

Therefore, for any $A \subset \mathbb{R}^n$,

$$\begin{aligned} \gamma_l(\mathbb{R}^n \times A) &= \sum_{j,k \in \mathbb{Z}^n} \gamma(Q_j^l \times Q_k^l)(\mu_{j,l} \otimes \nu_{k,l})(\mathbb{R}^n \times A) \\ &= \sum_{(j,k):\mu(Q_j^l)>0} \gamma(Q_j^l \times Q_k^l)\nu_{k,l}(A) && \text{(By (1))} \\ &= \sum_{k \in \mathbb{Z}^n} \nu_{k,l}(A) \cdot \sum_{j \in \mathbb{Z}^n} \gamma(Q_j^l \times Q_k^l) && \text{(By (2))} \\ &= \nu(A). && \text{(By (3))} \end{aligned}$$

It can be shown analogously that $\gamma_l(A \times \mathbb{R}^n) = \mu(A)$, so $\gamma_l \in \Pi(\mu, \nu)$. $\square$

Lemma 2.18. *Let $\{Q_i\}_{i\in I}$ be a countable partition of $\mathbb{R}^n$ and suppose that for any $i \in I$, $diam(Q_i) \leq C < \infty$. If $\mu, \nu \in \mathcal{P}_2^r(\mathbb{R}^n)$ are such that $\mu(Q_i) = \nu(Q_i)$ $\forall i \in I$, then $W^2(\mu, \nu) \leq C^2$.*

Proof: Let $\tilde{I}$ be the subset of I where $\mu(Q_i) = \nu(Q_i) > 0$. Define the probability measures μ_i and ν_i, $\forall A \in \mathbb{R}^n$, as

$$\mu_i(A) := \frac{\mu(A \cap Q_i)}{\mu(Q_i)},\ \nu_i(A) := \frac{\nu(Q_i \cap A)}{\nu(Q_i)}.$$

For each $i \in \tilde{I}$, let $\gamma_i \in \Pi(\mu_i, \nu_i)$ be a transport plan between these two measures. By the assumption on the diameter of Q_i and since γ_i is a probability measure, it follows that

$$\begin{aligned}
(c, \gamma_i) &:= \int\limits_{\mathbb{R}^n \times \mathbb{R}^n} |x-y|^2 d\gamma_i(x,y) \\
&= \int\limits_{Q_i \times Q_i} |x-y|^2 d\gamma_i(x,y) \\
&\leq \int\limits_{Q_i \times Q_i} C^2 d\gamma_i(x,y) \\
&= C^2.
\end{aligned}$$

Since $\{Q_i\}_{i\in I}$ partitions $\mathbb{R}^n$, $\gamma \in \Pi(\mu, \nu)$ can be defined as the sum of the transport plans on each Q_i scaled by $\mu(Q_i)$, i.e. $\gamma := \sum\limits_{i\in\tilde{I}} \mu(Q_i)\gamma_i \in \Pi(\mu, \nu)$. Since W^2 is defined as the infimum over all transport plans, it follows that

$$W^2(\mu, \nu) \leq (c, \gamma) \leq \sum_{i\in\tilde{I}} \mu(Q_i)(c, \gamma_i) \leq \sum_{i\in\tilde{I}} \mu(Q_i) \cdot C^2 = C^2.$$

$\square$

Corollary 2.19. *$W^2(\gamma, \gamma_l) \leq 4nl^2$ and, as a consequence of Proposition 2.13, γ_l converges weakly to γ as $l \to 0^+$.*

Proof: It can be shown that the plan γ and the block approximation γ_l satisfy the conditions of Lemma 2.18. Each scaled n-cube Q_k^l has diameter equal to the distance between the vertices $(0, 0, \ldots, 0) \in \mathbb{R}^n$ and $(l, l, \ldots, l) \in \mathbb{R}^n$. Therefore,

$$\text{diam}(Q_k^l) = \sqrt{(l-0)^2 + \cdots + (l-0)^2} = l\sqrt{n}.$$

So it follows that $\text{diam}(Q_j^l \times Q_k^l) \leq 2l\sqrt{n}$ and by definition $\{Q_j^l \times Q_k^l\}_{j,k\in\mathbb{Z}^n}$ forms a partition of $\mathbb{R}^n \times \mathbb{R}^n$ such that $\gamma_l(Q_j^l \times Q_k^l) = \gamma(Q_j^l \times Q_k^l)$ on each block. By Lemma 2.18, this implies that $W^2(\gamma, \gamma_l) \leq 4nl^2$. $\square$

Corollary 2.20. *For the block approximations $(\gamma_l)_{l\in\mathbb{N}} \in \Pi(\mu, \nu)$,*

$$\lim_{l\to 0^+} (c, \gamma_l) = (c, \gamma).$$

Proof: The cost function $c(x, y) = |x-y|^2 \leq 2(|x|^2 + |y|^2)$. Therefore, since $W^2(\gamma, \gamma_l) \to 0$ by Corollary 2.19, the result follows from Proposition 2.13. $\square$

Proposition 2.21. *There are constants $C > 0$ and $\alpha \in (0,1)$ such that the entropy of the block approximation γ_l of $\gamma \in \Pi(\mu,\nu)$ is bounded above by*

$$H_1(\mu) + H_1(\nu) + C\left((2M(\mu) + 8nl^2 + 1)^\alpha + (2M(\nu) + 8nl^2 + 1)^\alpha\right) - 2n\log(l)$$

Proof: Consider the density $\mu_l(x) = \mu(Q_j^l) \cdot l^{-n}$, where j is uniquely determined by the partition member Q_j^l containing x. Then,

$$\mu_l(Q_j^l) = \int_{Q_j^l} \mu_l(x)dx = \int_{Q_j^l} \mu(Q_j^l) \cdot l^{-n} dx = \mu(Q_j^l) \cdot l^{-n}\lambda(Q_j^l) = \mu(Q_j^l)$$

and so

$$\sum_{j \in \mathbb{Z}^n} \mu(Q_j^l)\log(\mu(Q_j^l)) = \int_{\mathbb{R}^n} \mu_l(x)\log(\mu_l(x) \cdot l^n)dx = H_1(\mu_l) + n\log(l) \geq -C(M(\mu_l)+1)^\alpha + n\log l$$

by Proposition 2.14. Moreover, if ρ_l is the optimal map in $\Pi(\mu_l, \mu)$, then

$$\begin{aligned}
M(\mu_l) = \int_{\mathbb{R}^n} |x|^2 d\mu_l(x) &= \int_{\mathbb{R}^n \times \mathbb{R}^n} |x|^2 d\rho_l(x,y) \\
&\leq \int_{\mathbb{R}^n \times \mathbb{R}^n} \left(2|y|^2 + 2|x-y|^2\right) d\rho_l(x,y) \\
&= 2 \cdot \int_{\mathbb{R}^n \times \mathbb{R}^n} |y|^2 d\rho_l(x,y) + 2 \cdot \int_{\mathbb{R}^n \times \mathbb{R}^n} |x-y|^2 d\rho_l(x,y) \\
&= 2M(\mu) + 2W^2(\mu_l, \mu)
\end{aligned}$$

and by Corollary 2.19, $W^2(\mu_l, \mu) \leq 4nl^2$. Therefore,

$$\sum_{j \in \mathbb{Z}^n} \mu(Q_j^l)\log(\mu(Q_j^l)) \geq -C(2M(\mu) + 8nl^2 + 1)^\alpha + n\log l.$$

Then, substituting the density γ_l into the definition of entropy and using this result, it follows that

$$\begin{aligned}
H_2(\gamma_l) &= \sum_{j,k} \gamma(Q_j^l \times Q_k^l) \cdot \int_{Q_j^l \times Q_k^l} \frac{\mu(x)\nu(y)}{\mu(Q_j^l)\nu(Q_k^l)} \left[\log\left(\gamma(Q_j^l \times Q_k^l)\right) + \log\left(\frac{\mu(x)}{\mu(Q_j^l)}\right) + \log\left(\frac{\nu(y)}{\nu(Q_k^l)}\right)\right] dxdy \\
&= \sum_{j,k} \gamma(Q_j^l \times Q_k^l) \left[\log\left(\gamma(Q_j^l \times Q_k^l)\right) + \int_{Q_j^l} \frac{\mu(x)\log(\mu(x))}{\mu(Q_j^l)} dx - \log(\mu(Q_j^l)) + \int_{Q_k^l} \frac{\nu(y)\log(\nu(y))}{\nu(Q_k^l)} dy - \log(\nu(Q_k^l))\right] \\
&\leq \sum_{j,k} \gamma(Q_j^l \times Q_k^l) \left[\int_{Q_j^l} \mu(x)\log(\mu(x))dx + \int_{Q_k^l} \nu(y)\log(\nu(y))dy - \mu(Q_j^l)\log(\mu(Q_j^l)) - \nu(Q_k^l)\log(\nu(Q_k^l))\right] \\
&\leq H_1(\mu) + H_1(\nu) - \sum_j \mu(Q_j^l)\log\left(\mu(Q_j^l)\right) - \sum_k \nu(Q_k^l)\log\left(\nu(Q_k^l)\right) \\
&\leq H_1(\mu) + H_1(\nu) + C(2M(\mu) + 8nl^2 + 1)^\alpha + C(2M(\nu) + 8nl^2 + 1)^\alpha - 2n\log l.
\end{aligned}$$

□

2.2.2 Proof of main result

Theorem 2.22. *The sequence of functionals $(F_k)_{k\in\mathbb{N}}$ defined at Definition 2.9 is equicoercive.*

Proof: Equicoercivity of $(F_k)_{k\in\mathbb{N}}$ follows from the tightness of $\Pi(\mu,\nu)$ (Theorem 1.4). Indeed, for equicoercivity it needs to be shown that for each $k \in \mathbb{N}, \forall\, \alpha \in \mathbb{R}$, the set

$$A := \{\gamma \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n) : F_k(\gamma) \leq \alpha\} = \{\gamma \in \Pi(\mu,\nu) : F_k(\gamma) \leq \alpha\} \subseteq K_\alpha$$

for some compact $K_\alpha \subset \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$. Since $\Pi(\mu,\nu)$ is tight, it follows by Prokhorov's Theorem that $\Pi(\mu,\nu)$ is relatively compact. Since $\Pi(\mu,\nu)$ is also closed (Theorem 1.3), this implies that it is compact. The set A is contained in the compact space $\Pi(\mu,\nu)$ so any closed subset containing A is compact. □

Theorem 2.23. *The sequence of functionals $(F_k)_{k\in\mathbb{N}}$ Γ-converges to F (defined at 2.9), under the assumption that μ and ν have finite entropy and finite second moments.*

Proof (Liminf condition): It needs to be shown that for any sequence $(\gamma_k)_{k\in\mathbb{N}}$ in $\mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$ such that $\gamma_k \to \gamma$ weakly, $F(\gamma) \leq \liminf\limits_{k\to\infty} F_k(\gamma_k)$.
If $\gamma \notin \Pi(\mu,\nu)$ then this holds: since $\Pi(\mu,\nu)$ is closed, $\gamma_k \notin \Pi(\mu,\nu)$ for k big enough, so $F(\gamma) \leq \liminf\limits_{k\to\infty} F_k(\gamma_k) = \infty$. Similarly, it suffices to consider subsequences of $(\gamma_k)_{k\in\mathbb{N}}$ that are absolutely continuous with respect to the Lebesgue measure; otherwise, $H_2(\gamma) \to \infty$ by definition.
Since $M_2(\gamma_k) = \int\limits_{\mathbb{R}^n} |x|^2 d\mu(x) + \int\limits_{\mathbb{R}^n} |y|^2 d\nu(y) = M(\mu) + M(\nu) < \infty$, by Corollary 2.15 there exists a constant $C < \infty$ such that

$$\int\limits_{\mathbb{R}^n\times\mathbb{R}^n} |\min\{\gamma_k(x,y)\log(\gamma_k(x,y)), 0\}| d(x,y) > -C.$$

Hence, taking k large enough such that $\epsilon_k < \frac{1}{C}$, it follows that $\liminf\limits_{k\to\infty} \epsilon_k H_2(\gamma_k) \geq 0$.
By the lower semi-continuity of $\gamma \to (c,\gamma)$ (Proposition 1.7),

$$\liminf_{k\to\infty} F_k(\gamma_k) = \liminf_{k\to\infty} (c,\gamma_k) + \liminf_{k\to\infty} \epsilon_k H_2(\gamma_k) \geq (c,\gamma).$$

□

Proof (Limsup condition): It remains to be proved that for every $\gamma \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$, *there exists* a sequence $(\gamma_k)_{k\in\mathbb{N}}$ that converges weakly to γ and such that $F(\gamma) \geq \limsup\limits_{k\to\infty} F_k(\gamma_k)$. The proof of this result relies on the construction of the sequence of measures above called the block approximation of γ, along with the upper bound on H_2 from Proposition 2.21.

1. For any non-negative sequence $(l_k)_{k\in\mathbb{N}}$ converging to zero, the block approximation (γ_{l_k}) converges weakly to γ as $k \to \infty$. (Proved in Corollary 2.19)

2. For $\gamma \notin \Pi(\mu,\nu)$, the condition holds trivially since $F(\gamma) = \infty$. Consider then only $\gamma \in \Pi(\mu,\nu)$. Then, by Proposition 2.17, the sequence of measures $(\gamma_{l_k})_{k\in\mathbb{N}}$ given by the block approximation of γ is also in $\Pi(\mu,\nu)$.

3. Then, by the upper bound result on H_2 from Proposition 2.21,

$$\begin{aligned} H_2(\gamma_{l_k}) &\leq H_1(\mu) + H_1(\nu) + C\left((2M(\mu) + 8nl_k^2 + 1)^\alpha + (2M(\nu) + 8nl_k^2 + 1)^\alpha\right) - 2n\log(l_k) \\ \epsilon_k H_2(\gamma_{l_k}) &\leq \epsilon_k H_1(\mu) + \epsilon_k H_1(\nu) + \epsilon_k C\left((2M(\mu) + 8nl_k^2 + 1)^\alpha + (2M(\nu) + 8nl_k^2 + 1)^\alpha\right) - 2n\epsilon_k\log(l_k) \end{aligned}$$

   $\limsup\limits_{k\to\infty} \epsilon_k H_2(\gamma_{l_k}) \leq 0 + 0 + 0 + 0 = 0.$

   This last inequality holds since $H_1(\mu), H_1(\nu), M(\mu), M(\nu) < \infty$, by assumption.

4. The Limsup Condition follows because $\lim_{l \to 0^+} (c, \gamma_l) = (c, \gamma)$ (Corollary 2.20). Hence, for the block approximation (γ_l) of each γ,

$$F(\gamma) \geq \limsup_{k \to \infty} F_k(\gamma_k).$$

□

2.3 Profiting from Γ-convergence: the Sinkhorn Algorithm

Practically, the biggest advantage of entropy-regularised OT is that there exists an efficient algorithm for solving the problem, called the Sinkhorn Algorithm. The intention of Section 2.3 is to provide an idea of why this is the case, not to give an extensive review of the algorithm and its efficiency. A detailed account can be found in [7] and in [22], which this discussion follows. To provide an overview of the Sinkhorn Algorithm, it is necessary to briefly introduce the discretised version of entropy-regularised OT to establish a framework where an algorithm can be applied.

Let $U(A, B)$ denote the set of matrices representing the possible transport plans from the vector $a \in \mathbb{R}^{m_1}$ to $b \in \mathbb{R}^{m_2}$, where $a^T \mathbb{1}_{m_1} = 1$ and $b^T \mathbb{1}_{m_2} = 1$. That is,

$$U(a, b) := \{P \in \mathbb{R}^{m_1 \times m_2} : P^T \cdot \mathbb{1}_{m_1} = b, P \cdot \mathbb{1}_{m_2} = a\}.$$

$U(a, b)$ is the discrete representation of the set of transport plans $\Pi(\mu, \nu)$. Given a cost matrix $C \in \mathbb{R}^{m_2} \times \mathbb{R}^{m_1}$ and for $k > 0$, the discretized version of the entropy-regularised OT problem is to find

$$W_k := \min_{P \in U(a,b)} \left\{ \sum_{\substack{1 \leq i \leq m_1 \\ 1 \leq j \leq m_2}} P_{i,j} C_{j,i} + \frac{1}{k} \sum_{\substack{1 \leq i \leq m_1 \\ 1 \leq j \leq m_2}} P_{i,j} \log(P_{i,j}) \right\}.$$

The function $p \mapsto pc + \frac{1}{k} p \log p$ is strictly convex. Therefore, there is a unique solution to W_k. Following duality theory as explained in Sections 5.1 and 5.2 of [2], the constraints that $P^T \cdot \mathbb{1}_{m_1} - b = 0$ and $P \cdot \mathbb{1}_{m_2} - a = 0$ can be taken into the optimisation problem by forming the Lagrangian Λ, given by

$$\Lambda(P, \alpha, \beta) = \frac{1}{k} \sum_{i,j} P_{i,j} \log(P_{i,j}) + \sum_{i,j} P_{i,j} C_{j,i} + \alpha(P^T \cdot \mathbb{1}_{m_1} - b) + \beta(P \cdot \mathbb{1}_{m_2} - a).$$

The problem is then to find the unique matrix P that satisfies

$$\min\{\Lambda(P, \alpha, \beta) : P \in \mathbb{R}^{m_1 \times m_2}, \alpha \geq 0, \beta \geq 0\}.$$

This can be solved by taking the derivative of Λ with respect to P and setting it equal to zero:

$$\frac{d\Lambda}{dP_{i,j}} = \frac{1}{k} \log(P_{i,j}) + 1 + C_{j,i} + \alpha_i + \beta_j = 0.$$

This implies that, for some new variables α and β, the solution of the optimisation problem is of the form

$$P_{i,j} = \alpha_i \beta_j e^{-kC_{j,i}} \ .$$

Now, the Sinkhorn Algorithm just needs to find the optimal vectors $\alpha \in \mathbb{R}^{m_1}$ and $\beta \in \mathbb{R}^{m_2}$.

Notice that

$$b_i = \sum_{1 \leq j \leq m_2} P_{i,j} = \alpha_i \sum_{1 \leq j \leq m_2} \beta_j e^{-kC_{j,i}}.$$

So,

$$\alpha_i = b_i \left(\sum_{1 \leq j \leq m_2} \beta_j e^{-kC_{j,i}} \right)^{-1} \text{ and similarly } \beta_j = a_j \left(\sum_{1 \leq i \leq m_1} \alpha_i e^{-kC_{j,i}} \right)^{-1}.$$

The advantage of adding the entropy term to the OT problem is that by iterating over t after setting

$$\alpha_i^{(t+1)} = b_i \left(\sum_{1 \leq j \leq m_2} \beta_j^{(t)} e^{-kC_{j,i}} \right)^{-1} \text{ and then } \beta_j^{(t+1)} = a_j \left(\sum_{1 \leq i \leq m_1} \alpha_i^{(t)} e^{-kC_{j,i}} \right)^{-1},$$

this converges to a solution far quicker than finding a direct solution to the optimal transport problem. [12]

[12] From [7].

3 Schrödinger's Problem

The goal of Section 3 is to introduce the Schrödinger problem and to link it to the entropy-regularised Optimal Transport problem. In Section 3.1, Large Deviation Theory is introduced and utilised to formulate the dynamic Schrödinger Problem in terms of modern probability theory in Definition 3.7. In Section 3.3, the static version of Schrödinger's Problem is defined and Schilder's Theorem is used to establish the relationship between the static and dynamic problems for a special case when $Z_t^\epsilon = \epsilon B_t$. This static version of the Schrödinger Problem is equivalent to the entropy-regularised Optimal Transport problem. Schilder's Theorem is presented and proved in Section 3.2. The arguments in Section 3 follow [13], [14] and [5], with specific proofs and subsections from other sources referenced where relevant.

Let X be a polish space and let $\Omega = C([0,1], X)$ be the set of all continuous paths from $[0,1]$ to X. Consider the trajectories of a large number of individual particles in X for $t \in [0,1]$ as a collection of i.i.d. random variables $(Z_i) \in \Omega$. Suppose that the initial distribution of the particles is approximately distributed around the probability measure $\pi_0 \in \mathcal{P}(X)$. The law of large numbers says that at $t = 1$, based on the law of the random variables (Z_i), the configuration of particles should be close to the expected distribution with high probability.

Suppose that at $t = 1$ the configuration is observed to be some $\pi_1 \in \mathcal{P}(X)$ far away from the expected configuration. Schrödinger's question is the following:

> *Conditionally on this rare event, what is the most likely path of the whole system between $t = 0$ and $t = 1$?*

3.1 Schrödinger's Question

Schrödinger's question can be formalised to obtain what will be called the dynamic Schrödinger problem.

Definition 3.1. *A path $x = (x_1, x_2, \dots, x_n) \in X^n$ is categorised into a type $\nu \in \mathcal{P}(X)$ based on its empirical measure, given by*

$$L_x^n(\cdot) = \tfrac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i = \cdot\}}.$$

This is what will be meant when considering all paths in X^n of type ν. The measure $\nu \in \mathcal{P}(X)$ takes the value $\nu(x_j) = \frac{k}{n}$ where k is the number of times the point $x_j \in X$ appears in the path vector $x \in X^n$. That is, $n\nu(x_j)$ is the number of entries in the path vector x that are equal to x_j.
Let $V_n(\nu)$ denote the set of all the discrete paths in X^n of type ν.
Let $\mathcal{L}_n := \{\nu \in \mathcal{P}(X) : \nu$ is the type of some path in $X^n\}$.

The empirical measure $L^n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}$ associated to the sequence of random variables $\{Z_i\}_{i=1}^n$ is a random element of the set $\mathcal{L}_n$.

Take the initial and final distributions $\pi_0, \pi_1 \in \mathcal{P}(X)$ and assume that $M(\pi_0), M(\pi_1) < \infty$. Consider the δ-neighbourhoods in $\mathcal{P}(X)$, $C_0^\delta \ni \pi_0$ and $C_1^\delta \ni \pi_1$ (with respect to the Wasserstein metric, so $W^2(\mu, \pi_0) < \delta \ \forall \mu \in C_0^\delta$) . The purpose of introducing these δ-neighbourhoods is to avoid dividing by 0 in the next expression.

Definition 3.2. [13] *A formulation of Schrödinger's question in modern probability language is the following. Find the measurable sets $A \in \mathcal{P}(X)$ such that the conditional probability*

$$\mathbb{P}(L^n \in A : L_0^n \in C_0^\delta \ , \ L_1^n \in C_1^\delta) := \frac{\mathbb{P}(L^n \in A \ \cap \ L_0^n \in C_0^\delta \ \cap \ L_1^n \in C_1^\delta)}{\mathbb{P}(L_0^n \in C_0^\delta \ \cap \ L_1^n \in C_1^\delta)}$$

is maximal when $n \to \infty$ and $\delta \to 0$.

3.1.1 Large Deviation Theory

Large Deviation Theory is fundamental in deriving the dynamic Schrödinger Problem from Definition 3.2. Sanov's Theorem is the key result, presented here as in [13].

The Large Deviation Principle (LDP) characterizes the limiting behaviour as $\epsilon \to 0$ of a family of probability measures $\{\gamma_\epsilon\}$ on X in terms of a rate function. This characterisation is via asymptotic upper and lower exponential bounds on the values that γ_ϵ assigns to measurable subsets of X.[14]

Definition 3.3. *If X is a polish space, we say that a sequence of probability measures $(\gamma_k)_{k\in\mathbb{N}} \in \mathcal{P}(X)$ satisfies the Large Deviation Principle with scale n and rate function I if for each measurable set $A \subset X$,*

$$-\inf_{x\in intA} I(x) \le \liminf_{n\to\infty} \tfrac{1}{n} \log \gamma_n(A) \le \limsup_{n\to\infty} \tfrac{1}{n} \log \gamma_n(A) \le -\inf_{x\in clA} I(x)$$

where the rate function $I : X \to [0, \infty]$ is lower semi-continuous.

If γ_n satisfies the LDP and the measurable set $A \subset X$ is such that $\inf_{x\in intA} I(x) = \inf_{x\in clA} I(x)$ then these are all equalities. A set satisfying this property is called an I continuity set. It will be assumed that A is an I continuity set and the preferred notation for the LDP is to write

$$\gamma_n(A) \simeq \exp\left(-n \cdot \inf_{x\in A} I(x)\right), \text{ for large } n.$$

Theorem 3.4. ***(Sanov's Theorem)****: Let $Z_1, Z_2, \ldots$ be a sequence of i.i.d. X-valued random variables with law μ. Let $\mathbb{P}$ denote the probability measure associated to the sequence of random variables $\{Z_i\}$. Then, for the empirical measure $L^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}$,*

$$\mathbb{P}(L^n \in \cdot)$$

satisfies the LDP in $\mathcal{P}(X)$ with scale n and rate function $H(\cdot|\mu) : \mathcal{P}(X) \to [0, \infty]$, the relative entropy with respect to the reference measure μ, defined in Definition 2.2.

The proof of Sanov's Theorem follows [16] with some assistance from [8]. The theorem is proved for discrete X. Lemma 3.5 and Lemma 3.6 are needed first.

Lemma 3.5. *If $x \in V_n(\nu)$, then for i.i.d. random variables $Z_1, Z_2, \ldots Z_n$ with law $\mu \in \mathcal{P}(X)$,*

$$\mathbb{P}((Z_1, Z_2, \ldots, Z_n) = x) = e^{-n(H(\nu|\mu) - H(\nu|\#))}$$

where $\# \in \mathcal{P}(X)$ is the counting measure.

[13] From [13].
[14] From [8].

Proof:

$$
\begin{aligned}
\mathbb{P}((Z_1, Z_2, ..., Z_n) = x) &= \prod_{i=1}^{n} \mathbb{P}(Z_i = x_i) && \text{(by independence)} \\
&= \prod_{i=1}^{n} \mu(x_i) && \text{(μ is the law of Z)} \\
&= \prod_{x \in X} \mu(x)^{n\nu(x)} && \text{(Reordering, definition of $\nu(x)$)} \\
\log\left[\mathbb{P}((Z_1, Z_2, ..., Z_n) = x)\right] &= \log\left(\prod_{x \in X} \mu(x)^{n\nu(x)}\right) \\
&= \sum_{x \in X} n\nu(x) \log \mu(x) \\
\implies \mathbb{P}\left((Z_1, Z_2, ..., Z_n) = x\right) &= \exp\left(n \sum_{x \in X} \nu(x) \log \mu(x)\right).
\end{aligned}
$$

Note that

$$
\begin{aligned}
H(\nu|\mu) - H(\nu|\#) &= \sum_{x \in X} \nu(x) \log\left(\frac{\nu(x)}{\mu(x)}\right) - \sum_{x \in X} \nu(x) \log(\nu(x)) \\
&= -\sum_{x \in X} \nu(x) \log \mu(x).
\end{aligned}
$$

So indeed it follows that

$$
\mathbb{P}\left((Z_1, Z_2, ..., Z_n) = x\right) = \exp\left[-n\left(H(\nu|\mu) - H(\nu|\#)\right)\right].
$$

$\square$

Lemma 3.6. *There exists a function f such that $\frac{1}{n}\log f(n) \to 0$ as $n \to \infty$ and such that $\forall\ \nu \in \mathcal{L}_n$,*

$$
e^{-nH(\nu|\#))} \leq |V_n(\nu)| \leq f(n)e^{-nH(\nu|\#)}.
$$

Proof: A combinatorical argument will be used to calculate $|V_n(\nu)|$, the number of paths in X^n of type ν. Let $i_j = n\nu(x_j)$ denote the number of entries of path type ν that are equal to $x_j \in X$. Then $|V_n(\nu)|$ is in bijection to the number of ways of selecting n distinct objects from a collection of size $|X|$, such that this selection has i_j of object x_j, for all j. That is, partitions of X of size i_j, where $j \in \{1, \dots, N\}$ if we suppose there are $N \leq n$ distinct entries in path type ν.

The number of these selections is

$$
\begin{aligned}
\binom{n}{i_1} \cdot \binom{n - i_1}{i_2} \dots \binom{n - i_1 - \dots - i_{N-1}}{i_N} &= \frac{n!\cancel{(n-i_1)!}\dots\cancel{(n-i_1-\dots-i_{N-1})!}}{i_1!i_2!\dots i_N!\cancel{(n-i_1)!}\dots\cancel{(n-i_1-\dots-i_{N-1})!}(n-i_1-\dots-i_N)!} \\
&= \frac{n!}{i_1!i_2!\dots i_N!} \\
&= \frac{n!}{\prod_{x \in X} (n\nu(x)!)}
\end{aligned}
$$

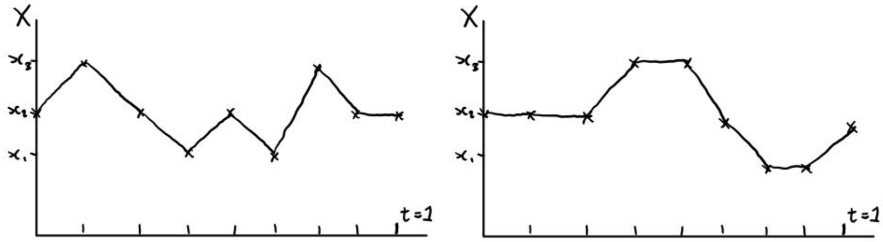

Figure 4: Two examples of a path starting at x_2 with the same empirical measure. The measure ν is defined by $\nu(x_1) = \frac{1}{4}, \nu(x_2) = \frac{1}{2}, \nu(x_3) = \frac{1}{4}$ and $\nu(x) = 0$ for all other $x \in X$.

where the second equality holds since $(n - i_1 - \cdots - i_N)! = 0! = 1$.

Stirling's formula holds for large values of $n \in \mathbb{N}$:

$$n! \approx \left(\tfrac{n}{e}\right)^n \sqrt{2\pi n}.$$

It follows that

$$\log n! \approx \log\sqrt{2\pi} + \log\sqrt{n} + n\log\frac{n}{e}.$$

So for large n,

$$n\log\frac{n}{e} \leq \log n! \leq n\log\frac{n}{e} + \log n + C,$$

where $C = \log\sqrt{2\pi}$. Applying this upper bound to $|V_n(\nu)| = \frac{n!}{\prod_{x\in X}(n\nu(x)!)}$, it follows that

$$\begin{aligned}
\log|V_n(\nu)| &= \log n! - \sum_{x\in X}\log\left(n\nu(x)!\right) \\
&\leq n\log\frac{n}{e} + \log n + C - \sum_{x\in X}\left(n\nu(x)\log\frac{n\nu(x)}{e} + \log\left(n\nu(x)\right) + C\right) \\
&= n\log\frac{n}{e} + \log n + C - n\sum_{x\in X}\nu(x)\log(\nu(x)) - n\log\frac{n}{e}\sum_{x\in X}\nu(x) - \sum_{x\in X}\left(\log\left(n\nu(x)\right) + C\right) \\
&= \log n + CnH(\nu|\#) - \sum_{x\in X}\left(\log\left(n\nu(x)\right) + C\right) \\
&=: \rho(n) - nH(\nu|\#) \\
\implies |V_n(\nu)| &\leq e^{\rho(n)}e^{-nH(\nu|\#)},
\end{aligned}$$

where $f(n) = e^{\rho(n)}$. So,

$$\begin{aligned}
\frac{1}{n}\log\left(f(n)\right) &= \frac{1}{n}\rho(n) \\
&= \frac{1}{n}\left(\log n + C - \sum_{x\in X}\left[\log\left(n\nu(x)\right) + C\right]\right) \\
&\to 0 \textit{ as } n \to \infty.
\end{aligned}$$

For the lower bound,

$$\begin{aligned}
\log|V_n(\nu)| &\geq n\log\frac{n}{e} - \sum_{x\in X} n\nu(x)\log\frac{n\nu(x)}{e} \\
&= n\log\frac{n}{e} - n\sum_{x\in X}\Big(\nu(x)\log(\nu(x)) + \nu(x)\log\frac{n}{e}\Big) \\
&= n\log\frac{n}{e} - nH(\nu|\#) - n\log\frac{n}{e}\cdot 1 \\
&= -nH(\nu|\#).
\end{aligned}$$

Therefore, $e^{-nH(\nu|\#)} \leq |V_n(\nu)| \leq f(n)e^{-nH(\nu|\#)}$. □

Proof of Theorem 3.4 (Sanov's Theorem) for discrete X: For any measurable set $A \subseteq \mathcal{P}(X)$,

$$\begin{aligned}
\mathbb{P}(L^n \in A) &= \sum_{\nu\in\mathcal{L}_n\cap A}\mathbb{P}(L^n = \nu) \\
&= \sum_{\nu\in\mathcal{L}_n\cap A}\sum_{x\in V_n(\nu)}\mathbb{P}(Z = x) && \text{(since } Z \text{ i.i.d.)} \\
&= \sum_{\nu\in\mathcal{L}_n\cap A}|V_n(\nu)|\cdot\mathbb{P}(Z = x) && \text{(all paths } x\in V_n(\nu) \text{ have equal probability)} \\
&\leq \sum_{\nu\in\mathcal{L}_n\cap A} f(n)\cdot e^{-nH(\nu|\#)}\cdot e^{-n(H(\nu|\mu)-H(\nu|\#))} && \text{(Lemmas 3.5 and 3.6)} \\
&= |\mathcal{L}_n\cap A|\cdot f(n)\cdot e^{-nH(\nu|\mu)} \\
&\leq |\mathcal{L}_n\cap A|\cdot f(n)\cdot e^{-n\inf\limits_{\nu\in A}H(\nu|\mu)} \\
\frac{1}{n}\log\left(\mathbb{P}(L^n \in A)\right) &\leq \frac{1}{n}\log\Big(|\mathcal{L}_n\cap A|\cdot f(n)\cdot e^{-n\inf\limits_{\nu\in A}H(\nu|\mu)}\Big) \\
&= \frac{1}{n}\log|\mathcal{L}_n\cap A| + \frac{1}{n}\log f(n) + \frac{1}{n}\log\Big(e^{-n\inf\limits_{\nu\in A}H(\nu|\mu)}\Big) \\
&\to -\inf_{\nu\in A}H(\nu|\mu) \ as \ n\to\infty, && \text{(Lemma 3.6 again)}
\end{aligned}$$

showing the upper bound of Sanov's Theorem. Similarly, the lower bound can be calculated. Let $(\nu_k)_{k\in\mathbb{N}}$ be a minimising sequence for $H(\nu|\mu)$. That is, $\forall\epsilon > 0, \exists\ N\in\mathbb{N}$ such that for all $k \geq N$, $\|H(\nu_k|\mu) - \inf\limits_{\nu\in\mathcal{L}_n}H(\nu|\mu)\|_\infty < \epsilon$. Suppose $\nu_k \in A$. Then,

$$\begin{aligned}
\mathbb{P}(L^n \in A) &\geq \mathbb{P}(L^n = \nu_k) \\
&= \sum_{x\in V_n(\nu_k)}\mathbb{P}(Z = x) && \text{(since } Z \text{ i.i.d.)} \\
&= |V_n(\nu)|\cdot e^{-n(H(\nu_k|\mu)-H(\nu_k|\#))} && \text{(as in proof of upper bound)} \\
&\geq e^{nH(\nu)}\cdot e^{-n(H(\nu_k|\mu)-H(\nu_k|\#))} && \text{(by Lemma 3.6)} \\
&= e^{-n(H(\nu_k|\mu))} \\
\implies \frac{1}{n}\log\mathbb{P}(L^n \in A) &\geq \frac{1}{n}\log\Big(e^{-nH(\nu_k|\mu)}\Big) \\
&= -H(\nu_k|\mu) \\
&\geq -\inf_{\nu\in A}H(\nu|\mu) - \epsilon. && (k\to\infty)
\end{aligned}$$

Taking $\epsilon \to 0$ yields the lower bound of Sanov's Theorem, completing the proof. $\square$

3.1.2 Statement of Schrödinger's Problem

In Section 3.1.2, the goal is to formulate the dynamic Schrödinger Problem in Definition 3.7, starting from Schrödinger's question in Definition 3.2 and using Sanov's Theorem (Theorem 3.4). The results presented are an intuition only rather than a rigorous formulation.

By an informal application of Sanov's Theorem, the probability $\mathbb{P}(L^n \in \cdot)$ is approximated for large n to obtain, for all H continuity sets $A \subset \mathcal{P}(X)$:

$$\begin{aligned}\mathbb{P}(L^n \in A : L_0^n \in C_0^\delta,\ L_1^n \in C_1^\delta) &\simeq \frac{\exp\left(-n \inf_{P\in\mathcal{P}(\Omega)} \left\{H(P|\mu) : P \in A, P_0 \in C_0^\delta, P_1 \in C_1^\delta\right\}\right)}{\exp\left(-n \inf_{P\in\mathcal{P}(\Omega)} \left\{H(P|\mu) : P_0 \in C_0^\delta, P_1 \in C_1^\delta\right\}\right)} \\ &= e^{-n\left(\inf_{P\in\mathcal{P}(\Omega)} \left\{H(P|\mu): P\in A, P_0\in C_0^\delta, P_1\in C_1^\delta\right\} - \inf_{P\in\mathcal{P}(\Omega)} \left\{H(P|\mu): P_0\in C_0^\delta, P_1\in C_1^\delta\right\}\right)}.\end{aligned}$$

Taking $n \to \infty$, it follows that

$$\lim_{n\to\infty} \mathbb{P}(L^n \in A : L_0^n \in C_0^\delta\ ,\ L_1^n \in C_1^\delta) = \begin{cases} 1 \text{ if } \hat{P}^\delta \in A \\ 0 \text{ if } \hat{P}^\delta \notin A \end{cases}$$

where $\hat{P}^\delta$ is the unique solution to

$$\text{minimise } \{H(P|\mu) : P \in \mathcal{P}(\Omega), P_0 \in C_0^\delta, P_1 \in C_1^\delta\}.$$

Indeed, if this solution is in the set A, then the infimum terms coincide, giving 1. If $\hat{P}^\delta \notin A$ then the term in the exponent goes to $-\infty$ as $n \to \infty$. The uniqueness of $\hat{P}^\delta$ follows by the strict convexity of $H(\,\cdot\,|\mu)$, shown in the proof of Theorem 2.7. By definition of the δ-neighbourhoods $C_0^\delta \ni \pi_0$ and $C_1^\delta \ni \pi_1$, $C_i^\delta \to \pi_i$ as $\delta \to 0$. Therefore, the solution to Schrödinger's question is given by

$$\lim_{\delta\to 0}\lim_{n\to\infty} \mathbb{P}(L^n \in A : L_0^n \in C_0^\delta\ ,\ L_1^n \in C_1^\delta) = \begin{cases} 1 \text{ if } \hat{P} \in A \\ 0 \text{ if } \hat{P} \notin A \end{cases}$$

where $\hat{P}$ is the unique solution to:

$$\text{minimise } \{H(P|\mu) : P \in \mathcal{P}(\Omega), P_0 = \pi_0, P_1 = \pi_1\}.$$

This inspires the following definition.

Definition 3.7. [15] *The* dynamic Schrödinger Problem *is to find the unique measure $\hat{P} \in \mathcal{P}(\Omega)$ such that*

$$H(\hat{P}|\mu) = \inf\{H(P|\mu) : P \in \mathcal{P}(\Omega), P_0 = \pi_0, P_1 = \pi_1\},$$

where $\mu \in \mathcal{P}(\Omega)$ is the reference measure. The solution $\hat{P} \in \mathcal{P}(\Omega)$ is called the Schrödinger bridge from π_0 to π_1.

[15] From [13].

3.2 Schilder's Theorem

Schilder's Theorem is another application of Large Deviation Theory; it is the key for connecting a special case of the dynamic Schrödinger Problem to the static Schrödinger problem and hence to entropy-regularised Optimal Transport. Section 3.2 follows [8] and [21].

The focus from Section 3.2 onwards will be on the dynamic Schrödinger Problem where each particle follows the process

$$Z_t^{\epsilon,x_0} = x_0 + \epsilon B_t,$$

where B is a Brownian motion, $\epsilon > 0$ and $t \in [0,1]$. Let Z_t^ϵ denote the problem with starting point $x_0 = 0$.

Theorem 3.8. ***(Statement of Schilder's Theorem)***: *$(Z_t^\epsilon)_{\epsilon\geq 1}$ satisfies the LDP on $\Omega = C([0,1], X)$ equipped with the topology of uniform convergence with scale ϵ^{-2} and rate function*

$$J_1(x) = \int_0^1 \tfrac{1}{2}|\dot{x}_t|^2 dt \in [0,\infty],$$

if the path $x \in \Omega$ is absolutely continuous with respect to λ and $J_1(x) = \infty$ otherwise. That is, letting μ^ϵ be the law of Z_t^ϵ, for all measurable sets $A \subset \Omega$,

$$-\inf_{x\in int(A)} J_1(x) \leq \liminf_{\epsilon\to\infty} \epsilon^2 \log \mu^\epsilon \leq \limsup_{\epsilon\to\infty} \epsilon^2 \log \mu^\epsilon \leq -\inf_{x\in cl(A)} J_1(x).$$

For all J_1 continuity sets $A \subset \Omega$,

$$\mu^\epsilon(A) \simeq \exp\left(-\tfrac{1}{\epsilon^2}\inf_{x\in A} J_1(x)\right) \text{ for } \epsilon \text{ small.}$$

Definition 3.9. *The* Cameron-Martin space *$\mathcal{H}_1$ consists of all continuous paths $\omega \in C([0,1],\mathbb{R}^d)$ where $\omega(0) = 0$, ω has a derivative $\dot{\omega}$ almost everywhere and $\int_0^1 |\dot{\omega}_s|^2 ds < \infty$. The norm of $\omega \in \mathcal{H}_1$ given by $\|\omega\|_J = J_1(\omega) = \int_0^1 \frac{1}{2}|\dot{\omega}_t|^2 dt$ makes the Cameron-Martin space $\mathcal{H}_1$ a Hilbert space.*

Theorem 3.10. *For every $j > 0$, the level set $K_1(j) := \{\omega \in C([0,1],\mathbb{R}^d) : J_1(\omega) \leq j\}$ is compact.*

Proof See Lemma 27.7 in Olav Kallenberg, Foundations of Modern Probability. Probability and its Applications (New York) Springer-Verlag, New York, Second edition, (2002).

Lemmas 3.11 and 3.12 are needed to show the lower bound of Schilder's Theorem, i.e. to show that

$$-\inf_{x\in int(A)} J_1(x) \leq \liminf_{\epsilon\to\infty} \epsilon^2 \log \mu^\epsilon.$$

Lemma 3.11. *Fix a function $f \in \mathcal{H}_1$ and let $Y^\epsilon = Z^\epsilon - f$. Suppose Y^ϵ has the law ν and Z^ϵ has the law μ. Then $\nu \ll \mu$ and*

$$\tfrac{d\nu}{d\mu}(\epsilon x) = \exp\left(-\tfrac{1}{\epsilon}\int_0^1 \dot{x}(t) dB_t - \tfrac{1}{2\epsilon^2}\int_0^1 |\dot{x}(t)|^2 dt\right).$$

Proof: Girsanov's Theorem says that if the process $M \in \Omega$ is a local Martingale under a measure μ with $M_0 = 0$, then as long as the stochastic exponential $\mathcal{E}(M)_t := \exp(M_t - \frac{1}{2}[M]_t)$ is uniformly integrable, a new probability measure ν can be defined on Ω and $\frac{d\nu}{d\mu} = \mathcal{E}(M)_\infty$. Moreover, if X is a μ-local martingale, then $X - [X, M]$ is a ν-local martingale. Let $x \in \mathcal{H}_1$. Then $M_t := -\frac{1}{\epsilon}\int_0^t \dot{x}(s)dB_s$ is a local martingale under the measure μ since the integral of a locally bounded process with respect to a martingale is a local martingale. Note also that $M_0 = 0$ and $\mathcal{E}(M)_t$ is uniformly integrable since $[M] = \frac{1}{2\epsilon^2}\int |\dot{x}(t)|^2 dt$ is bounded because $x \in \mathcal{H}_1$. By Ito's Isometry, $[M]_\infty = \frac{1}{2\epsilon^2}\int_0^1 |\dot{x}(t)|^2 dt$ and therefore, by Girsanov's Theorem, there exists a measure $\nu \ll \mu$ such that

$$\frac{d\nu}{d\mu}(\epsilon x) = \exp\left(-\frac{1}{\epsilon}\int_0^1 \dot{x}(t)dB_t - \frac{1}{2\epsilon^2}\int_0^1 |\dot{x}(t)|^2 dt\right).$$

Moreover, for the local martingale Z^ϵ,

$$Z_t^\epsilon - [Z^\epsilon, M]_t = \epsilon B_t - \int_0^t \dot{x}(s)ds = Z_t^\epsilon - f_t = Y_t^\epsilon,$$

where $f_t := \int_0^t \dot{x}(s)ds = x(t)$ is a continuous path in $\mathcal{H}_1$ and the first equality follows from the Kunita-Watanabe Theorem. $\square$

Lemma 3.12. *For any positive constants δ, γ and K, there exists an $\epsilon_0 > 0$ such that $\forall\ 0 < \epsilon \leq \epsilon_0$,*

$$\mathbb{P}(\|Z^\epsilon - x\|_\infty \leq \delta) \geq \exp\left(\frac{-J_1(x)+\gamma}{\epsilon^2}\right)$$

provided $x_0 = 0$ and $J_1(x) < K$.

Proof: Changing the measure and inserting the Radon-Nikodym derivative as in Lemma 3.11, it follows that

$$\begin{aligned}\mathbb{P}_\mu(\|Z^\epsilon - x\|_\infty \leq \delta) &= \mathbb{P}_\nu(\|Y^\epsilon\|_\infty \leq \delta) \\ &= \int\limits_{\|\epsilon x\|_\infty < \delta} \frac{d\nu}{d\mu}(\epsilon x) \cdot d\mu(\epsilon x) \\ &= e^{-\frac{J_1(x)}{\epsilon^2}} \int\limits_{\|\epsilon x\|_\infty < \delta} \exp\left(-\frac{1}{\epsilon}\int_0^1 \dot{x}_t dB_t\right) d\mu(x).\end{aligned}$$

Claim: For the Brownian motion B, $\mathbb{P}(\|\epsilon B\|_\infty < \delta) \to 1$ as $\epsilon \to 0$.
Indeed, $\|\epsilon B\|_\infty = |\epsilon| \sup\{|B_t| : t \in [0,1]\}$ and by Doob's maximal inequality,

$$\mathbb{P}\left(\sup_{t\in[0,1]} |B_t| \geq \frac{\delta}{\epsilon}\right) \leq \frac{\epsilon}{\delta}\mathbb{E}(|B_1|) \to 0 \text{ as } \epsilon \to 0.$$

So, for ϵ sufficiently small, $\mathbb{P}(\|\epsilon B\|_\infty < \delta) \geq \frac{3}{4}$. Now,

$$\begin{aligned}
\mathbb{P}\left(-\frac{1}{\epsilon}\int_0^1 \dot{x}_t \cdot dB_t \leq -\frac{2\sqrt{2}}{\epsilon}\sqrt{J_1(x)}\right) &\leq \mathbb{P}\left(\left|\frac{1}{\epsilon}\int_0^1 \dot{x}_t \cdot dB_t\right| \geq \frac{2\sqrt{2}}{\epsilon}\sqrt{J_1(x)}\right) \\
&\leq \frac{\epsilon^2 \mathbb{E}\left(\left(\int_0^1 \dot{x}_t dB_t\right)^2\right)}{8\epsilon^2 J_1(x)} \qquad \text{(by Chebyshev's inequality)} \\
&= \frac{\int_0^1 |\dot{x}_t|^2 \cdot dt}{8J_1(x)} \qquad \text{(by Itô's Isometry)} \\
&= \frac{2J_1(x)}{8J_1(x)} = \frac{1}{4}
\end{aligned}$$

Therefore, because the exponential function is strictly increasing,

$$\mathbb{P}\left[\exp\left(-\tfrac{1}{\epsilon}\int_0^1 \dot{x}_t dB_t\right) \geq \exp\left(-\tfrac{2\sqrt{2}}{\epsilon}\sqrt{J_1(x)}\right)\right] \geq \tfrac{3}{4}.$$

By the Markov inequality it follows that

$$\begin{aligned}
\frac{3}{4} \leq \mathbb{P}\left[\exp\left(-\frac{1}{\epsilon}\int_0^1 \dot{x}_t dB_t\right) \geq \exp\left(-\frac{2\sqrt{2}}{\epsilon}\sqrt{J_1(x)}\right)\right] &\leq \frac{\mathbb{E}\left[\exp\left(-\frac{1}{\epsilon}\int_0^1 \dot{x}_t dB_t\right)\right]}{\exp\left(-\frac{2\sqrt{2}}{\epsilon}\sqrt{J_1(x)}\right)} \\
\implies \mathbb{E}\left[\exp\left(-\frac{1}{\epsilon}\int_0^1 \dot{x}_t dB_t\right)\right] &\geq \frac{3}{4}\exp\left(-\frac{2\sqrt{2}}{\epsilon}\sqrt{J_1(x)}\right).
\end{aligned}$$

Now, multiplying by $\mathbb{P}(\|\epsilon B\|_\infty < \delta)$ and using the independence of these events, it can be concluded that

$$\begin{aligned}
&\mathbb{P}(\|\epsilon B\|_\infty < \delta)\mathbb{E}\left[\exp\left(-\frac{1}{\epsilon}\int_0^1 \dot{x}_t dB_t\right)\right] \geq \frac{3}{4}\cdot\frac{3}{4}\exp\left(-\frac{2\sqrt{2}}{\epsilon}\sqrt{J_1(x)}\right) \\
\implies &\int_{\|\epsilon x\|_\infty < \delta} \exp\left(-\frac{1}{\epsilon}\int_0^1 \dot{x}_t dB_t\right) d\mu(x) \geq \exp\left(-\frac{2\sqrt{2}}{\epsilon}\sqrt{J_1(x)}\right),
\end{aligned}$$

for $\epsilon \to 0$. Substituting this result back into the expression at the beginning of the proof yields the desired result,

$$\mathbb{P}_\mu(\|Z^\epsilon - x\|_\infty \leq \delta) \geq \exp\left(-\frac{J_1(x)}{\epsilon^2} - \frac{2\sqrt{2}}{\epsilon}\sqrt{J_1(x)}\right).$$

The term $-2\epsilon\sqrt{2J_1(x)}$ becomes smaller than any $\gamma > 0$ as $\epsilon \to 0$. $\square$

Proof of Theorem 3.8 (Lower Bound): From Lemma 3.12, letting $A = \{x \in \Omega : \|Z^\epsilon - x\|_\infty \leq \delta\}$, it follows that $\forall x \in A$, recalling that μ is the law for the probability denoted by $\mathbb{P}$,

$$\mu(A) \geq \exp\left(\frac{-J_1(x) + \gamma}{\epsilon^2}\right)$$
$$\implies \log(\mu(A)) \geq \frac{1}{\epsilon^2}(-J_1(x) + \gamma)$$
$$\implies \liminf_{\epsilon \to 0} \epsilon^2 \log(\mu(A)) \geq -\inf_{x \in A} J_1(x),$$

proving the lower bound of Schilder's Theorem. $\square$

The following Lemma 3.13 is the key to prove the upper bound of Theorem 3.8.

Lemma 3.13. *For every $j, \delta > 0$, let $U(j,\delta)$ be the open δ-neighbourhoods of $K_1(j) := \{\omega \in C([0,1], \mathbb{R}^d) : J_1(\omega) \leq j\}$. Then, for any $\gamma > 0$, there exists an $\epsilon_0 > 0$ such that, if $0 < \epsilon < \epsilon_0$, then*

$$\mathbb{P}\left(Z^\epsilon \notin U(j,\delta)\right) \leq \exp\left(-\tfrac{j-\gamma}{\epsilon^2}\right).$$

Proof: Leaving the subscript ϵ out for ease of notation, choose $n \in \mathbb{N}$ and let $Y_n(t)$ be the linear approximation of Z at discrete times in the set $\{t \in [0,1] : t = \frac{i}{n}, 0 \leq i \leq n\}$. Denote by T_i the interval $[\frac{i-1}{n}, \frac{i}{n}]$. First, observe that

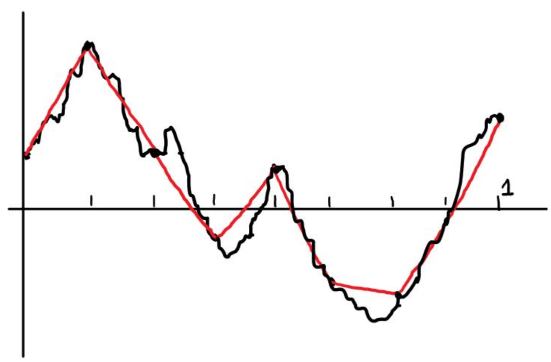

Figure 5: Example of Y_n approximating the process Z, with $n = 8$.

$$\begin{aligned}
\mathbb{P}\left(Z \notin U(j,\delta)\right) &= \mathbb{P}\left(Z \notin U(j,\delta) \cap \|Z - Y_n\|_\infty < \delta\right) + \mathbb{P}\left(Z \notin U(j,\delta) \cap \|Z - Y_n\|_\infty \geq \delta\right) \\
&\leq \mathbb{P}\left(Z \notin U(j,\delta) \cap \|Z - Y_n\|_\infty < \delta\right) + \mathbb{P}\left(\|Z - Y_n\|_\infty \geq \delta\right) \\
&\leq \mathbb{P}\left(J_1(Y_n) \geq j\right) + \mathbb{P}\left(\|Z - Y_n\|_\infty \geq \delta\right).
\end{aligned}$$

To find an upper bound on the probability $\mathbb{P}\left(J_1(Y_n) \geq j\right)$, observe that

$$\begin{aligned}
J_1(Y_n) &= \frac{1}{2}\int_0^1 |\dot{Y}_n(t)|^2 dt \\
&= \frac{1}{2}\sum_{i=1}^{n}\left|Y_n\left(\frac{i}{n}\right) - Y_n\left(\frac{i-1}{n}\right)\right|^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\left|\epsilon B\left(\frac{i}{n}\right) - \epsilon B\left(\frac{i-1}{n}\right)\right|^2 \\
&= \frac{\epsilon^2}{2}\sum_{i=1}^{n}\left|\mathcal{N}\left(0, \frac{1}{n}\right)\right|^2 \\
&= \frac{\epsilon^2}{2n}\chi_n. && \text{(where } \chi_n = \textstyle\sum_{i=1}^{n} |\mathcal{N}\left(0,1\right)|^2\text{)}
\end{aligned}$$

Hence, $\mathbb{P}\left(J_1(Y_n) \geq j\right) = \mathbb{P}(\chi_n \geq \frac{2jn}{\epsilon^2})$, for the chi-squared random variable χ_n. By Markov's inequality, Chernoff's bound can be obtained and it follows that

$$\begin{aligned}
\mathbb{P}\left(\chi_n \geq \frac{2jn}{\epsilon^2}\right) &= \mathbb{P}\left(e^{t\chi_n} \geq e^{\frac{2jnt}{\epsilon^2}}\right) \\
&\leq e^{-\frac{2jnt}{\epsilon^2}} \cdot \mathbb{E}\left(e^{t\chi_n}\right) \\
&= e^{-\frac{2jnt}{\epsilon^2}} \cdot (1-2t)^{-\frac{n}{2}} && \text{(For } t < \tfrac{1}{2}\text{, using the m.g.f. of } \chi_n\text{)} \\
&\leq \frac{1}{2}\exp\left(-\frac{2jnt}{\epsilon^2}\right).
\end{aligned}$$

This can be bounded above by $\frac{1}{2}\exp\left(-\frac{j-\gamma}{\epsilon^2}\right)$ by taking n large enough.

For the other part of the expression, it holds that

$$
\begin{aligned}
\mathbb{P}\left(\|Z - Y_n\|_\infty \geq \delta\right) &\leq \mathbb{P}\left(\sup_{t\in T_1} |Z_t - Y_n(t)| \geq \delta\right) + \mathbb{P}\left(\sup_{t\in T_2} |Z_t - Y_n(t)| \geq \delta\right) + \cdots + \mathbb{P}\left(\sup_{t\in T_n} |Z_t - Y_n(t)| \geq \delta\right) \\
&= n\mathbb{P}\left(\sup_{0\leq t\leq \frac{1}{n}} |Z_t - Y_n(t)| \geq \delta\right) \\
&\qquad\qquad (Z \text{ has stationary, independent increments}) \\
&\leq n\mathbb{P}\left(\sup_{0\leq t\leq \frac{1}{n}} \left|\epsilon B_t - \epsilon B\left(\frac{1}{n}\right)\right| \geq \delta\right) \\
&\leq n\mathbb{P}\left(\sup_{0\leq t\leq \frac{1}{n}} |\epsilon B_t| \geq \frac{\delta}{2}\right) \\
&= n\mathbb{P}\left(\sup_{0\leq t\leq \frac{1}{n}} B_t^2 \geq \frac{\delta^2}{4\epsilon^2}\right) \\
&\leq nd\mathbb{P}\left(\sup_{0\leq t\leq \frac{1}{n}} (B_t)_1^2 \geq \frac{\delta^2}{4d\epsilon^2}\right) \\
&\qquad\qquad (\text{where } (B_t)_1 \text{ is a one-dimensional Brownian Motion}) \\
&\leq nd\mathbb{P}\left(\sup_{0\leq t\leq 1} |(B_t)_1| \geq \frac{\delta\sqrt{n}}{2\epsilon\sqrt{d}}\right) && \left(\text{since } \tfrac{1}{\sqrt{n}} B_{tn} \text{ has same law as } B_t\right) \\
&= 2nd\mathbb{P}\left(\sup_{0\leq t\leq 1} (B_t)_1 \geq \frac{\delta\sqrt{n}}{2\epsilon\sqrt{d}}\right) && (\text{since } -B_t \text{ has same law as } B_t) \\
&= 4nd\mathbb{P}\left((B_t)_1 \geq \frac{\delta\sqrt{n}}{2\epsilon\sqrt{d}}\right) && (\text{reflection principle}) \\
&\leq \frac{1}{2}\exp\left(-\frac{j-\gamma}{\epsilon^2}\right) && (\text{by Lemma 3.14})
\end{aligned}
$$

So, it can be concluded that for any $\gamma > 0$, there exists an $\epsilon_0 > 0$ such that if $0 < \epsilon < \epsilon_0$, then

$$
\mathbb{P}\left(Z_\epsilon \notin U(j,\delta)\right) \leq \mathbb{P}\left(J_1(Y_n) > j\right) + \mathbb{P}\left(\|Z - Y_n\|_\infty \geq \delta\right) \leq \exp\left(-\frac{j-\gamma}{\epsilon^2}\right).
$$

□

Lemma 3.14. [16] *Let B_t be a one-dimensional Brownian motion and $t \in [0,1]$. Then,*

$$
4nd\mathbb{P}\left(B_t \geq \tfrac{\delta\sqrt{n}}{2\epsilon\sqrt{d}}\right) \leq \tfrac{1}{2}\exp\left(-\tfrac{j-\gamma}{\epsilon^2}\right).
$$

[16] The proof uses Section 7.1 of [10] as well as [8].

Proof
Let $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ be the density function of B_t, $0 \le t \le 1$. Note that $f'(x) = -xf(x)$ (1). Observe that

$$\begin{aligned}
f(x) &< f(x) + \frac{f(x)}{x^2} && \text{(since } f(x) > 0\text{)} \\
\int_y^\infty f(x)dx &< \int_y^\infty \frac{x^2 f(x) + f(x)}{x^2} dx \\
\mathbb{P}(B_t \ge y) &< \int_y^\infty \frac{d}{dx}\left(-\frac{f(x)}{x}\right) dx && \text{(using (1))} \\
&= \frac{f(y)}{y}.
\end{aligned}$$

Hence,

$$4nd\mathbb{P}\left(B_t \ge \tfrac{\delta\sqrt{n}}{2\epsilon\sqrt{d}}\right) \le \tfrac{8\epsilon d\sqrt{nd}}{\delta\sqrt{2\pi}} \exp\left(-\tfrac{\delta^2 n}{8\epsilon^2 d}\right).$$

Taking $n \to \infty$ and ϵ small enough, this can be made less than $\frac{1}{2}\exp\left(\frac{j-\gamma}{\epsilon^2}\right)$. $\square$

Proof of Theorem 3.8 (Upper Bound): To prove the upper bound of Schilder's Theorem, pick any closed set $C \subseteq \Omega$ and any $\gamma > 0$. Let $s = J_1(C) - \gamma$ and set $K = K_1(s) := \{\omega \in \Omega : J_1(\omega) \le s\}$, which is compact (Theorem 3.10). Note that by construction, $C \cap K = \emptyset$ and this implies that $\delta := \inf_{x \in C, y \in K} \|x - y\|_\infty > 0$. Denote by U the closed δ-neighbourhood of K. Then,

$$\begin{aligned}
\mathbb{P}(Z \in C) &\le \mathbb{P}(Z \notin U) && \text{(by construction)} \\
&\le \exp\left(-\frac{s-\gamma}{\epsilon^2}\right) && \text{(Lemma 3.13)} \\
&= \exp\left(-\frac{J_1(C) - 2\gamma}{\epsilon^2}\right) && \text{(definition of } s\text{)} \\
\implies \log \mathbb{P}(Z \in C) &\le -\frac{J_1(C) - 2\gamma}{\epsilon^2} \\
\implies \limsup_{\epsilon \to 0} \epsilon^2 \log \mu(C) &\le -J_1(C) + 2\gamma
\end{aligned}$$

for any arbitrary $\gamma > 0$ and closed $C \subset \Omega$. $\square$

3.3 The Dynamic and Static Schrödinger Problems

Section 3.3 follows [13]. Let $\Omega = C([0,1], X)$. The Contraction Principle (Theorem 3.15) is used to derive a large deviation result for the static Schrödinger Problem, via the projection of $\mu \in \mathcal{P}(\Omega)$ onto the product space X^2. The static Schrödinger Problem is defined in Definition 3.16 and its connection to the dynamic problem is shown in the special case where $Z_t^\epsilon = \epsilon B_t$. The equivalence of the static Schrödinger Problem and the entropy-regularised Optimal Transport problem is also shown for this case.

Recall from Definition 3.7 that the dynamic Schrödinger Problem is:

$$\text{minimise } \{H(P|\mu^\epsilon) : P \in \mathcal{P}(\Omega), P_0 = \pi_0, P_1 = \pi_1\},$$

where $\mu^\epsilon \in \mathcal{P}(\Omega)$ is the reference measure, $P_0, P_1 \in \mathcal{P}(X)$ are the time marginals of P at $t = 0$ and $t = 1$ and $\pi_0, \pi_1 \in \mathcal{P}(X)$. Consider the case where the reference measure μ^ϵ is the law of the process $Z_t^\epsilon = \epsilon B_t$, where B is Brownian motion.

Recall that Schilder's Theorem (Theorem 3.8) gives a large deviation result for $\mu^\epsilon \in \mathcal{P}(\Omega)$. For all J_1 continuity sets $A \in \Omega$,

$$\mu^\epsilon(A) \simeq \exp\left(\frac{1}{\epsilon^2} \inf_{x \in A} J_1(x)\right) \text{ for small } \epsilon,$$

where the function $J_1 : \Omega \to [0, \infty]$ is defined as $J_1(x) := \frac{1}{2} \int\limits_0^1 |\dot{x}_t|^2 dt$.

Theorem 3.15. ***(Contraction Principle)***[17] *Let X and Y be polish spaces and let $f : X \to Y$ be a continuous function. Consider the rate function $I : X \to [0, \infty]$ that is also coercive.*
(a) For each $y \in Y$, define $I'(y) := \inf\{I(x) : x \in X, y = f(x)\}$. Then I' is a rate function on Y and it is coercive.
(b) If I controls the LDP associated with a family of probability measures $\{\mu^\epsilon\}$ on X, then I' controls the LDP associated with the family of probability measures given by the push-forward measures $\mu^\epsilon \circ f^{-1}$ on Y.

Proof: I' is nonnegative. Since I is coercive, for all $y \in Y$ such that $y = f(x)$ for some $x \in X$, the infimum in the definition of I' is obtained at some point of X. So, the level sets of I', $K'(\alpha) = \{y \in Y : I'(y) \leq \alpha\}$, are given by

$$K'(\alpha) = \{f(x) : I(x) \leq \alpha\} = f(K(\alpha))$$

where $K(\alpha)$ are the corresponding level sets of I. Since $K(\alpha)$ are compact in X, $K'(\alpha) \subset Y$ compact.
(b) By the definition of I', for any $A \subset Y$, it holds that

$$\inf_{y \in A} I'(y) = \inf_{x \in f^{-1}(A)} I(x). \tag{3}$$

f is continuous, so if $A \subset Y$ is closed then $f^{-1}(A) \subset X$ is closed and if $A \subset Y$ is open then $f^{-1}(A) \subset X$ is open. Therefore, the LDP for $\mu^\epsilon \circ f^{-1}$ follows as a consequence of the LDP for μ^ϵ and (3). □

[17] From page 126 of [8], Theorem 4.2.1.

Let $f : \Omega \to X^2$ be the continuous map given by the projection onto the initial and final time marginals. Then, by Theorem 3.15 and Schilder's Theorem, $\mu_{01}^{\epsilon} := \mu^{\epsilon} \circ f^{-1}$ satisfies the LDP in X^2 with the coercive rate function

$$c(x, y) := \inf\{J_1(\omega) : \omega \in \Omega, \omega_0 = x, \omega_1 = y\}.$$

By Jensen's inequality, $c(x, y) = \frac{1}{2}|x - y|^2$ since

$$\tfrac{1}{2}|x-y|^2 = \tfrac{1}{2}\left(\int_0^1 \dot{\omega}_t dt\right)^2 \leq \int \tfrac{1}{2}\dot{\omega}_t{}^2 dt = J_1(\omega), \text{ for any path } \omega \in \Omega.$$

Therefore, for all c continuity sets $(x, y) \subset X^2$ the measure $\mu_{01}^{\epsilon} \in \mathcal{P}(X^2)$ is approximated by

$$\mu_{01}^{\epsilon}(x, y) \simeq \exp\left(-\frac{1}{2\epsilon^2}|x - y|^2\right), \text{ for small values of } \epsilon. \tag{4}$$

Definition 3.16. *The* static Schrödinger Problem *is to find the plan* $\gamma \in \mathcal{P}(X^2)$ *satisfying:*

$$\textit{minimise } \{H(\gamma|\mu_{01}^{\epsilon}) : \gamma \in \mathcal{P}(X^2), \gamma_0 = \pi_0, \gamma_1 = \pi_1\}$$

where $\pi_0, \pi_1 \in \mathcal{P}(X)$ *are the initial and final distributions of the dynamic problem and* $\mu_{01}^{\epsilon} \in \mathcal{P}(X^2)$ *is the reference measure. Recalling the notation for the family of transport plans from Section 1, the problem is to find* $\inf\{H(\gamma|\mu_{01}^{\epsilon}) : \gamma \in \Pi(\pi_0, \pi_1)\}$.

It will be shown that the solution to the dynamic Schrödinger Problem can be found via the static Schrödinger Problem. Let P_{01} and μ_{01} denote the joint law of the initial and final positions of P, μ. That is, $P_{01} = (X_0, X_1)_{\#}P \in \mathcal{P}(X^2)$ and $\mu_{01} = (X_0, X_1)_{\#}\mu \in \mathcal{P}(X^2)$. Let P^{xy} and μ^{xy} denote the conditional laws of P, μ given $X_0 = x$ and $X_1 = y$. Explicitly, $P^{xy}(\cdot) = P(\cdot|X_0 = x, X_1 = y)$ and $\mu^{xy}(\cdot) = \mu(\cdot|X_0 = x, X_1 = y)$. Then, the tensorisation property of relative entropy implies that

$$H(P|\mu) = H(P_{01}|\mu_{01}) + \int_{X^2} H(P^{xy}|\mu^{xy}) dP_{01}(x, y).$$

Indeed,

$$\begin{aligned}
\text{RHS} &= \mathbb{E}_{P_{01}}\left(\log \frac{dP_{01}}{d\mu_{01}}\right) + \mathbb{E}_{P_{01}}\left(\mathbb{E}_{P^{xy}}\left(\log \frac{dP^{xy}}{d\mu^{xy}}\right)\right) \\
&= \mathbb{E}_{P_{01}}\left(\mathbb{E}_{P^{xy}}\left(\log \frac{dP_{01}}{d\mu_{01}} \frac{dP^{xy}}{d\mu^{xy}}\right)\right) \quad \text{(By linearity of } \mathbb{E}_{P_{01}} \text{ and } P^{xy}\text{-measurability of } \tfrac{dP_{01}}{d\mu_{01}}) \\
&= \mathbb{E}_P\left(\log \frac{dP}{d\mu}\right) \qquad \text{(Independence of } \tfrac{dP_{01}}{d\mu_{01}} \text{ and } \tfrac{dP^{xy}}{d\mu^{xy}})
\end{aligned}$$

Now, the dynamic Schrödinger problem can be rewritten by decomposing the marginal restraint into $P_{01} = \gamma \in \mathcal{P}(X^2)$, where the marginal distributions of γ are $\pi_0, \pi_1 \in \mathcal{P}(X)$. That is,

$$\inf\{H(P|\mu) : P \in \mathcal{P}(\Omega), P_0 = \pi_0, P_1 = \pi_1\}$$

can be rewritten as:

$$\inf\{\inf\,[H(P|\mu) : P \in \mathcal{P}(\Omega), P_{01} = \gamma] : \gamma \in \Pi(\pi_0, \pi_1)\}.$$

By applying the above tensorisation property, notice that

$$\begin{aligned}\inf\left[H(P|\mu) : P \in \mathcal{P}(\Omega), P_{01} = \gamma\right] &= H(\gamma|\mu_{01}) + \inf\left[\int_{X^2} H(P^{xy}|\mu^{xy})dP_{01}(x,y) : P_{01} = \gamma\right] \\ &= H(\gamma|\mu_{01}).\end{aligned}$$

This holds because $\inf\left[\int_{X^2} H(P^{xy}|\mu^{xy})dP_{01}(x,y) : P_{01} = \gamma\right] = 0$ and this is achieved when $P^{xy} = \mu^{xy}$. Hence, for each $\gamma \in \mathcal{P}(X^2)$,

$$\inf\left[H(P|\mu) : P \in \mathcal{P}(\Omega), P_{01} = \gamma\right] = H(\mu^{\gamma}|\mu)$$

where $\mu^{\gamma} = \int_{X^2} \mu^{xy} d\gamma(x,y) \in \mathcal{P}(\Omega)$ is the disintegration of μ with respect to $\gamma \in \mathcal{P}(X^2)$.

Therefore,

$$\inf\left[H(P|\mu) : P \in \mathcal{P}(\Omega), P_{01} = \gamma\right] = H(\mu^{\gamma}|\mu) = H(\gamma|\mu_{01}).$$

Taking the infimum of this final term will give the solution $\hat{\gamma} \in \mathcal{P}(X^2)$ to the static Schrödinger Problem from Definition 3.16. This plan is used to find the measure $\hat{P} = \mu^{\hat{\gamma}} = \int_{X^2} \mu^{xy} d\hat{\gamma}(x,y) \in \mathcal{P}(\Omega)$, the unique solution to the dynamic problem.

It remains to be shown that solving the static Schrödinger problem is equivalent to solving the entropy-regularised Optimal Transport problem. To see this, observe that

$$\begin{aligned}H(\gamma|\mu_{01}) &= \int_{X^2} \log\left(\frac{d\gamma}{d\mu_{01}}\right) d\gamma \\ &= \int_{X^2} \log\left(\frac{d\gamma}{d\lambda}\right) d\gamma + \int_{X^2} \log\left(\frac{d\lambda}{d\mu_{01}}\right) d\gamma && \text{(Radon-Nikodym chain rule)} \\ &= H_2(\gamma) + \frac{1}{2\epsilon^2}\int_{X^2} |x-y|^2 d\gamma(x,y). && \text{(recalling (4))}\end{aligned}$$

It has been informally shown that the solution $\hat{P} \in \mathcal{P}(\Omega)$ of the dynamic Schrödinger problem for the system driven by $Z_t^{\epsilon} = \epsilon B_t$ can be found by finding $\gamma \in \mathcal{P}(X^2)$ that solves the static Schrödinger problem. Moreover, finding the measure $\gamma \in \mathcal{P}(X^2)$ that solves the static problem is equivalent to finding the optimal plan $\gamma \in \mathcal{P}(X^2)$ solving the entropy-regularised Optimal Transport problem at Definition 2.1. In [13], the relationship between these problems is presented more generally and the Γ-convergence of Schrödinger's problem to the Optimal Transport problem is shown directly.

References

[1] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport, Modelling and Optimisation of Flows on Networks. Lecture Notes in Mathematics, vol 2062. Springer, Berlin, Heidelberg, 2013.

[2] Stephen Boyd and Lieven Vandenberghe. Convex Optimisation, Cambridge University Press, 2004.

[3] Andrea Braides. A Handbook of Γ-convergence. Handbook of Differential Equations. Stationary Partial Differential Equations, Volume 3 (M. Chipot and P. Quittner, eds.), Elsevier, 2006.

[4] Guillaume Carlier, Vincent Duval, Gabriel Peyré and Bernhard Schmitzer. Convergence of Entropic Schemes for Optimal Transport and Gradient Flows, 2017. SIAM J. Math. Anal. 49 (2017), no. 2, 1385–1418.

[5] Yongxin Chen, Tryphon Georgiou and Michele Pavon. On the relation between optimal transport and Schrodinger bridges: A stochastic control viewpoint, 2014. J. Optim. Theory Appl. 169 (2016), no. 2, 671–691.

[6] Christian Clason, Dirk Lorenz, Hinrich Mahler, Benedikt Wirth, Entropic regularization of continuous optimal transport problems, eprint arXiv:1906.01333v2, June 2019.

[7] Marco Cuturi. Sinkhorn Distances: Lightspeed computation of optimal transportation distances, Advances in Neural Information Processing Systems 26, pages 2292–2300, 2013.

[8] A. Dembo and O. Zeitouni. Large deviations techniques and applications. Second edition. Applications of Mathematics (New York), 38. Springer-Verlag, New York, 1998.

[9] Simone Di Marino and Augusto Geralin, An Optimal Transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm, arXiv:1911.06850v1 [math.PR] 15 Nov 2019.
https://arxiv.org/pdf/1911.06850.pdf

[10] William Feller, An introduction to probability theory and its applications, Third Edition, 1950, John Wiley and Sons.
https://www.ime.usp.br/ fmachado/MAE5709/FellerV1.pdf

[11] Aude Genevay. Entropy-Regularized Optimal Transport for Machine Learning, Dauphine Universite Paris, 13 March 2019.

[12] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.

[13] Christian Léonard. From the Schrödinger problem to the Monge-Kantorovich problem, Journal of Functional Analysis, 262 (2012), pp. 1879–1920.
https://arxiv.org/abs/1011.2564v1

[14] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport, Discrete Contin. Dyn. Syst. A, 34 (2014), pp. 1533–1574.

[15] Christian Léonard. Some Properties of Path Measures, 2013. Séminaire de Probabilités XLVI, 207–230, Lecture Notes in Math., 2123, Springer, Cham, 2014.
https://arxiv.org/abs/1308.0217

[16] Peter Mörters, Large deviation theory and applications, University of Bath Lecture Notes, November 10 2008.
https://people.bath.ac.uk/maspm/LDP.pdf

[17] T. Muthukumar. Introduction to Γ-convergence, Indian Institute of Technology Kanpur, Lecture Notes 2010.
http://home.iitk.ac.in/ tmk/courses/minicourse/Gamma/tmkMSAH.pdf

[18] Gabriel Peyre and Marco Cuturi. Computational Optimal Transport, Foundations and Trends in Machine Learning, vol. 11, no. 5-6, pp. 355-607, 2019.

[19] Filippo Santambrogio. Optimal Transport for Applied Mathematicians, Calculus of Variations, PDEs and Modeling, version of May 2015.

[20] Erwin Schrödinger. Über die Umkehrung der Naturgesetze, Sitzungsberichte Press. Akad. Wiss. Berlin. Phys. Math.,144 (1931), pp. 144–153.

[21] Cosma Shalizi, lecture notes Freidlin-Wentzell Thoery Chapter 35: Large Deviations for Stochastic Differential Equations, Carnegie Mellon University, 2006.
https://www.stat.cmu.edu/-cshalizi/754/2006/notes/lecture-35.pdf

[22] François-Xavier Vialard. An elementary introduction to entropic regularization and proximal methods for numerical optimal transport. Doctoral. France. 2019. hal-02303456f

[23] Cedric Villani, Optimal Transport: Old and New, vol. 338 of Grundlehren der mathematischen Wissenschaften, Springer, 2009.