4 Problems in optimal transportation

by

Joaquín Sánchez García

A thesis submitted in conformity with the requirements for the degree of PhD. in Mathematics, Graduate Department of Mathematics, in the University of Toronto

Contents

General Introduction

This thesis is devoted to the study of 4 different problems for which the theory of optimal transportation is well-suited:

1. The generalization of the Schrödinger Problem to synthetic Lorentzian geometries.
2. The small time existence for solutions for the aggregation equation on compact Riemannian manifolds for non-regular interaction potentials via the minimizing movement scheme.
3. The technique of measure pre-conditioning general Machine-Learning tasks and Domain Adaptation transfer learning.
4. The generalization of an economic model of Roy for partition of labor including occupational choice as a constraint.

For the 4 problems we use the theory of optimal mass transportation which has been widely developed in the last years.

Part 1: The Schrödinger problem in synthetic Lorentzian geometries

The Schrödinger problem refers to the minimization of relative entropy with respect to a reference measure. The Schrödinger problem is usually analyzed in two related formulations: the static and the dynamic Schrödinger problems. In this document we study both approaches. One of the main questions of the Schrödinger problem is whether or not the solutions to the entropically regularized optimal transport problem converges to solutions of the optimal transport problem. In the dynamical setting, this property amounts to study the Large Deviation Principles of the reference measure.
In the Riemannian case, the law of Brownian motion is a Markov measure which satisfies a large deviation principle directly related to geodesic flow. So far, there is no analogue of the Riemannian Brownian motion in the Lorentzian setting. We study a Levy-like construction proposed by Dr. McCann which emulates the behaviour of Brownian bridges. With this construction we recover a partial version of the entropic convergence in the non-smooth Lorentzian case.

Main Take away 1. *In the Lorentzian case bridges fail to be Markovian and satisfy large deviation principles for geodesic flow. The absence of a heat kernel impedes us from using the elliptic theory. A Levy-like construction allows us to prescribe large deviation principles to use in entropic regularizations.*

Part 2: The aggregation equation via the minimizng movement scheme in compact Riemannian manifolds

The aggregation equation for non-regular potentials on Riemannian manifolds is a very active area of research. In this work we study the small-time existence of solutions for non-smooth potentials via the minimizing movement scheme. The theory of gradient flows in metric spaces does not consider a potential non-regularity of potentials in the cut-locus. The presence of the cut-locus presents a difficulty for the JKO scheme to choose a direction, nevertheless we show explicitly a time bound for which we can flow the minimizing movement scheme. The minimizing movement scheme is essential in numerical algorithms, so understanding it's scope is fundamental.

Main Take away 2. *The minimizing movement scheme is shown to converge to a limiting path measure satisfying the aggregation equation up to a time determined by the distance to the cut locus from points in the support of the initial measure.*

Part 3: Measure Pre-conditioning in Machine-Learning

We study a new technique to improve convergence of algorithms for specific ML-tasks. We show that if the modifications of the problem at level n (sample size) is done in a specific way (full learner recovery systems) we can show analytical convergence of subsequence to the original model. This technique seems to be specifically important for Domain Adaptation in transfer learning as the modifications can ensure existence of analytical tools otherwise unavailable.

Main Take away 3. Γ*-convergence shows that small modifications (in uniform ways) yield good approximations of machine-learning models. The hypothesis for this convergence amounts to checking a double-sided Fatou Lemma.*

Part 4: Generalizing an economic model of Roy for labor partition using occupational choice as a constraint

We study the analytical properties of a generalization of the economic model for labor force partition studied by Dr. Roy. The new model proposed by Dr. Siow, includes occupational choice as a constraint rather than a consequence. This difference allows us to rewrite the problem in an analytically useful way. We explain how this formulation relates to the identification problem and compare it's consequences to previous known conclusions from the linear model of Roy.

Main Take away 4. *Roy's model assumes a linear separation function for the partition of labor. Taking away such assumption is technically difficult but the non-linear version is more realistic and still allows economically interesting results.*

Chapter 1

The non-smooth relativistic Schrödinger Problem

1.1 Introduction

In this paper we continue the investigation of C. Léonard on the Schrödinger problem as we adapt it to the synthetic relativistic setting of [McCann2019] ([McCann2023], [Kunziger-Saemann],[Cavalletti-Mondino],[Braun] and others). We manage the difficulty of bridges being non-timelike which is a non-physical condition.
We extend the work of [Leonard2014], [Leonard2012], [Tamanini] to the framework of synthetic Lorentz geometry of low regularity. Recent work [McCann2023], [Kunziger-Saemann], [Eckstein-Miller], [Cavalletti-Mondino] has shown the synthetic approach to Lorentzian geometry in terms of optimal transportation to be incredibly fruitful.
In [Leonard2014], the theory of the Schrödinger Problem in general polish spaces was developed (see [Leonard2014], [Leonard2012] or [Tamanini]). We adapt this theory to the physically relevant generalizations of Lorentzian manifolds on which the underlying topology generated by cones is assumed to be only metrizable, so that the apparent dependence on the underlying metric is only through the topology it generates and its set of rectifiable curves. Informally speaking, globally hyperbolic chrono-regular Lorentzian length spaces are spaces with a chronological and causal structure whose underlying chronological topology is assumed to only be metrizable, each causal emerald is compact and ℓ-curves do not contain non-constant null subsegments (see [McCann2023]). We study the Schrödinger problem in such spaces.
The Schrödinger Problem has been extensively studied in the context of Polish spaces in the seminal work of C. Leonard, see [Leonard2014], [Leonard2012], [Leonard2001] and in the general case of $RCD(K,N)$ spaces see [Tamanini]. The Schrödinger problem can also be posed on euclidean phase-space (see [Chiarini-Conforti-Greco]). In the study of general relativity (and more generally non-smooth Lorentzian length and pre-length spaces) there is no underlying Hilbertian structure nor a canonical heat semigroup. The absence of such semigroup is associated to the absence of a canonical Brownian motion. The Wiener measure (law of the Brownian motion) is essential to the study of the Schrödinger Problem (see for example [Leonard2014, Section 5]).

The focus of this work is to study how the techniques on the Schrödinger Problem and the non-smooth causality theory interact.
One of the main tools for studying the "classical" Schrödinger problem is a Brownian bridge measure. Along our study, we find two types of problems when adapting the framework of Schrödinger bridges to general spacetimes. The first type of problem is the inclusion of a physical constraint for probability measures. These problems are typically resolved by a convexity property on the set of feasible probability measures. The latter, somewhat more unsatisfactory is the problem of the external parameter for curves and the underlying un-physicality of some intrinsic concepts of metric spaces.
The static Schrödinger problem consists in finding among all probability measures on the product space with fixed marginals, the one that minimizes relative entropy with respect to a reference measure. To make the Schrödinger problem physical, we add the restriction of the support of the measures being causal, meaning that they are concentrated on pairs of points on which one can physically travel. This imposition turns out to be somewhat immaterial for the static problem. We then consider the dynamical Schrödinger problem on which we study the paths travelled by particles on the Schrödinger problem; travelled paths must be future d, adding a constraint to the problem. In the context of [Leonard2014] and [Tamanini], some of the most important developments occur when the reference measure on the dynamic problem is Markovian, meaning that the past and the future are dependent only through the present. A subtlety here is presented, the Markov property can be thought of in two different ways for the Schrödinger problem: i.) by physically considering past and present with respect to the chronological and causal relations on the space or ii.) unphysically by using the external time parametrization of curves. It turns out both approaches yield relatively similar consequences. We study both methods and explain their differences.
In the Riemannian case, the Wiener measure in path space (the law of Brownian motion) satisfies a large deviation principle which can be used to establish convergence of solutions for the dynamic Schrödinger problem. Motivated by this program, we analyze a process defined originally by R. Dudley in Minkowski space in the seminal work [Dudley1966]. This process behaves in several ways as a Brownian Motion in phase space. We analyze the bridges of this process and relate it to a new construction in globally hyperbolic chrono-regular Lorentzian length spaces similar in nature to the Brownian bridge construction of Levy. This construction replicates the idea of bridges, even in curved geometries (where Dudley's process has also been generalized [Franchi-LeJan2007], [Dunkel-Hanggi], [Chevalier-Debbasch]).

1.1.1 Organization of the paper

In section 1.2.1 we set the framework of the underlying space of study. We recall some results on causality and the general framework of Lorentzian pre-length spaces developed by [Kunziger-Saemann], [McCann2023], [McCann2019], [Eckstein-Miller] and others. We define the " relativistic" version of the Static Schrödinger problem (RSch). In section 1.3 we study bridge spaces. A bridge space consists of curves (continuous/cadlag) with beginning and endpoint fixed. We start by recalling some properties of the Brownian Bridge in $\mathbb{R}^n$ and Riemannian manifolds as studied in [Hsu]. In 1.3.1 we start a construction on Minkowski space that resembles Levy's construction of Brownian bridges. We also study the conditional version of the process of Dudley (1.3.2) and use his definition of the Markov property. We formulate the apparently non-physical Markov Property in (74). In section 1.4 we study the large deviations principle, it's connections to this work and study the convergence of bridges related to the dynamical Schrödinger problem.

The adaptation of the Schrodinger Problem to the relativistic setting encounters the difficulty of the absence of the canonical heat semigroup. This absence is dealt by studying it's fundamental properties separately. That is, we study bridge measures by Markovianity and by small time asymptotics. This separation breaks the field in two interesting branches that we study in this document.

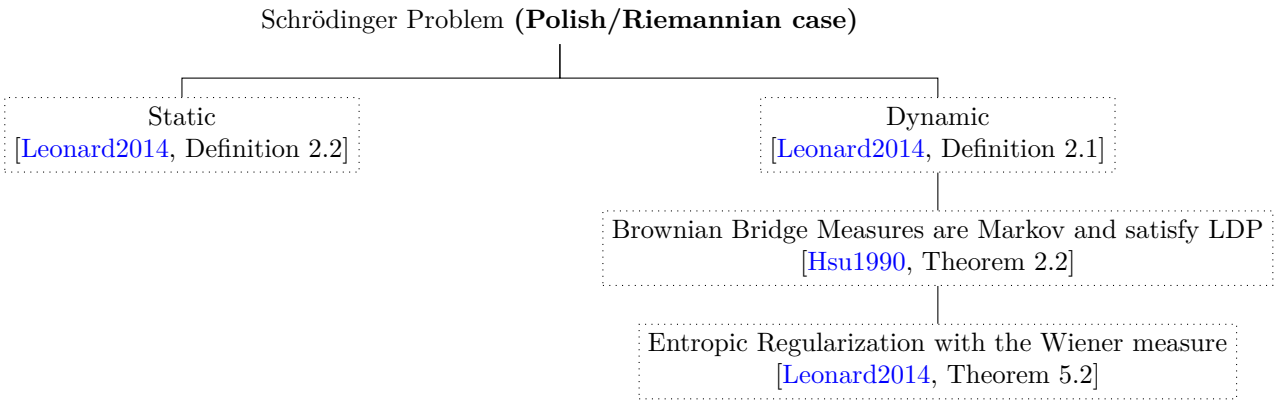

Figure 1.1: In the polish space case, the entropic regularization of optimal transport problem uses the Wiener measure as reference measure. Weiner's measure satisfies at the same time LDP properties and a canonical Markov property.

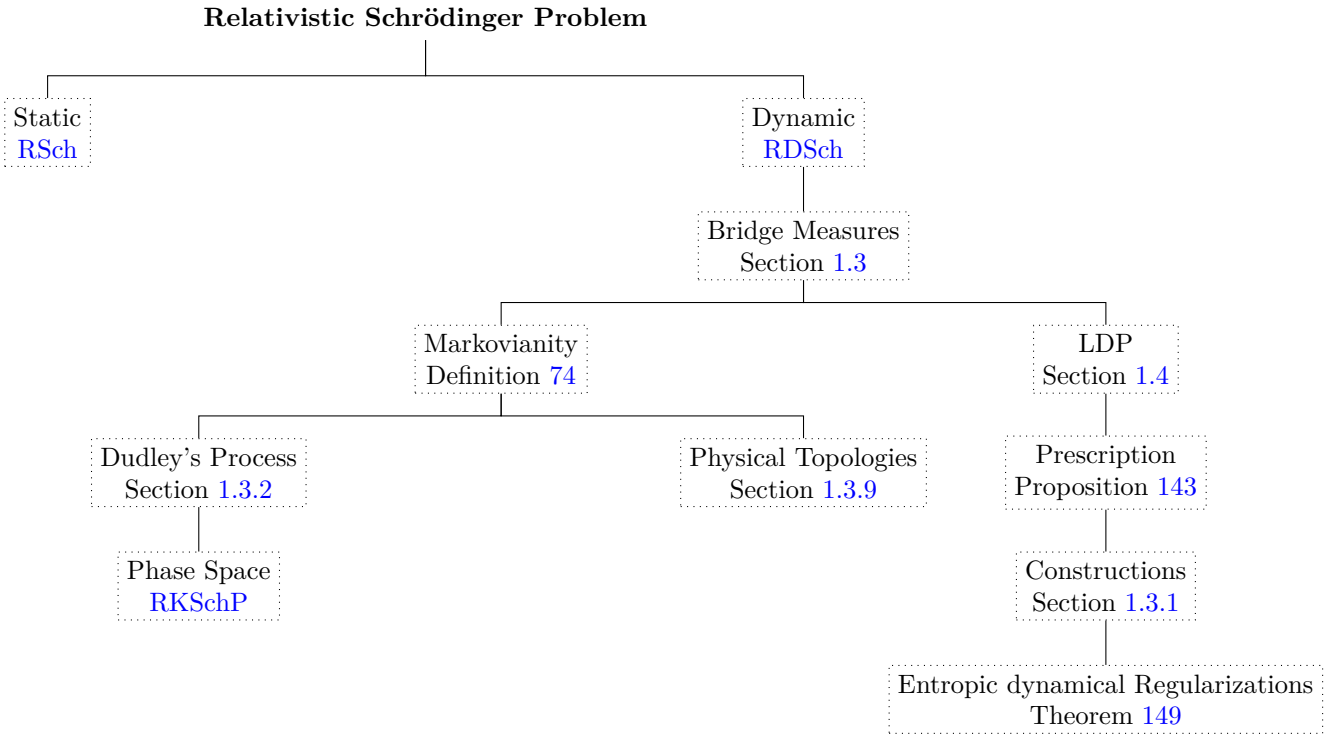

Figure 1.2: Outline of the document: Different to the Riemannian case, there seems to be no canonical generalization of the Brownian bridge. Generalizations fail to be either completely causal, Markovian or satisfy small time asymptotics. We study both the Markov property and the large deviation principle separately and use the latter to study entropic regularizations of Lorentzian costs.

1.2 Statement of the problem

1.2.1 Background: Lorentzian length spaces and fuzzy events

We mainly follow the conventions on [McCann2019] and [Kunziger-Saemann] different to [McCann2023] where sign of the time separation is assumed to be negative towards the benefit of using the more familiar notation of metric (and length) spaces. The first part of this introduction replicates the introduction of [McCann2023] but with positive separation function. Even though most of the concepts are well-known in non-smooth theory, each convention (inclusion of infinities, signs and terms) is included here for the purpose of clarity and self-containment. A reader familiar with [McCann2023] can easily skip this introductory section.

Definition 1. *(Time separation function ℓ)*
For a set M, a time separation function on M is a function $\ell : M \times M \to \{-\infty\} \cup [0, \infty]$ satisfying

1. $\ell(x, x) \geq 0$ *for all* $x \in M$
2. $\ell(x, y) \geq \ell(x, z) + \ell(z, y)$ *for all* $x, y, z \in M$.

Observe that both conditions imply $\ell(x, x) \in \{0, \infty\}$ and we will always assume (unless explicitly stated) that $\ell^{-1}(\{\infty\}) = \varnothing$ is imposed in which case $\ell(x, x) = 0$ is implied.

Definition 2. *(Causal and Chronological relations)*
Given a set M and a time separation function ℓ on M, the chronological and causal relations between pairs of points in M are described via

1. $x \ll y$ *if and only if* $(x, y) \in M^2_{\ll} := \ell^{-1}((0, \infty])$
2. $x \leq y$ *if and only if* $(x, y) \in M^2_{\leq} := \ell^{-1}([0, \infty])$.

In the first case we say x belongs to the timelike or chronological past of y and in the second that x belongs to the causal past of y. As convention we say y belongs to the future of x in any of both cases with the possibility of making the precision of causal or chronological case.

Causal and chronological relations are interpreted as the possibility to travel between spacetime events.

Definition 3. *(Causal and chronological pasts and futures)*
Given a set M endowed with a time separation function ℓ we call the sets

$$I^+(x) := \ell(x, \cdot)^{-1}((0, \infty]) \text{ and } J^+(x) = \ell(x, \cdot)^{-1}([0, \infty])$$

the timelike and causal futures of x respectively.
Similarly for $y \in M$, the sets

$$I^-(y) := \ell(\cdot, y)^{-1}((0, \infty]) \text{ and } J^-(y) = \ell(\cdot, y)^{-1}([0, \infty])$$

the timelike and causal pasts of y.

Following the notation of [McCann2019] and [McCann2023], a sufficient condition for $M^2_{\ll}$ and $M^2_{\leq}$ to be antisymmetric relations is that $\min\{\ell(x, y), \ell(y, x)\} > -\infty$ if and only if $x = y$. This antisymmetry, which is essentially the identity of indiscernibles, is shown to be satisfied for the

spaces we will work on throughout this paper (see [McCann2023]).
The relations $M^2_{\ll}$ and $M^2_{\leq}$ induced by ℓ satisfy the so called *push-up* property: $x \ll z$ and $z \leq y$ imply $x \ll y$. Observe that antisymmetry can always be obtained by studying instead an appropriate quotient space in terms of equivalence classes for the time separation function $\tilde{\ell}(\tilde{x}, \tilde{y}) = \sup_{x \in \tilde{x}, y \in \tilde{y}} \ell(x, y)$, see [McCann2023, Lemma 1] for details.
For this reason we will always assume the antisymmetry condition holds. One fundamental difference between the work of [McCann2023] and previous descriptions of Lorentzian pre-length spaces is ability to study *rough paths*. The idea is to explore all timelike and causal paths (without assumptions on their continuity) and derive that the underlying topologies (to be described briefly) yield continuity properties with respect to an assumed topology induced by a distance. The framework of Lorentzian length spaces developed in [Kunziger-Saemann] allows us to relate the assumed underlying topology to the ones induced by the structure of cones and diamonds.

Definition 4. *((Rough) Causal paths)*
A path on M is just a function from an interval $A \subseteq \mathbb{R}$ onto M. A path $\sigma : A \to M$ is said to be causal if $s < t$ implies $\sigma(s) \leq \sigma(t)$.

Observe that we can not yet describe continuity in any way as we haven't explored the topologies on M. We will define continuous causal paths when we endow M with a topology. The word *rough* used in [McCann2023] is used to emphasize this (a-priori) lack of continuity, the word *curve* will be saved to emphasize the regularity of paths. The definition of a timelike path is that of Definition 4 using the timelike relation instead.

Definition 5. *(ℓ-length)*
For a causal path $\sigma : A \to M$, we define it's ℓ-length via

$$L_\ell(\sigma) = \inf \left\{ \sum_{k=0}^{N} \ell(\sigma(t_k), \sigma(t_{k+1})) \right\} \tag{1.1}$$

where the infimum is taken over all finite partitions of A. If A is not compact, the L_ℓ-length is defined via exhaustion of A by compact sets. Note that by the reverse triangle inequality we have $L_\ell(\sigma) \in [0, \ell(\sigma(0), \sigma(1)]$.

Definition 5 is somewhat unintuitive because of our sign convention. Towards making the work as understandable as possible we define $\ell^- := -\ell$ and use ℓ and ℓ^- depending on what we think makes each result more understandable. With this notation (1.1) reads

$$L_{\ell^-}(\sigma) = \sup_{\{t_k\}} \left\{ \sum_{k=0}^{N} \ell^-(\sigma(t_k), \sigma(t_{k+1})) \right\}$$

which couples better with our notion of length and corresponds to the convention used in [McCann2023].
In this notation we have $L_{\ell^-}(\sigma) \in [\ell^-(\sigma(0), \sigma(1)), 0]$.
Notice that with the previous conventions, the definition of L_{ℓ^-} resembles exactly the definition of length we know (with the obvious difference of being negative).
The following definition generalizes the idea of straight-lines. The connection between geodesics, geometry and transport are well-known so it is fundamental to have the correct definition of geodesics.

Definition 6. *(ℓ-paths and timelike ℓ-path spaces)*
A timelike path $\sigma : [0,1] \to M$ is called an ℓ-path if

1. $\ell(\sigma(0), \sigma(1)) \in (0, \infty)$
2. $\ell(\sigma(s), \sigma(t)) = (t-s)\ell(\sigma(0), \sigma(1))$ *for all* $t \geq s, t, s \in [0,1]$.

and the convention is that ℓ-paths are parametrized proportional to proper time.

Remark 7. *The parameter $t \in [0,1]$ will be relevant in the following sections, for now we just think about it as defining the curves.*

Definition 8. *(Timelike path space)*
(M, ℓ) is called a timelike path space if for every pair of points $(x, y) \in M^2_{\ll}$ there exists a time-like ℓ-path with $\sigma(0) = x, \sigma(1) = y$. (M, ℓ) is called timelike non-branching unless there exist two distinct ℓ-paths that coincide in an open interval.

Definition 9. *(Metric spacetime and Lorentzian pre-length spaces)*
If (M, d) is a metric space and ℓ is a separation function on M, we call (M, d, ℓ) a metric spacetime. A metric spacetime on which $\ell^+ := \max\{\ell, 0\}$ is lower semicontinuous is called a Lorentzian Pre-Lenth Space(LPS).
A Lorentizan Pre-length (LPL) space is called causally closed if $M^2_{\leq}$ is closed (with respect to the topology induced by d).

Remark 10. *Observe that in a LPL, $\ell^{-1}([0, \infty)) = M^2_{\ll}$ is open by lower semi-continuity. Note also that $\ell^+(x, y) = \ell(x, y)$ unless $\ell(x, y) = -\infty$ in which case $\ell^+(x, y) = 0$.*

Definition 11. *(Causal curve and ℓ-recifiability)*
On (M, d, ℓ), a non-constant causal path which is locally-d-Lipschitz continuous is called a causal curve. A timelike curve is a causal curve which is also a timelike path.
A causal curve σ satisfying $L_{\ell^-}(\sigma|_{[a,b]}) < 0$ for every $a < b, a, b \in A$ is called ℓ-rectifiable.
A causal curve is called ℓ^--minimizing or (ℓ-maximizing) if it minimizes L_{ℓ^-} (resp. maximizes L_ℓ) among all causal curves sharing it's endpoints.
An ℓ-curve is a causal curve $\sigma : [0,1] \to M$ with $L_\ell(\sigma) = \ell(\sigma(0), \sigma(1))$.

ℓ-rectifiability is usually called rectifiability and d-rectifiability and d-minimizing are defined as in Defintion 11 mutandis mutatis. Clearly, every ℓ-curve is ℓ-maximizing by triangle inequality.

Definition 12. *(Globally hyperbolic)*
Given a metric spacetime (M, ℓ, d) we say it is globally hyperbolic if

1. *It is non-totally imprisioning i.e for each compact set there exists a uniform upper-bound on the d-lengths of causal curves contained in the set.*
2. *$J^+(x) \cap J^-(y)$ is compact for every x, y.*

The set $J^+(x) \cap J^-(y)$ is denoted $J(x, y)$ and called a causal diamond.

The definition of $\mathcal{K}$-globally hyperbolic is similar to global hyperbolicity but requiring that $J^+(X) \cap J^-(Y)$ is compact for every pair of compact subsets X, Y of M instead of single points where $J^+(X)$ is just the union over elements in X of the sets $J^+(x)$. In the above setting the space is called causally curve connected if $x \leq y$, $x \neq y$ implies the existence of a causal curve joining x and y. It is called timelike curve-connected if $x \ll y$ implies the existence of a timelike curve with x and y as it's endpoints.

Definition 13. *(Geodesic, Lorentzian geodesic space and regularity)*
An $M^2_{\ll}$ geodesic space will be a metric spacetime where every pair of points in $M^2_{\ll}$ can be joined by an ℓ-curve.
A Lorentzian geodesic space is a LPLS in which any distinct points on $M^2_{\leq}$ can be joined by an ℓ-curve.
It is called regular if $L_\ell(\sigma) = \ell(\sigma(0), \sigma(1))$ implies there are no non-constant subsegments $B \subset A$ for σ which are lightlike ($L_\ell(\sigma|_B) = 0$).

In a regular metric spacetime, after reparametrization, ℓ-curves become timelike.

Lemma 14. *(McCann's automatic regularity of ℓ-paths)*
In regular Lorentzian geodesic spaces on which ℓ^+ is continuous and all causal diamonds are compact, each ℓ-path is d-continuous.

See [McCann2023, Lemma 5]. Evidently, to obtain d-continuity it is essential for ℓ-paths to not have non-constant null subsegments. Motivated by this Lemma, McCann made the folowing definition.

Definition 15. *(ℓ-geodesic)*
In a metric spacetime (M, d, ℓ) an ℓ-geodesic is a d-continuous ℓ-path.

Lemma 14 says that under a continuity property of ℓ and compactness of causal diamonds, every ℓ-path is an ℓ-geodesic.

Definition 16. *(Timelike Geodesics and Affine Timelike Geodesics)*
In a metric spacetime, we define the uniform metric on $C([0,1], M)$ to be

$$d^\infty(\sigma, \tilde{\sigma}) = \sup_{s\in[0,1]} d(\sigma(s), \tilde{\sigma}(s)). \tag{1.2}$$

We define $\mathrm{TGeo}^\ell(M) = \{\sigma \in C([0,1], M) : \sigma$ is an ℓ-path according to Definition 6$\}$ and by $\mathrm{Geo}^\ell(M)$ it's closure under d^∞.

A useful observation, needed for our study of the Markov property in section 1.12 is that the definition of $\mathrm{TGeo}^\ell(M)$ depends only on the topology induced by d. As a d^∞-closure, $\mathrm{Geo}^\ell(M)$ does depend on d^∞ and not only on the topology it induces.

Lemma 17. *(Characterization of globally hyperbolic Lorentzian length spaces)*
A globally hyperbolic metric spacetime (M, d, ℓ) is a Lorentzian length space if and only if

1. *It is timelike curve-connected,*
2. *it is $M^2_{\ll}$- geodesic space (as in Definition 13)*
3. *$I^\pm(x) \neq \emptyset$ for every $x \in M$,*
4. *$(\ell)^{-1}(-\infty)$ is open,*
5. *ℓ^+ is continuous and real-valued on M^2.*

See [McCann2023, Lemma 7].

Lemma 18. *(K-global hyperbolicity)*
Any globally hyperbolic Lorentzian length space is K-globally hyperbolic.

See [Cavalletti-Mondino, Lemma 1.5] or [McCann2023, Remark 12].
The seminal work of [Kunziger-Saemann] studied continuity of paths on metric spacetimes. The underlying topology induced by d seems a-priori an abstract imposition to the theory. It is reasonable to try to understand topologies within the framework of physical dynamics, that is, topologies that can be generated with only the knowledge of causal diamonds and time-separation functions. In [Kunziger-Saemann] the following topologies were studied:

Definition 19. *(Spacetime topologies)*

1. *The Alexandrov topology: Coarsest topology containing all diamonds $I^+(x) \cap I^-(y)$ for $x, y \in M$*
2. *The chronological topology: Coarsest topology containing all cones $I^{\pm}(x)$ for $x \in M$.*
3. *The metric topology: Topology generated by d.*

As explained in [Kunziger-Saemann], the topologies in LPS are ordered. A LPS is called strongly causal if the three topologies coincide. It has been shown that globally hyperbolic Lorentzian length spaces are strongly causal. If two different metrics are given in a globally hyperbolic non-totally imprisioning LLS, their topologies coincide by strong causality. As seen in [McCann2023], they become equi-globally hyperbolic in the sense that either both spaces are g.h. or none of them is. This means that the metrics are only relevant through the topology they generate and the set of d-rectifiable curves. Motivated by this property on rough paths, McCann defined our final object of interest.

Definition 20. *(Globally hyperbolic chrono-regular Lorentzian length space)*
A globally hyperbolic Lorentzian length-space (M, d, ℓ) on which (M, d) is complete and separable is called chrono-regular if all rough paths which are continuous with respect to the chronological topology and satisfy $L_\ell(\sigma) = \ell(\sigma(0), \sigma(1)) < 0$ contain no non-constant null segments.

In [McCann2023], the author realized that chrono-regularity satisfies the same property as above: given any two metrics on M the associated spacetimes are equi-chronoregular i.e. either both are chrono-regular or none of them is. We aim to study these spaces being careful of studying which properties correspond to the metric and which correspond to the topology they generate.

Remark 21. *It is well known that completeness is not a topological property but separability is. In the previous context, completeness is an assumption depending on the metric but separability is shared among all the metrics that define the same topology.*

Now let us recall the manifold structure from [McCann2019].

Definition 22. *(Spacetime)*
Let (M, g) be a smooth, connected, Hausdorff, time-oriented, Lorentzian manifold with signature $(+, -, ..., -)$, we know M is second countable and admits a Riemannian metric $\tilde{g}$. In this case we call (M, g) a spacetime.

A tangent vector is called timelike if $v^a g_{ab} v^b > 0$, spacelike if $v^a g_{ab} v^b < 0$ and null if equality holds. For $q \in (0, 1]$ we use the convex Lagrangian $L(x, v; q) = -(v^a g_{ab} v^b)^{q/2}/q$ with the convention

that $L(x,v;q)=\infty$ unless v is *future-directed*. Similarly we define the action of a curve $A(\sigma,q)$ as the integral (with respect to t) of $L(\sigma(t),\sigma'(t);q)$, and finally the q-Lorentz distance as

$$\ell(x,y)=-\inf\{A(\sigma;q):\sigma\in C^1([0,1],M),\sigma(0)=x,\sigma(1)=y\}. \tag{1.3}$$

In [McCann2019] it is proved that ℓ is independent of q and satisfies a reverse triangle inequality (it is a separation function). Global hyperbolicity ensure the infimum in (1.3) is attained.
We denote by $\mathcal{P}(M)$ the set of Borel measures on M and $\mathcal{P}_c(M)$ the subset of Borel measures with compact support. Analogous to the Euclidean and Riemannian version we define the ℓ_q distance between probability measures as

$$\ell_q(\mu,\nu)=\sup_{\pi\in\Gamma_{\leq}(\mu,\nu)}\left(\int\ell(x,y)^q d\pi(x,y)\right)^{1/q}$$

where $\Gamma_{\leq}(\mu,\nu)$ denotes the set of probability measures on $M\times M$ which are positive $\pi\geq 0$ and $\mathrm{spt}(\pi)\subseteq\ell^{-1}([0,\infty])$. A joint measure π is called ℓ^q-optimal if it satisfies the equality. Further, ℓ_q satisfies the reverse triangle inequality as proved in [McCann2019, Proposition 2.9].

Remark 23. *(Rewriting conditions on causality)*
By definition $(x,y)\in\ell^{-1}([0,\infty])$ if and only if $\ell(x,y)\geq 0$ which means that $y\in J^+(x)$ or in other words $(x,y)\in M^2_{\leq}$, hence $\mathrm{spt}(\pi)\subseteq\ell^{-1}([0,\infty])$ can be written (if $M^2_{\leq}$ is Borel) as $\pi(M^2_{\leq})=1$

Lemma 24. *In a ghcrlls (Definition 20), $M^2_{\leq}$ is d-Borel measurable.*

Proof. By assumption $\max\{\ell,0\}$ is continuous on M^2 and so $M^2_{\leq}=\ell^{-1}([0,\infty))$ is Borel. ∎

Further $M^2_{\leq}$ would be closed in a causally closed space.
Given a metric space (X,d) we will always denote by $\mathcal{P}(X)$ the space of Borel probability measures on X and by $M_+(X)$ the set of positive Borel measures on X. In the case where (X,τ) is just a topological space, $\mathcal{P}^\tau(X)$ denotes the set of Borel (with respect to τ) probability measures.
In the smooth spacetime case of above, where the manifold is time orientable, we say that a function $f:M\to\mathbb{R}$ is a causal functional if it is non-decreasing along any future-directed causal curve. [Eckstein-Miller, Theorem 5] states that in the case of smooth spacetimes, $(p,q)\in M^2_{\leq}$ if and only if $f(p)\leq f(q)$ for every smooth bounded causal functional f. The set of continuous causal functionals from (M,d,ℓ) is denoted $\mathcal{C}(M,\mathbb{R})$.

Definition 25. *(Causal order or measures via smooth causal functionals)*
If (M,ℓ,d) is a smooth spacetime, for $\mu,\nu\in\mathcal{P}(M)$ we say that $\mu\preceq\nu$ if for every smooth bounded causal functional f we have

$$\int_M f d\mu\leq\int_M f d\nu.$$

Observe that functionals refer to maps from M to $\mathbb{R}$ and are different to causal curves and paths. The reason behind defining causal functionals is to obtain a duality characterization of the ordering $\preceq$.
Following the conventions of [Leonard2014], we define entropy with respect to general measures and not only probability measures. Let $h:[0,\infty)\to\mathbb{R}$ be given by

$$h(x)=x\log(x)-x+1,\ \ h(0)=1$$

Observe that $h(x)\geq 0\ \forall x\in[0,\infty)$ and achieves it's minimum uniquely at $x=1$ with $h(1)=0$.

Definition 26. *(Relative Entropy for probability measures)*
In a metric space (M, d), given $\mu \in \mathcal{P}(M)$ and ν absolutely continuous with respect to μ we define the relative entropy of ν with respect to μ,

$$\mathrm{Ent}(\nu|\mu) = \int_M h\left(\frac{d\nu}{d\mu}\right) d\mu$$

whenever the integral exists, we define $\mathrm{Ent}(\nu|\mu) = \infty$ in any other case.

One can also modify Definition 26 to be ∞ if $\mathrm{spt}(\mu)$ is not compact (as in [McCann2023]). We do not make this assumption with the idea of incorporating more general measures.

Remark 27. *Note that by our use of the function h we obtain*

$$\mathrm{Ent}(\nu|\mu) = 0 \Leftrightarrow \mu = \nu$$

So far we have only defined $\mathrm{Ent}(\nu|\mu)$ in the case $\mu \in \mathcal{P}(M)$, we aim to define it for general positive measures $M_+(M)$. To this end, following [Leonard2014] we restrict the domain of definition of entropy.

Definition 28. *(Relative Entropy for positive measures)*
Let μ be a $\sigma-$finite measure on M and let $W : M \to [0, \infty)$ be such that

$$z_W := \int_M e^{-W(x)} d\mu(x) < \infty$$

For any $\nu \in \mathcal{P}(M)$ such that

$$\int_X W(x) d\nu(x) < \infty$$

we define the relative entropy of ν with respect to μ by the formula

$$\mathrm{Ent}(\nu|\mu) = \mathrm{Ent}(\nu|z_W^{-1} e^{-W}\mu) - \int_M W\, d\nu - \log(z_W)$$

In [Leonard2014, Appendix A] it is shown that such W always exists and the definition of entropy is independent of the choice of W.
For simplicity, we will mostly focus on the case where the measures are in fact probability measures but we will comment on the adaptations required to generalize to positive measures when necessary. The case of positive measures is justified by the study of Wiener's measure (see [Leonard2014], [Tamanini]) although we will restrict to the simpler case of probabilities because of the nature of the framework of Lorentzian spaces.

Lemma 29. *(Leonard's topological version of Gibb's duality)*
Let μ be a σ-finite measure on a topological space (Y, τ) endowed with it's Borel σ-algebra, assume there exists a Borel measurable function $W : Y \to [-\infty, \infty)$ such that

$$\int_Y e^{-W} d\mu < \infty$$

then for every π Borel (w.r.t. τ) probability measure which is a.c. with respect to μ one has

$$\begin{aligned}\mathrm{Ent}(\pi \mid \mu) &= \sup\left\{\int_Y f d\pi - \log\left(\int_Y e^f d\mu\right) : f : Y \to [-\infty, \infty), \int_Y e^f d\mu < \infty\right\} \\ &= \sup_{u \in C_W}\left\{\int_Y u d\pi - \log\left(\int_Y e^f d\mu\right)\right\}\end{aligned}$$

where $C_W = \{u \in C(Y, \mathbb{R}) : \sup|u|/W < \infty\}$.

For a proof see [Leonard2014, A.5,A.6]. Observe that Lemma 29 immediately yields lower semi-continuity of entropy with respect to the weak convergence of continuous bounded functions on Y. This fact is essential to our study of the Schrödinger problem and it's formulation in topological spaces (rather than only Polish spaces) will be relevant in section 1.3.6.

Remark 30. *The topological space in Lemma 29 not necessarily being a polish space will be fundamental in problem (τ−RDSCH).*

In most of the cases we will use μ as a probability measure on which the Lemma 29 simplifies and recovers it's more known form, set $W = 1$ in the previous case to get:

Lemma 31. *(Leonard's topological version of Gibb's duality for probability measures)*
Let μ be a probability measure on a topological space (Y, τ) endowed with it's Borel σ-algebra, then for every π Borel (w.r.t. τ) probability measure which is a.c. with respect to μ one has

$$\begin{aligned}\mathrm{Ent}(\pi \mid \mu) &= \sup\left\{\int_Y f d\pi - \log\left(\int_Y e^f d\mu\right) : f : Y \to [-\infty, \infty), \int_Y e^f d\mu < \infty\right\} \\ &= \sup_{u \in C_b}\left\{\int_Y u d\pi - \log\left(\int_Y e^f d\mu\right)\right\}\end{aligned}$$

where $C_b = \{u \in C(Y, \mathbb{R}) : \sup|u| < \infty\}$ the set of continuous bounded functions from Y to $\mathbb{R}$.

1.2.2 The Benamou-Brenier formula for Lorentzian Manifolds

In the context of [McCann2019] we can obtain an analogue of the Benamou-Brenier formula [Benamou-Brenier] which in the euclidean case with respect to quadratic cost states that the Wasserstein 2-distance can be alternatively computed as the minimization of kinetic or dissipation energy among all solutions of the continuity equation.

Theorem 32. *(Lorentzian Benamou-Brenier)*
Consider a smooth spacetime M as in Definition 22. Take $\mu_0, \mu_1 \in \mathcal{P}_c(M)$ and assume that $\mathrm{spt}(\mu_0 \times \mu) \subseteq \{\ell > 0\}$ (or more generally (μ_0, μ_1) are q-separated [McCann2019, Definition 4.1]), then

$$\ell(\mu_0, \mu_1)^q = \sup\left\{\int_0^1 \int_M \frac{1}{q}(g_{ab}(x) v_t^a v_t^b)^{q/2} d\mu_t(x) dt\right\} \tag{1.4}$$

where the supremum is taken over all pairs (v_t, μ_t) satisfying

1. $\bigcup_{0 \le t \le 1} \mathrm{spt}(\mu_t)$ *is bounded,*

2. (v_t, μ_t) *satisfy the continuity equation*

$$\partial_t \mu_t + \nabla_M \cdot (v_t \mu_t) = 0, \tag{1.5}$$

where $\nabla_M \cdot$ *denotes divergence with respect to the Lorentzian connection,*

3. $\mu_t|_{t=0} = \mu_0, \mu_t|_{t=1} = \mu_1$,

4. $v \in L^q(\mu_t dt)$,

5. $t \to \mu_t$ *is a.c.*

6. $v_t(\cdot)$ *is future-directed and* $g_{ab} v_t^a v_t^b > 0$.

Proof. Given that (M, d, ℓ) is a smooth spacetime, by [DeLellis, Theorems 1.3,1.4,1.7] it costs no generality to assume that v_t, μ_t solves the continuity equation (1.5) and that

$$\frac{dT_t}{dt} = v_t(T_t) \tag{1.6}$$

can be solved in classical sense. In this case, let $\mu_t = T_t \# \mu_0$ and one can see that if $\dot{T}_t$ exists, it is future-directed and $g_{ab}(\dot{T}^a, \dot{T}^b) > 0$, then

$$\ell^q(\mu_0, \mu_1) \geq \int \ell^q(x, y) d(Id \times T_1)\mu_0 = -\int_0^1 \int L\left(\frac{dT_t}{dt}\right) d\mu_t dt$$

where $L(v) = -(g_{ab} v^a v^b)^{q/2}/q$ if v is future directed, $g_{ab} v^a v^b > 0$ and ∞ otherwise. Concluding that

$$\ell^q(\mu_0, \mu_1) \geq \sup_{(v_t, \mu_t) \in \mathrm{V_{sm}}} \int_0^1 \int_M \frac{1}{q} (g_{ab} v_t^a v_t^b)^{q/2} d\mu_t dt \tag{1.7}$$

where $\mathrm{V_{sm}}$ correspond to the pairs (v_t, μ_t) satisfying conditions (1.5) and (1.6). To obtain equality, let T be the unique map of [McCann2019, Theorem 5.8], i.e.

$$T_t(x) = \exp_x(tDH(D\overline{u}(x), x, q)) \tag{1.8}$$

define $v_t = \left(\frac{dT_t}{dt}\right) \circ T_t^{-1}$ then $(v_t, T_t \# \mu_0)$ achieves the equality. ∎

1.2.3 Static and dynamic problems

The main object of study of this work is an adaptation of the "Schrödinger problem" from Polish spaces ([Leonard2014], [Leonard2001],[Leonard2012] and references therein) and $RCD^*(K, N)$ spaces ([Tamanini],[Gigli-Tamanini]) to the framework of Lorentzian length spaces of [McCann2019],[McCann2023] and [Kunziger-Saemann]. In this section we define the relativistic static Schrödinger problem (RSch) in short, on which we think about the Schrödinger problem as in the static formulation of [Leonard2014] but we add the constraint of the target/final measure being supported in the causal future of the initial measure. This constraint makes the relativistic problem physical.

Definition 33. *(Relativistic Static Schrödinger Problem)*
Let (M, ℓ, d) be a globally hyperbolic chrono-regular Lorentzian length space (as described in Definition 20), we define the Relativistic Schrödinger problem with respect to a fixed reference measure $r \in \mathcal{P}(M^2)$ as the minimization problem of entropy along causal transference plans, that is, given $\mu_0, \mu_1 \in \mathcal{P}(M)$,

$$\min_{\pi \in \Gamma_{\leq}(\mu_0, \mu_1)} \mathrm{Ent}(\pi | r) \tag{RSch}$$

where $\Gamma_{\leq}(\mu_0, \mu_1)$ denotes the set of Borel probability measures $\pi \in \mathcal{P}(M^2)$ with $\mathrm{spt}(\pi) \subseteq M^2_{\leq}$ and whose marginals are μ_0 and μ_1 respectively i.e.

$$\Gamma_{\leq}(\mu_0, \mu_1) = \{\pi \in \mathcal{P}(M^2) : \pi(M^2_{\leq}) = 1, \mathrm{Proj}_1 \#\pi = \mu_0, \mathrm{Proj}_2 \#\pi = \mu_1\} \tag{1.9}$$

where for $(x, y) \in M^2$, $\mathrm{Proj}_1(x, y) = x$ and $\mathrm{Proj}_2(x, y) = y$.

Remark 34. *Note that so far (RSch) is only defined for $r \in \mathcal{P}(M^2)$ and that this problem only differs from the case of polish spaces through the physical condition $\pi \in \Gamma_{\leq}(\mu_0, \mu_1)$. We'll see how this condition is immaterial in terms of existence and uniqueness of the minimizers in Propositions 38, 41 and 44.*

Definition 35. *(Relativistic Dynamical Schrödinger Problem)*
Let (M, d, ℓ) be a globally hyperbolic chrono-regular Lorentzian length space (as described in Definition 20), let Ω be the set of continuous causal paths, that is

$$\Omega = \{\sigma \in C([0, 1], M) : \sigma(s) \preceq \sigma(t), \forall t, s \in [0, 1], s \leq t\}$$

We define the dynamical relativistic Schrödinger problem associated to $R \in \mathcal{P}(\Omega)$ as the minimization problem

$$\min_{\substack{\Pi \in \mathcal{P}(\Omega) \\ e_0 \#\Pi = \mu_0, e_1 \#\Pi = \mu_1}} \mathrm{Ent}(\Pi | R) \tag{RDSch}$$

where e_t denotes the evaluation map i.e. $e_t(\gamma) = \gamma(t)$ and $\mu_0, \mu_1 \in \mathcal{P}(M)$ are given and $\mathcal{P}(\Omega)$ is the set of Borel-probability measures when Ω is endowed with d^∞ (from (1.2)).

Remark 36. *The choice of d^∞ is natural in the context of abstract metric spaces (and the common assumption on the polish space case) but does not incorporate our physical understanding of causal curves for which we will discuss different topologies in section 1.3.6. Note that when Ω is endowed with the topology of d_∞ every evaluation map e_t is Borel and there is no parametrization restriction.*

Remark 37. *In (RDSch) we defined the relativistic Schrödinger problem for a reference probability measure ($R \in \mathcal{P}(\Omega)$). This definition is simpler but disallows the familiar case of the Wiener measure with respect to Lebegue, in the general case we want to consider*

$$\hat{R}(\cdot) = \int \mathcal{W}^x(\cdot) dx$$

where $\mathcal{W}^x$ is the Wiener measure started at x. Because this expression is a positive measure in Ω the seminal work of [Leonard2014] is based on the description of the problem in the set of positive measures $M^+(\Omega)$. Just as in Definition 28 we will work on the case of probability measures but describe the modifications required for the $M^+(\Omega)$ case.

Proposition 38. *(Uniqueness of solutions)*
Problem(RSch) *admits at most one minimizer.*

Proof. If the existence of a minimizer is assumed, note that the objective function is strictly convex by definition of entropy via integral of a strictly convex function and $\Gamma_{\leq}(\mu,\nu)$ is convex as $t\pi_1(M^2_{\leq}) + (1-t)\pi_2(M^2_{\leq}) = 1$. ∎

One of the main tools in optimal transportation is the push-forward of measures, naturally the study of the Schrödinger problem uses extensively the following result.

Lemma 39. *(Entropy and Push-forward)*
Assume that $\mu,\nu \in \mathcal{P}(X)$ where X is a polish space, then for any measurable function f we have

$$\mathrm{Ent}(f\#\mu|f\#\nu) \leq \mathrm{Ent}(\mu|\nu). \tag{1.10}$$

This result is a consequence of convexity and Jensen's inequality. We will have a detailed version in Lemma 133.

Proposition 40. *(Necessity)*
In a ghcrLLS (definition 20) (M,ℓ) it is necessary that

$$\max\{\mathrm{Ent}(\mu_0|\mathrm{Proj}_1\,\#r), \mathrm{Ent}(\mu_1|\mathrm{Proj}_2\,\#r)\} < \infty \tag{1.11}$$

for RSch to have a solution.

Proof. Direct consequence of Lemma 39, as if the hypothesis were not true, then every plan in $\Gamma_{\leq}(\mu_0,\mu_1)$ has infinite entropy. ∎

The following propositions are direct adaptations of the results of [Leonard2014].

Proposition 41. *(Equi-existence)*
If $(e_0,e_1)\#R = r$ then the minimization values of RSch and RDSch are equal i.e.

$$\min_{\pi\in\Gamma_{\leq}(\mu_0,\mu_1)} \mathrm{Ent}(\pi|r) = \min_{\substack{\Pi\in\mathcal{P}(\Omega)\\ e_0\#\Pi=\mu_0, e_1\#\Pi=\mu_1}} \mathrm{Ent}(\Pi|R)$$

Furthermore, the problems have solutions if and only if either of the two values of minimization is finite or equivalently if and only any of the problems is feasible with an element of finite entropy.

We delay the proof of Proposition 41.

Proposition 42. *(Leonard's sufficient conditions for existence for the static problem in ghcrLLS)*
Let (M,ℓ) be a globally hyperbolic chrono-regular Lorentzian length space whose underlying metric space (M,d) is Polish and suppose that $\mu_0 \preceq \mu_1$ and $r \in \mathcal{P}(\Omega)$ and $\Gamma_{\leq}(\mu_0,\mu_1) \neq \emptyset$, if $M^2_{\leq}$ is σ-compact then there exists a solution for RSch.

Proof. By strict convexity of the functional, it is enough to show that $\Gamma_{\leq}(\mu_0,\mu_1)$ is uniformly tight. By Polishness of (M,d) we know that $\{\pi \in \mathcal{P}(M^2) : \pi_0 = \mu_0, \pi_1 = \mu_1\}$ is tight so it is enough to show that $\Gamma_{\leq}(\mu_0,\mu_1)$ is closed with respect to weak convergence but this is the assumption of σ-compactness of $M^2_{\leq}$. ∎

Note that in the smooth spacetime case of condition 7 from [Eckstein-Miller, Theorem 8] ensures that if $\mu_0 \preceq \mu_1$ (as in Definition 25) then $\Gamma_{\leq}(\mu_0, \mu_1) \neq \emptyset$ so the problem is feasible. Because the function h is strictly convex, it is enough to verify that the set of constraints are closed (with respect to weak convergence) and convex. Note also that from Remark 23 $\pi_1, \pi_2 \in \Gamma_{\leq}(\mu_0, \mu_1)$ then $\pi_1(M^2_{\leq}) = 1 = \pi_2(M^2_{\leq})$ from which convexity follows.

Remark 43. *The assumption of σ-compactness is not too restrictive. In [Eckstein-Miller, Theorem 4] it is shown that every smooth spacetime satisfies this restriction.*
Another approach to existence can be defining RSch only for chronological measures, i.e. $\pi(M^2_{\ll}) = 1$ as this set is always open by Remark 10 and so any weak limit of measures in the set satisfies $\pi(M^2_{\leq}) = 1$ by Prokhorov's Theorem.

Proposition 44. *(Existence and Uniqueness for Dynamical problem)*
Let (M, ℓ, d) be a ghcrll space (as in Definition 20) if $\mu_0 \preceq \mu_1$ then RDSch admits a unique solution.

Proof. Convexity of the set of constraints is direct in this case, to observe that the set of constraints is not empty, define for A in the Borel-σ-algebra of continuous paths,

$$\Pi(A) := \int_{M\times M} \delta_{\gamma_{x,y}}(A) d\mu_0 \times \mu_1(x, y)$$

where $\gamma_{x,y}$ denotes an ℓ-curve joining x and y (given that the space is timelike curve connected) and $\delta_{(\cdot)}$ denotes the Dirac delta. It follows that Π has the correct marginals and belongs to $\mathcal{P}(\Omega)$. ∎

Remark 45. *Recall that the assumption of curve-connectedness can be obtained directly from the existence of an ℓ-path as every ℓ-path becomes a ℓ-curve after a continuous increasing reparametrization by [McCann2023, Lemma 7].*

Proposition 46. *(Relation between Static and Dynamic problems)*
Given a solution π for RSch with reference measure $r := (e_0, e_1)\#R$, we obtain a solution for RDSch with reference measure R via the formula

$$\Pi(A) = \int_{M^2} R^{x,y}(A) d\pi(x, y) \tag{1.12}$$

where $R^{x,y}$ denotes the $x - y$ bridge for R, defined by $R^{x,y}(\cdot) = R(\cdot | e_0 = x, e_1 = y)$ the regular conditional probability for R. Conversely, given a solution Π for (RDSch) with reference measure R we obtain a solution for (RSch) with reference measure $(e_0, e_1)\#R$ by projecting into endpoints i.e.

$$\pi = (e_0, e_1)_{\#}\Pi$$

Proof. Following the proof in [Leonard2014], to show optimality, disintegrate π as follows, because (M, d) is a Polish space, for any given Y polish and for any measurable function $\phi : M \times M \to Y$

$$\int_{M\times M} f(x, y) d\pi(x, y) = \int_{M\times M} \int_{\phi^{-1}(z)} f(z) d\pi_z(x, y) d\phi_{\#}\pi(z)$$

then note that entropy is additive with respect to disintegration by (133) which combined with $\mathrm{Ent}(\phi_{\#}\pi \mid \phi_{\#}R) \leq \mathrm{Ent}(\pi \mid R)$ yields the result.

$$\mathrm{Ent}(\Pi|R) = \mathrm{Ent}(\pi|r) + \int_{M^2} \mathrm{Ent}(\Pi^{x,y}|R^{x,y}) d\pi(x, y)$$

where $R^{x.y} = R(\cdot | X_0 = x, X_1 = y)$ the R bridge which means that Π and R share bridges as it was shown by Léonard in [Leonard2014] and also in [Leonard2014b]. ∎

1.2.4 Duality and convergence of the static relativistic Schrödinger problem

Proposition 47. *(Causal duality of the static problem)*
Let (M,ℓ) be a ghcrlls, suppose $\mu_0 \preceq \mu_1$ with $\mu_0,\mu_1 \in \mathcal{P}_{ac}(M)$ and $r \in \mathcal{P}(M^2)$ is such that $r(M^2_{\leq}) = 1$ and $\Gamma_{\leq}(\mu_0,\mu_1) \neq \emptyset$ then

$$\min_{\pi\in\Gamma_{\leq}(\mu_0,\mu_1)} \mathrm{Ent}(\pi|r) = \sup_{\substack{(\phi,\psi)\in C_b(M)^2 \\ \phi\leq^r\psi}} \left\{ \int_M \phi d\mu_0 + \int_M \psi d\mu_1 - \log \int_{M^2} e^{\phi+\psi} dr(x,y) \right\}$$

where $C_b(M)$ is the set of continuous bounded functions on M and $\leq^r$ denotes,

$$\phi \leq^r \psi \Leftrightarrow \phi(x)+\psi(y) = 0 \text{ unless } x \leq y, \ r\text{-a.e.} \tag{1.13}$$

In the general case where $r \in M^+(M^2)$, we replace the continuous bounded functions for

$$C_B = \{f \in C(M,\mathbb{R}) : \sup|f|/B < \infty\}$$

where $B : M \to \mathbb{R}$ (assumed to exist for now) satisfies

$$\int B d\mu_0 < \infty, \int B d\mu_1 < \infty, \int e^{-(B\oplus B)} dr < \infty \tag{1.14}$$

Proof. The constraint $\Gamma_{\leq}(\mu_0,\mu_1)$ can be replaced by $\Pi(\mu_0,\mu_1)$ the set of probability measures on the product space with μ_0 and μ_1 as marginals because $\pi \ll r$ and $r(M^2_{\leq}) = 1$ imply that $\pi(M^2_{\leq}) = 1$.The relative entropy is defined to be $+\infty$ outside of $\mathcal{P}_{ac,r} = \{\pi : \pi \ll r\} \neq \emptyset$ (as $\mu_0 \times \mu_1$ belongs to this set), so the infimum is finite and thus we can restrict to $\mathcal{P}_{ac,r}$. This observation allows us to apply the duality result [Leonard2001b, Theorem 3.4], given that RSch depends only on the topology of the spacetime and on the metric only through the reference measure r. The condition (1.13) is forced by $r(M^2_{\leq}) = 1$, if not the right-handside would be ∞. ∎

Remark 48. *The statement of Proposition 47 is written for the case where $r \in M^+(M^2)$. The existence of such B is immaterial in the case where $r \in \mathcal{P}(M^2)$ or has finite measure as in this case any constant function satisfies (1.14) and hence C_B is just continuous bounded functions. We prefer this statement for Proposition 47 in accordance with Remark 37. In such case $r(M^2_{\leq}) = 1$ should be replaced with $r(M^2 \setminus M^2_{\leq}) = 0$ where the measure could be infinite in $M^2_{\leq}$.*

The following Lemma is a simple calculation that elucidates the general idea and approach of section 1.4. In order to obtain a desired cost function in a limit of entropic regularizations, we must construct measures converging in exponential rate to that cost.

Lemma 49. *(Static computation for general cost)*
Let (M,d,ℓ) be a ghcrlls, then for every lower-semicontinuous function $c : M \times M \to \mathbb{R}$ for given $r \in \mathcal{P}(M^2), r(M^2_{\leq}) = 1$ and $\epsilon > 0$ we have

$$\inf_{\pi\in\Gamma_{\leq}(\mu_0,\mu_1)} \left\{ \int_{M\times M} c(x,y) d\pi(x,y) + \epsilon\, \mathrm{Ent}(\pi|r) \right\} = \epsilon \inf_{\pi\in\Gamma_{\leq}(\mu_0,\mu_1)} \mathrm{Ent}(\pi|r_c^\epsilon) \tag{1.15}$$

where $r_c^\epsilon = e^{-c(x,y)/\epsilon} r$ in the sense of measures.

Proof. The proof is a direct computation,

$$\int_{M\times M} c(x,y)d\pi(x,y) + \epsilon \int_{M\times M} \log\left(\frac{d\pi}{dr}\right) d\pi(x,y) = \epsilon \int_{M\times M} \log\left(e^{c(x,y)/\epsilon}\frac{d\pi}{dr}\right) d\pi$$
$$= \epsilon \int_{M\times M} \log\left(\frac{d\pi}{dr_c^\epsilon} d\pi\right),$$

as minimization on both sides yields the result. ∎

Remark 50. *Observe that the left hand side of (1.15) should converge to the solution of optimal transport with respect to c as $\epsilon \to 0$. On the other hand, the right hand-side of (1.15) corresponds to a relativistic Schrödinger problem with respect to reference measure r_c^ϵ. Our goal will be to study the features of the dynamical version of (1.15) together with the convergence of it's minimizing arguments. We will do so in Theorem 149. Note that, to obtain a Lorentzian cost as in [McCann2019], one would set $c(x,y) = -\ell_q(x,y)$ which will motivate (1.137).*

1.2.5 Low temperature entropic limit for the static problem

Proposition 51. *(Static slow down and the limit as optimal transport)*
Assume that $r \in \mathcal{P}(M^2)$ is such that $r(M^2_{\leq}) = 1$, let $\mu_0 \preceq \mu_1$ and that $\mu_0 \times \mu_1 \ll R$, let $c : M\times M \to \mathbb{R}$ be lower semi-continuous, bounded below and suppose $\mu_0 \times \mu_1$ has finite total c-cost, i.e.

$$\int_{M\times M} c(x,y)d\mu_0 \times \mu_1(x,y) < \infty.$$

Assume that $M^2_{\leq}$ is closed and let π_ϵ be the unique solution for the Schrödinger (RSch) problem with respect to reference measure r_c^ϵ given by

$$dr_c^\epsilon(x,y) = e^{-c(x,y)/\epsilon} dr(x,y),$$

then there exists a measure π and a sub-sequence $\epsilon_n \to 0$ such that as $n \to \infty$

$$\pi_{\epsilon_n} \rightharpoonup \pi.$$

where $\pi \in \mathcal{P}(\Omega)$ solves the optimal transport problem

$$\inf_{\pi\in\Gamma_{\leq}(\mu_0,\mu_1)} \int_{M\times M} c(x,y)d\pi(x,y)$$

Proof. By the assumption of σ-compactness of $M^2_{\leq}$, the family $\{\pi^\epsilon\}_{\epsilon>0}$ is uniformly tight. Using Prokhorov's theorem there exist a weak sub-sequential limit $\pi \in \mathcal{P}(M^2)$, by Remark 10 $\pi(M^2_{\leq}) = 1$. Observe that by lower semi-continuity of entropy,

$$\mathrm{Ent}(\pi|r) \leq \liminf_{\epsilon_n\to 0} \mathrm{Ent}(\pi^\epsilon|r).$$

And so if $\tilde{\pi} \in \Gamma_{\leq}(\mu_0,\mu_1) \cap \{\pi : \pi \ll r\}$, for every $\epsilon > 0$ optimality of π_ϵ yields

$$\int_{M\times M} c(x,y)d\pi_\epsilon + \epsilon\,\mathrm{Ent}(\pi_\epsilon|r) \leq \int_{M\times M} c(x,y)d\tilde{\pi} + \epsilon\,\mathrm{Ent}(\tilde{\pi}|r),$$

where we have used Lemma 49. As $0 \leq \mathrm{Ent}(\pi^\epsilon|r) < \infty$,

$$\begin{aligned}\int_{M\times M} c(x,y)d\pi_\epsilon &\leq \int_{M\times M} c(x,y)d\tilde{\pi} + \epsilon\left(\mathrm{Ent}(\tilde{\pi}|r) - \mathrm{Ent}(\pi_\epsilon|r)\right)\\ &\leq \int_{M\times M} c(x,y)d\tilde{\pi} + \epsilon\left(\mathrm{Ent}(\tilde{\pi}|r)\right).\end{aligned} \tag{1.16}$$

Taking the limit as $\epsilon \to 0$ on both sides, due to lower-semicontinuity of c, weak convergences establishes

$$\int_{M\times M} c(x,y)d\pi(x,y) \leq \liminf_{\epsilon\to 0}\int_{M\times M} c(x,y)d\pi_\epsilon(x,y) \leq \int_{M\times M} c(x,y)d\tilde{\pi}(x,y).$$

∎

Remark 52. *Proposition 51 is a relatively direct statement, nevertheless it captures the main ideas of convergence of the Schrödinger problem (on the primal side). One can also study convergence with careful analysis of the dual problem (see [Leonard2014], [Tamanini]). The main reason to study* (RDSch) *on primal side only is the absence of a canonical elliptic heat semigroup. In the polish (and $RCD(K,N)$) case, the study of the convergence of the dual side relies on regularity properties of the Schrödinger potentials (maximizing arguments of the dual) inherited from the heat kernel.*
Up to this point we have dealt with the static and the dynamic problem simultaneously but the analogue of Proposition 51 for the dynamical setting corresponds to controlling probability of large events (see section 1.4 on Large Deviation Principles) which will occupy most of the focus of the present work.

1.3 Bridge spaces and Markovianity

So far in our analysis of the Relativistic Schrödinger problems (RSch and RDSch) the situation has not been very different to the case where the underlying space is a Polish space, an $RDC(K,N)$ space or a Riemannian manifold. In this classical setting, one usually restricts to the case of a Markovian reference measure. Markovianity and it's known connections with diffusion semigroups allow one to develop the theory of regularity and convergence of Schrödinger potentials (see [Tamanini]). Furthermore, the large deviation principles satisfied by the laws of Brownian bridges ([Hsu1990]) and Brownian motion (see [Varadhan], [Hsu]) yield convergence in slowed-down limit to optimal transference plans (with respect to cost induced by their rate function). With this in mind, we start our exploration of Markov-like measures in pre-length spaces satisfying large deviation principles.
The natural measure to consider in Riemannian manifolds and $RCD(K,N)$ spaces is the law associated to the diffusion generated by the heat kernel. Ellipticity and regularity principles of the heat semigroup imply convergence properties of Schrödinger potentials (see [Leonard2014], [Tamanini]). The absence of an elliptic heat kernel in Lorentzian manifolds raises the question of what the natural bridge-measure to use as reference in (RDSch) should be.
The study of generalizations of Brownian motion to Lorentzian manifolds is an active area of research. Following the seminal paper of [Dudley1966] (and contemporaneous [Hakim]), many researchers have successfully studied Brownian-like paths in Lorentzian manifolds (see [Franchi-LeJan2007], [Franchi], [Chevalier-Debbasch], [Dunkel-Hanggi] and references therein). As we aim to find analogues of bridge measures adapted to RSch and RDSch, our exploration starts a little differently. In section

1.3.1 we first study a bridge construction proposed to the author by Dr. McCann in Minkowski space with the property that we have enough control in "parameter" space to induce large deviation principles. We observe that this bridge construction generalizes the idea of Levy's first construction of Brownian bridges.
In sections 1.3.2 and 1.3.3 we study the original construction of Dudley in Minkowski space and compare it to this new construction. We explain the problems of the large deviation principles of these processes. Dudley's process exists naturally in phase-space which opens the questions for investigation of the Schrödinger problem there. We address the phase-space problem in section 1.3.4 and show some connections to the original Schrodinger problem. Finally, Dudley's theorem on Lorentz-invariant Markov processes on Minkowski space is addressed by studying bridge spaces taking into account the topology of causal curves in section 1.3.6.

1.3.1 Construction of Brownian-bridge-Like processes

Schilder's original theorem on the large deviation principle satisfied by the laws of slowed-down Brownian bridges is a fundamental result in the analysis of diffusions. In [Hsu1990], the author is able to exploit reversibility and other properties of the heat semigroup on Riemannian manifolds to obtain a large deviation principle for Riemannian Brownian bridges. The non-existence theorem of Dudley (Theorem 71 below) relies on the non-compactness of the hyperboloid; in short, this lack of compactness of the group prohibits the existence of a Lorentz-invariant finite Borel measure on the group (see section 1.3.2 for details). The idea of this section is to create a specific process in every causal diamond to exploit the compactness ensured by global hyperbolicity, i.e. given that in a globally hyperbolic Lorentzian manifold every causal diamond is compact, we will use this compactness to create path measures that allow us to control a parametric limit (resembling noise in the usual Brownian diffusion and $1/b$ in the context of [Bismut]). We observe that this parametric limit is not the slow-down parameter in the Riemannian case [Hsu1990] but we explain how to use this construction to get deviation principles.

A Levy-like construction on Minkowski Space

In this section, we refer to $\mathbb{R}^{1,n}$ with signature convention $(+,-,\cdots,-)$ as Minkowski space $\mathbb{M}^n$ following Dudley's work ([Dudley1966],[Dudley1967],[Dudley1973]).
To begin we focus on the case $n=3$ and we let $\ell : \mathbb{M}^3 \times \mathbb{M}^3$ to be defined via

$$\ell((t_1,z_1),(t_2,z_2)) = \left((t_2-t_1)^2 - ||z_2-z_1||_2^2\right)^{1/2} \tag{1.17}$$

if the argument of the square root is positive, where $||\cdot||_2$ refers to Euclidean norm in $\mathbb{R}^3$ matching the conventions on [McCann2019].

Definition 53. *(ℓ-mid set)*
Let $x, y \in \mathbb{M}^n$ chronologically ordered, i.e. $x \ll y$ we define the ℓ-mid set between x and y via

$$\mathrm{MID}(x,y) = \{x \leq z \leq y : \ \ell(x,z) = \ell(z,y)\}.$$

Notice that $\mathrm{MID}(x,y)$ consists only of points in the causal future of x and in the causal past of y whose proper time is the same, physically it corresponds to achievable points z on which the proper time elapsed from x to z equates the proper time from z to y. It takes the same (proper) time to get from x to z than it takes to get from z to y.

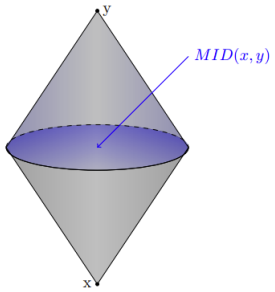

Figure 1.3: $MID(x,y)$ in $J^+(x) \cap J^-(y)$.

Proposition 54. *(Compactness of* $\mathrm{MID}(x,y)$*)*
Let $x \ll y$ *be fixed, then* $\mathrm{MID}(x,y)$ *is compact.*

Proof. Observe that $\mathrm{MID}(x,y) \subseteq J^+(x) \cap J^-(y)$ so it is enough to show it is closed (by global hyperbolicity). By Definition 53, $\mathrm{MID}(x,y)$ is closed when $\ell(x,\cdot)$ and $\ell(\cdot,y)$ are both continuous. ∎

Remark 55. *Observe that we have written the previous proof using the general notation of the first section. The aforementioned continuity is direct in* $\mathbb{M}^3$ *as seen from* (1.17).

Proposition 56. *(Distance from a* $\mathrm{MID}(a,b)$*)*
Assume $z \in \mathrm{MID}(a,b)$ *where* $a \ll b$ *then*

$$||z-a||_2 \leq ||a-b||_2 \tag{1.18}$$

where again $||\cdot||_2$ *refers to Euclidean distance in* $\mathbb{M}^n$.

Proof. In $\mathbb{M}^1$ the conclusion is an application of Cauchy-Schwartz inequality in equation (1.83) below. Note that if $s \in [0,1]$ parametrizes the mid-set between (t_1,x_1) and (t_2,x_2), let s be such that $z = M_s$ in such parametrization, then

$$d(a,M_s) = \sqrt{(1/2 - s + s^2)\, d(a,b)^2 + (1-2s)(x_2-x_1)(t_1-t_2)}. \tag{1.19}$$

Using Cauchy-Schwartz, since $s \in [0,1]$ we get that

$$d(a,M_s) \leq d(a,b)\sqrt{1-2s+s^2} = d(a,b)|1-s| \leq d(a,b). \tag{1.20}$$

For the general case in $\mathbb{M}^n$, the result follows from looking at a 2-dimensional plane and applying the $n=2$ case: consider $a,b \in \mathbb{M}^n$ and $z \in \mathrm{MID}(a,b)$, the span of $\{z-a, b-a\}$ is isometric to $\mathbb{M}^2$ (which can be seen by boosting) and $||b-a||_2$ corresponds to the diagonal in the rectangle so (1.18) follows. ∎

Construction 1. *(Levy-like construction on* $\mathbb{M}^n$ *with deterministic restriction on* $\mathrm{MID}(x,y)$*)*
Assume that $x \ll y$ *have been fixed, consider for any points* $a \leq b$ *where* $a,b \in J^+(x) \cap J^-(y)$ *a collection of measures* $\sigma(a,b) \in \mathcal{P}(\mathrm{MID}(a,b))$.

i.) Define ${}_0P_0 = x$ *and* ${}_0P_1 = y$.

ii.) *Let* ${}_1P_0 = x$, ${}_1P_1 \sim \sigma(x, y)$ *and* ${}_1P_2 = y$.

iii.) *Let* ${}_2P_0 = x$, ${}_2P_1 \sim \sigma({}_1P_0, {}_1P_1)$, ${}_2P_2 = {}_1P_1$, ${}_2P_3 \sim \sigma({}_1P_1, {}_1P_2)$ *and* ${}_2P_4 = y$,

and continue inductively, namely

$$ {}_nP_m = \begin{cases} {}_nP_{2k} = {}_{n-1}P_k & \text{if } m = 2k \\ {}_nP_{2k+1} \sim \sigma({}_{n-1}P_k, {}_{n-1}P_{k+1}) & \text{if } m = 2k+1. \end{cases} \tag{1.21} $$

The formula 1.21 indicates that we randomly chose a point in $\mathrm{MID}(x, y)$ *and use it as base point of a causal diamond at the next level. The prescript indicates the level, i.e.* ${}_nP_m$ *is the* m*-the element of the construction generated at level* n*. At level* n *we obtain* 2^n *points including our original* x *and* y *as endpoints,* ${}_0P_m = x, {}_{2^n}P_n = y$ *for every level* n.

We aim to find a probability measure on the space of continuous paths from $[0, 1] \to J^+(x) \cap J^-(y)$ such that

$$ X(i/2^n) \sim {}_iP_n. $$

We find general conditions for the collection of measures $\sigma(a, b)$ in the following propositions, first we depict the process (in $n = 2$ for simplicity) in Figure 1.4:

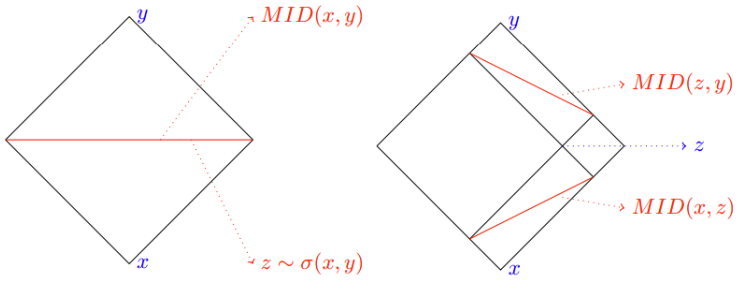

Figure 1.4: First 2 steps of the construction depicted in $\mathbb{R}^{1,1}$. Repeat the process in every new causal diamond once elements in the mid-set have been randomly chosen.

Although the notation of (1.21) is not very enlightening, a good way to visualize the process is to note the dependence of the random variables, similar to Levy's construction of Brownian Bridge, we are finding intermediate points at each level which will correspond to the dyadic times. Our goal for the section is to generate a path-measure using Construction 1 from section 1.3.1 . Not every collection of probabilities $\sigma(a, b)$ where $a, b \in J^+(x) \cap J^-(y)$ would yield measures concentrated on continuous paths. We start by making the simplified assumption of contracting every mid-set to avoid giving positive mass to light-like related points. Observe that when $J^+(a) \cap J^-(b)$ consists on a unique light-like geodesic, that is $\ell(a, b) = 0$, then every point in the light-like segment between a and b belongs to $\mathrm{MID}(a, b)$ because $\ell(a, \gamma(t)) = 0 = \ell(\gamma(t), b)$ for all proper times t. This phenomena of the light-like related points will stop us from having continuity of limiting paths with respect to

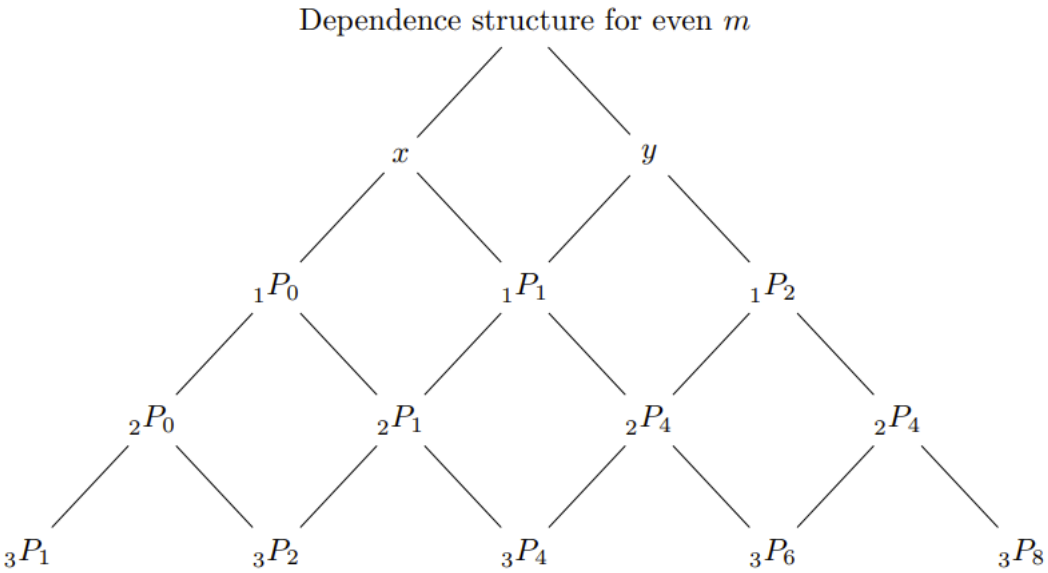

Figure 1.5: Dependence structure for even m up to level 3, for odd m the element corresponds to an element on the previous level. The edges in the drawing indicate from which mid-set we are choosing the new element.

the underlying topology: the euclidean distance can be very big while the proper time is 0. We depict this observation figure 1.6 in $\mathbb{R}^{1,1}$.

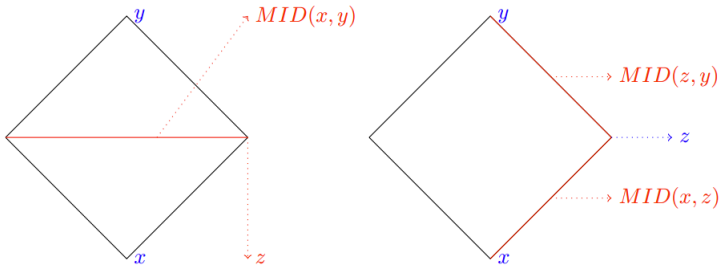

Figure 1.6: In the light-like case, the mid-set is all the points in the light-like geodesic

To start the analysis of collections of measures yielding path-measures concentrated on continuous paths, we start by controlling uniformly the distances between sequential elements on Construction 1 from section 1.3.1 . We start by studying only measures which are completely concentrated on contracting the mid sets by a uniform constant.

Definition 57. *(c-Contracted Mid set)*
Let $0 < c < 1$ be fixed and let $a \ll b$ where $a, b \in J^+(x) \cap J^-(y)$, we define the c-contracted mid set between a and b via

$$c - \mathrm{MID}(a,b) = \left\{ a \leq z \leq b : z = c\left(\tilde{z} - \left(\frac{b+a}{2}\right)\right) + \frac{b+a}{2} \text{ for some } \tilde{z} \in \mathrm{MID}(a,b) \right\}. \tag{1.22}$$

Definition 57 uses the specific nature of $\mathbb{M}^n$ as we are writing the contracted set by translating to the origin, scaling by c and translating back. The fact that the contracting constant c is uniform, allows us to control distances uniformly.

Proposition 58. *(Uniform control on distances to $\mathrm{MID}(a,b)$)*
Let $0 < c < 1$, $a \ll b$ and assume $z \in c - \mathrm{MID}(a,b)$ then we have

$$\max\{||z-a||_2, ||b-z||_2\} \leq \left(\frac{1}{2} + \frac{c}{2}\right) ||b-a||_2 \tag{1.23}$$

Furthermore, with the notation of Construction 1 from section 1.3.1 , for any $i \in \{0, 1, \ldots, 2^n - 1\}$

$$||{}_nP_i - {}_nP_{i+1}||_2 \leq \left(\frac{1+c}{2}\right)^n ||x-y||_2 \tag{1.24}$$

Proof. By symmetry of $c - \mathrm{MID}(a,b)$ it is enough to show the inequality for the distance between z and a. By definition of $c - \mathrm{MID}(a,b)$ let $\tilde{z}$ as in (57), then

$$\begin{aligned} ||z-a||_2 &= ||c(\tilde{z} - (b+a)/2) + (b+a)/2 - a||_2 \\ &= ||c(\tilde{z}-a) + (1-c)((a+b)/2 - a)||_2 \\ &\leq c||\tilde{z}-a||_2 + \frac{(1-c)}{2}||b-a||_2 \\ &\leq \left(\frac{1}{2} + \frac{c}{2}\right)||b-a||_2 \end{aligned} \tag{1.25}$$

where the last inequality is due to Proposition 56. In the notation of Construction 1 from section 1.3.1 if i is even then ${}_nP_i \in c - \mathrm{MID}({}_{n-1}P_{i/2}, {}_{n-1}P_{i/2+1})$ so the conclusion (1.24) follows by applying the previous inequality recursively. ∎

Lemma 59. *(Uniform distance control on causal diamonds in Minkowski space)*
Let $x, y \in \mathbb{M}^n$ and $x \ll y$ there exists $C_{x,y} > 0$ such that for all $a, b \in J^+(x) \cap J^-(y)$ with $a \ll b$ we have

$$\sup_{z_1, z_2 \in J^+(a) \cap J^-(b)} ||z_1 - z_2||_2 \leq C_{x,y}||b-a||_2 \tag{1.26}$$

Proof. Consider $x \ll y$ and $a, b \in J^+(x) \cap J^-(y)$ if $z_1, z_2 \in J^+(a) \cap J^-(b)$ in Minkowski space, let us show that $||z_1 - z_2||_2 \leq 2 \cdot ||b-a||_2$, so that $C_{x,y} = 2$ for all $x, y \in \mathbb{M}^n$. Let us consider global coordinates on $\mathbb{M}^n$ of the form (t, x) where $x \in \mathbb{R}^n$. If $z \in J^+(a) \cap J^-(b)$, then

$$(t_z - t_a)^2 \geq ||x_z - x_a||_2^2, \tag{1.27}$$

$$(t_b - t_z)^2 \geq ||x_b - x_z||_2^2. \tag{1.28}$$

Consequently,

$$\begin{aligned}||b-a||_2^2 =&(t_b-t_a)^2+||x_b-x_a||_2^2\\ &=(t_b-t_z)^2+(t_z-t_a)^2+2(t_b-t_z)(t_z-t_a)\\ &+||x_b-x_z||_2^2+||x_z-x_a||_2^2+2(x_b-x_z)\cdot(x_z-x_a)\\ &=||z-a||_2^2\\ &+(t_b-t_z)^2+||x_b-x_z||_2^2+2\left\{(t_b-t_z)(t_z-t_a)+(x_b-x_a)\cdot(x_z-x_a)\right\}\end{aligned} \tag{1.29}$$

By Cauchy-Schwartz, the sum of terms in (1.29) is non-negative, from which we conclude that if $z\in J^+(a)\cap J^-(b)$

$$||z-a||_2\leq||b-a||_2. \tag{1.30}$$

Given any points $z_1,z_2\in J^+(a)\cap J^-(b)$ we have

$$||z_1-z_2||_2\leq||z_1-a||+||a-z_2||\leq 2||b-a|| \tag{1.31}$$

where we used (1.30) twice in the last step. Taking the supremumum yields (1.26). ∎

Towards the developments in section 1.3.5, let us reformulate (1.26) with the notation of general spacetimes, (M,d,ℓ).

Definition 60. *(Uniform control on causal diamonds for general spacetimes)*
Let (M,ℓ,d) be a ghcrlls (definition 20), we say (M,d,ℓ) satisfies uniform control on causal diamonds if whenever $x,y\in M$, $x\ll y$ there exists $C_{x,y}>0$ such that for all $a,b\in J^+(x)\cap J^-(y)$ with $a\ll b$ we have

$$\sup_{z_1,z_2\in J^+(a)\cap J^-(b)} d(z_1,z_2)\leq C_{x,y}d(a,b). \tag{1.32}$$

Remark 61. *Lemma 59 says Minkowski space satisfies uniform control on causal diamonds (Definition 60). It is not clear whether Definition 60 holds in all ghcrlls.*

Proposition 62. *(Existence for measures concentrated on contracted mid-sets)*
The Levy-like construction from 1 yields a unique Borel (with respect to d^∞) probability measure on $\Omega_{x,y}=\{\gamma\in\mathcal{C}([0,1],M):\gamma(0)=x,\gamma(1)=y\}$ for every collection $\{\sigma_{a,b}\}_{a\leq b,a,b\in J^+(x)\cap J^-(y)}$. satisfying that for fixed $0<c<1$,

$$\sigma_{a,b}\left(\mathrm{MID}(a,b)\setminus c-\mathrm{MID}(a,b)\right)=0 \text{ for all } a,b\in J^+(x)\cap J^-(y). \tag{1.33}$$

Proof. We aim to define a Borel random variable X on the space $(\Omega_{x,y},d^\infty)$. We start by defining X on dyadic times and argue the existence of a unique limit for every converging sequence of dyadic times. If $n\in N$ and $k\in\{0,1,\ldots,2^n\}$ we set $X(k/2^n)={}_nP_k$ from Construction 1 from section 1.3.1 , to ensure the existence of $\mu\in\mathcal{P}(\Omega_{x,y})$, for any sequence of dyadic times converging to $t\in[0,1]$, define

$$X(t)=\bigcap_{n=1}^{\infty}J^+(X(\lfloor 2^nt\rfloor))\cap J^-(X(\lfloor 2^nt\rfloor+2^{-n})).$$

which in our notation rewrites

$$X(t) = \bigcap_{n=1}^{\infty} J^+({}_nP_i) \cap J^-({}_nP_{i+1}). \tag{1.34}$$

Using first (1.26) and then (1.24) we see

$$\begin{aligned} diam(J^+(X(\lfloor 2^n t \rfloor)) \cap J^-(X(\lfloor 2^n t \rfloor + 2^{-n}))) &\leq C_{x,y} ||{}_nP_i - {}_nP_{i+1}||_2 \\ &\leq C_{x,y} \left(\frac{1}{2} + \frac{c}{2}\right)^n ||y - x||_2 \end{aligned} \tag{1.35}$$

so that the diameters go uniformly to 0 and $X(t)$ is unique and well-defined by Cantor's theorem because every element in the intersection is compact by global hyperbolicity and nested by causality. Using Kolmogorov's theorem [Durrett, Chapter 7 Theorem 1.3] there exists a law $\mu \in \mathcal{P}(\mathcal{D}, \sigma(\{\{e_t\}_{t\in D^n}\}_{n\in\mathbb{N}})$ and (1.35) ensures it assigns measure 1 to uniformly continuous curves, hence the extension via (1.34) is well-defined as the map is measurable with respect to $(C([0,1], M), \sigma(e_t))$ and so there exists $\mu \in \mathcal{P}(\Omega_{x,y})$ such that $X \sim \mu$ as desired. ∎

The equivalence on the σ-algebras generated by $\{e_t\}_{t\in[0,1]}$ and d^∞ is explained in Theorem 125. Condition (1.33) means that the probability measures on the mid-sets don't give positive probability to elements close to null futures and pasts of a, b respectively. We use the uniform bound c but will argue that it is not necessary. We show a more general condition in Construction 2.

Remark 63. *Although continuity and the existence of a path measure could be proved in simpler ways (because of the linear structure on $\mathbb{M}^n$), we choose this method as it relies only on the nestedness and compactness of causal diamonds, together with the equalities (1.24) and (1.26) which can easily be formulated in more general frameworks (like ghcrlls from Definition 20).*

Remark 64. *Note that always $X(0) = x$ and $X(1) = y$. Consequently, $X(1/2) \sim \sigma(x, y)$ where everything is deterministic as x and y are fixed. The realization of $X(1/4)$ depends on the observed value of $X(1/2)$, so we know the conditional distribution*

$$X(1/4) \mid (X(1/2) = z) \sim \sigma(x, z) \tag{1.36}$$

So far, we have used the notation $\sigma(x, y)$ to describe a probability measure in $\mathrm{MID}(x, y)$. For notational purpose, on the following formula we write $\sigma_{x,y}$ instead of $\sigma(x, y)$ so we can rewrite (1.36) as

$$\mathbb{P}(X(1/4) \in A) = \int \int_A d\sigma_{x,z} d\sigma_{x,y}(z)$$

where we have now used the parenthesis to indicate the variable of integration. Analogously, if m is odd

$$\mathbb{P}(X(m/2^n) \in A \mid X((m-1)/2^n), X((m+1)/2^n)) = \sigma_{X((m-1)/2^n), X((m-1)/2^n)}(A).$$

Remark 65. *(Comparison with Levy's construction of Brownian motion in $[0, 1]$)*
A famous construction of Levy for Brownian motion is to set Gaussian measures at dyadic times in $[0, 1]$ (see [Durrett, Section 7.1, Theorems 1.3,1.3]). By consistency and Kolmogorov's theorem one obtains a measure which is later shown to have almost surely continuous paths. We observe that Construction 1 from section 1.3.1 exactly replicates this idea. Global hyperbolicity allows us to think

of $J^+(x) \cap J^-(y)$ as $[0,1]$ and the mid-sets in our construction correspond to vertical lines at dyadic times. The weight of each Gaussian on Levy's construction is proportional to the dyadic level which is emulated here by $\sigma_{a,b}$ being a probability measure on $\mathrm{MID}(a,b)$.

Remark 66. *We haven't assumed the existence of any temporal functions and we are parametrizing curves on $J^+(x) \cap J^-(y)$ with a external parameter t. This parameter is not proper time and we will discuss it's unphysicality throughout the rest of the work.*

The restriction (1.33) is not necessary. It was certainly helpful for the inequality (1.24) which yielded the convergence to zero of diameters of causal diamonds.

Construction 2. *(Levy-like construction with probabilistic restriction on $\mathrm{MID}(x,y)$)*
Let as before $\{\sigma(a,b)\}_{a,b \in J^+(x) \cap J^-(y)}$ be a collection of measures where $\sigma(a,b) \in \mathcal{P}(\mathrm{MID}(a,b))$. Define ${}_nP_k$ recursively as in Construction 1 from section 1.3.1 , assume further that

$$\lim_{s_n, t \in \mathcal{D}, s_n \to t} d(X_s, X_t) = 0 \tag{1.37}$$

uniformly on $t \in \mathcal{D}$ where $\mathcal{D}$ denotes the set of endpoints of dyadic intervals on $[0,1]$ and

$$X(i/2^n) \sim {}_iP_n, \tag{1.38}$$

then there exists a measure $\mu \in \mathcal{P}(\Omega_{x,y})$ such that if $X \sim \mu$ then $X(k/2^n) \sim {}_nP_k$.

Remark 67. *The uniformity condition is on $t \in [0,1]$ and the modulus of continuity may depend on the randomness ω.*

Construction 3. *(Levy-like)*
Let as before $\{\sigma(a,b)\}_{a,b \in J^+(x) \cap J^-(y)}$ be a collection of measures where $\sigma(a,b) \in \mathcal{P}(\mathrm{MID}(a,b))$. Define ${}_nP_k$ recursively as in Construction 1 from section 1.3.1 , assume further that

$$\lim_{s_n, t \in \mathcal{D}, s_n \to t} d(X_s, X_t) = 0 \tag{1.39}$$

for all $t \in \mathcal{D}$ where $\mathcal{D}$ denotes the set of endpoints of dyadic intervals on $[0,1]$ then there exists a measure $\mu \in \mathcal{P}(\Omega_{x,y})$ such that if $X \sim \mu$ then $X(k/2^n) \sim {}_nP_k$.

Construction 4. *(Levy-like)*
Let as before $\{\sigma(a,b)\}_{a,b \in J^+(x) \cap J^-(y)}$ be a collection of measures where $\sigma(a,b) \in \mathcal{P}(\mathrm{MID}(a,b))$. Define X_t recursively as in Construction 1 from section 1.3.1 for $t \in \mathcal{D}$, assume further that there exists $\alpha, \beta > 0$

$$\mathbf{E}[d(X_s, X_t)^\beta] \leq K|t-s|^{1+\alpha} \tag{1.40}$$

for all $s,t \in \mathcal{D}$ where $\mathcal{D}$ denotes the set of endpoints of dyadic intervals on $[0,1]$ then there exists a measure $\mu \in \mathcal{P}(\Omega_{x,y})$ such that if $X \sim \mu$ then $X(k/2^n) \sim {}_nP_k$.

Note that (1.40) ensures the hypothesis of Kolmogorov-Centsov and so the probability measure on dyadics of Kolmogorov's extension Theorem assigns probability 1 to uniformly continuous paths and so can be extended to $(\Omega_{x,y}, d^\infty)$ by the afore-mentioned measurability of the map that takes uniformly continuous maps on $\mathcal{D}$ to $\Omega_{x,y}$ (see [Durrett, Section 7.1, Theorem 1.4]).

Note that the proof of Proposition 62 works for all of these constructions, as we have set the hypothesis so that we can replicate it. As the construction is only done for uniformly (in time) continuous functions, the probability measure lifts from the cylindrical σ-algebra $(\mathcal{D}, Cyl(\mathcal{D})$ to $(C[0,1], \mathcal{B})$ where $\mathcal{B}$ is the σ-algebra generated by finite-dimensional evaluation maps. Further, observe that we have changed notation in Construction 2 of $||\cdot||_2$ to $d(\cdot,\cdot)$ as we are thinking of the formulation in more general frameworks (see 1.3.5).

Remark 68. *The condition of hypothesis* (1.37) *avoids the concentration of mass on light-like paths, as it was first observed by Dudley on his seminal work [Dudley1966] mass on light-like related points gives processes concentrated on lower-dimensional subsets (see Figure 1.6).*

The constructions 1 and 2 are relatively simple as the collection of measures $\{\sigma(a,b)\}$ is fairly general. The main point is that this simple construction can be carried out in much more general frameworks as we will detail in section 1.3.5. We will see that the constructions 1 and 2 are well adapted to the study of the abstract Schrödinger problem. Before we study these properties we need to study the known processes in Minkowski space $\mathbb{M}^n$ and general Lorentzian manifolds that resemble the behaviour of Brownian Bridges. Constructions 1 and 2 are not completely satisfactory as generalizations of Brownian Bridges (as we will see in the next chapters) but are suitable tools for large deviations principles (see section 1.4) which are deeply connected to the Schrödinger problem. The previous observation indicates one thing: to construct bridges in Lorentzian spaces we should focus on satisfying the uniform continuity condition with respect to the underlying distance d as we will do in Construction 1.3.5.

1.3.2 Dudley's random motions in Minkowski space

The seminal work of Richard Dudley [Dudley1966] started the investigation of stochastic Markov processes in Minkowski space $\mathbb{M}^3$. Dudley's work [Dudley1966], [Dudley1967] and [Dudley1973] started a fundamental area of research for mathematical physics. Dudley dealt with the concept of random motions (stochastic processes) in special relativity invariant with respect to the Lorentz group $\mathcal{L}$. In this work, the author showed the non-existence of certain $\mathcal{L}$-invariant stochastic processes. In essence, Dudley showed causality in Minkowski space is just a Lipschitz condition on space variables with respect to time variable which is incompatible with Markovianity. The fact that motions must have "speed" bounded by the speed of light enforces an almost-everywhere differentiability condition which is incompatible with the standard notion of strong Markovianity: if the past and the future are to be independent, how could the process be differentiable?
This incompatibility lead the author to study random motions where instant velocities are Brownian-like. In the next sections we study Dudley's observations to relate them to our setting. We defer the generalizations of Dudley's work to more general curved geometries for section 1.3.4 where a vast literature exists ([Franchi-LeJan2007],[Chevalier-Debbasch],[Dunkel-Hanggi]).
The reason to study Minkowskian space first is twofold: first, in order to study bridge constructions in ghcrlls spaces, we should understand the case of special relativity, then generalize to manifolds and only then to the non-smooth setting; second, just as Riemannian manifolds can be embedded in high dimensional euclidean spaces, smooth Lorentzian manfiolds can be embedded in Minkowskian spaces.

Dudley's observation in Minkowski Space and the physical Markov property

We start by presenting the details on [Dudley1966] on random motions on $\mathbb{M}^3$. It is important that we clearly state the construction together with it's assumptions (i.e. on choices of measurable spaces) as these subtleties are where we can take advantage of the structure of more general frameworks. Let us recall the construction on [Dudley1966].
Assume that a curve $t \in [a,b] \to \tilde{\gamma}(t)$ is parametrized with respect to the time variable in $\mathbb{M}^3$, i.e.

$$\gamma(t) = (t, \tilde{\gamma}(t)).$$

Causality in $\mathbb{M}^3$ means that for $t_1 \geq t_2$

$$(t_1 - t_2)^2 - ||\tilde{\gamma}_1(t_1) - \tilde{\gamma_2}(t_2)||_2^2 \geq 0.$$

Equivalently, for every $t_1 \geq t_2$

$$||\tilde{\gamma}(t_1) - \tilde{\gamma}(t_2)||_2 \leq |t_1 - t_2|. \tag{1.41}$$

Equation (1.41) is Lipschitz continuity of the curve when parametrized with the time variable. From this observation, set $\mathcal{A}$ to be the set of Lipschitz functions with constant 1, whose derivative is strictly bounded by 1 whenever it's defined,

$$\mathcal{A} = \{f : [0,\infty) \to \mathbb{R}^3 : ||f(t) - f(s)||_2 \leq |t-s|, ||f'(s)||_2 < 1 \text{ when it exists}\}.$$

Definition 69. *We say that a collection of measures $\{\mathbb{P}_x\}_{x\in\mathbb{R}^3}$ on a σ-algebra $\mathcal{S}$ of subsets of $\mathcal{A}$ are starting probabilities on $\mathbb{M}^3$ "itself" if for every x*

$$\mathbb{P}_x(\{f \in \mathcal{A} : f(0) = x\}) = 1. \tag{1.42}$$

Let $\mathbb{H}^3$ denote the hyperboloid, we say that a collection of measures $\{\mathbb{P}_x^v\}_{x\in\mathbb{R}^3, v\in\mathbb{H}^3}$ are starting probabilities (on phase-space) if

$$\mathbb{P}_x^v(\{f \in \mathcal{A} : f(0) = x, f'(0^+) = v\}) = 1 \tag{1.43}$$

where $f'(0+)$ corresponds to the derivative from the right.

Given $\{\mathbb{P}_x^v\}$ starting probabilities, a first Markov property can be formulated on $\mathbb{M}^3$. Observe that the use of $\mathcal{A}$ (Lipschitz functions with speed strictly smaller than one) corresponds to restricting to causal curves $(t, f(t))$ which do not have null segments of positive volume. The use of the parameter t instead of proper-time is a choice made by the author and justified only after the analysis of strong Markov properties (Markov property with respect to random time changes). The use of the domain $[0,\infty)$ is justified through the existence of global coordinates on $\mathbb{M}^3$.
To follow Dudley's notation, set $e_t(f) = f(t)$ for $t \in [0,\infty)$ and

$$\begin{aligned}\mathcal{B}_t^s &= \sigma(\{\{f \in \mathcal{A} : f(r) \in A\} \text{ where } s \leq r \leq t, A \in \mathcal{B}(\mathbb{R}^3)\}) \\ &= \sigma(\{e_r^{-1}(A)\} \text{ where } s \leq r \leq t, A \in \mathcal{B}(\mathbb{R}^3)\}) \\ &= \sigma(\{e_r\}_{s\leq r\leq t})\end{aligned} \tag{1.44}$$

Still following [Dudley1966], set the future-from s $\sigma-$algebra to be

$$\mathcal{B}^s = \sigma(\{\mathcal{B}_t^s\}_{t>s}) \tag{1.45}$$

where $\sigma(\cdot)$ refers to the σ-algebra generated by that collection.

Definition 70. *(Markov Process with starting probabilities in Minkowski Space)*
Let $\{\mathbb{P}_x^v\}$ be starting probabilities (on phase-space) we say that the collection corresponds to a Markov Process in $\mathbb{M}^3$ if for every $t > 0, x \in \mathbb{R}^3, v \in \mathbb{H}^3$

$$\mathbb{P}_x^v(\{f : f'(t) \text{ exists } \}) = 1$$

and for any $A \in \mathcal{B}^t(\mathcal{A})$ we have

$$\mathbb{P}_x^v(A|\mathcal{B}_t^0(\mathcal{A})) = \mathbb{P}_{f(t)}^{f'(t)}(\theta_t A)$$

where θ_t denotes translation by t, $\theta_t(f)(x) = f(x+t)$.

We say that a process is trivial if it is concentrated in a single function $f \in \mathcal{A}$. In [Dudley1966, Theorem 11.1] the author proved the following theorem:

Theorem 71. *(Dudley's non-existence)*
Let $\mathcal{J} = \{f : [0,\infty) \to \mathbb{M}^3 : |f(t) - f(s)|_M \leq |t-s|\}$. There is no non-trivial process with Lorentz-invariant with starting probabilities on $\mathbb{M}^n$ itself i.e. on $(\mathcal{J}, \mathbb{M}^3, \mathcal{B}(\mathbb{M}^3))$, every process satisfying

1. *(Translation invariance)*
   For $t_i \geq 0, z \in \mathbb{M}^3, A_i \in \mathcal{B}(\mathbb{M}^3)$

$$\mathbb{P}_0(\{f : f(t_i) \in A_i, i = 1, \ldots, n\}) = \mathbb{P}_z(\{f : f(t_i) \in A_i + z, i = 1, \ldots, n\}) \tag{1.46}$$

2. *($\mathcal{L}$-invariance)*
   For any $L \in \mathcal{L}$ and $A \in \mathcal{B}^0(\mathcal{J})$

$$\mathbb{P}_0(A) = \mathbb{P}_0(L(A)) \tag{1.47}$$

   must be concentrated on a single function on $\mathcal{J}$.

For a proof see [Dudley1966, Theorem 11.3], intuitively, the existence of such a process would yield a finite $\mathcal{L}$-invariant Borel probability measure, which is impossible by non-compactness of $\mathcal{L}$.

Corollary 72. *No $\mathcal{L}$-invariant Markov process (according to Definition 70) exists with starting probabilities $\mathbb{M}^3$ itself (as in Definition 69).*

Proof. The non-existence is independent of the Markov property, Definition 70 requires the existence of starting probabilities on M itself according to Definition 69 but Theorem 71 prevents that. ∎

Remark 73. *Theorem 71 shows the non-existence of Markov processes whose starting probabilities are $\mathcal{L}$-invariant on $\mathbb{M}^3$. Because our focus is the study of the Schrödinger problem, our main goal is to study stochastic processes whose small time equivalences we can control. In the Euclidean and Riemannian settings, the Wiener measure (together with Brownian motion and Brownian Bridges) satisfy both Markov properties **and** large deviation principles. A completely satisfactory generalization to Lorentzian manifolds would correspond to a Markovian process with $\mathcal{L}$-invariance and small time asymptotics approximating those of geodesic flow. Theorem 71 shows such generalization is impossible when the (strong) Markov property is interpreted in the classical sense of [Dudley1966]. We will therefore explore the asymptotics of several bridge constructions (like 1 and 2), their application to the Schrödinger problem and different ways to study the Markov property taking the physically inherent properties of the underlying space into account.*

A look into Theorem 71 and Corollary 72 reveals it's dependence on the choices of classes of functions and σ-algebras chosen. Although these choices are well justified, they open the question whether or not there are other possible ways to formulate Markov Properties well-adapted to the study of RSch and RDSch. The seminal work of C. Léonard [Leonard2014] introduces an altervative formulation of the Markov property than that of [Dudley1966] (Definition 70). In the next section we present this definition and connect it to Dudley's.

1.3.3 The extrinsic (un-physical) Markov Property

We aim to use tools from optimal transportation as in [Leonard2014] so it is important to make the distinction that the external parameter we will use for causal curves is not (a priori) related to the time variable in a space time. For example, in the definition of q-geodesics of [McCann2019], if $\{\mu_s\}_{s\in[0,1]}$ interpolates ℓ^q-optimally between causally related μ_0 and μ_1, the parameter s does not represent physical time. We use this idea from transportation theory to adapt the definition of Markovianity from [Leonard2014] and we call it the un-physical Markov property.

Definition 74. *(Markov process on causal bridges)*
On a complete, separable metric spacetime (M,ℓ,d), denote by

$$\Omega_{x,y} := \{\gamma \in \mathcal{C}([0,1],M), \gamma(0)=x, \gamma(y)=1\}$$

let $\nu \in \mathcal{P}(\Omega_{x,y})$ the set of Borel probability measures on $\Omega_{x,y}$, with the topology induced by d^∞. Denote $\nu_0 := e_0\#\nu$, following [Leonard2014] we say that ν is Markov in terms of causal bridges if ν is conditionable with respect to $\{e_t\}$ and for every $t\in[0,1]$ one has

$$\nu(e_{[0,t]} \in A, e_{[t,1]} \in B \mid e_t) = \nu(e_{[0,t]} \in A \mid e_t)\nu(e_{[t,1]} \mid e_t), \tag{1.48}$$

where e_t denotes the evaluation process, $e_t(\gamma)=\gamma(t)$ and where $A \in \sigma(\{e_s : 0\le s\le t\})$ and $B\in\sigma(\{e_s : t\le s\le 1\})$.

If ν is conditionable then equation (1.48) is equivalent to

$$\nu(e_{[t,1]} \in A | e_{[0,t]}) = \nu(e_{[t,1]} \in A | e_t). \tag{1.49}$$

The Markov property aims to encapsulate the idea that past and future are independent *given the present.*

Remark 75. *Note that in order for equation (1.48) to make sense, one needs $(e_{[0,t]})^{-1}(A)\in\mathcal{B}(\Omega_{x,y})$. This measurability of $t\to e_t$ is usually implicitly assumed in definitions of Markov properties and has remarkable consequences (see section 1.3.9 for details). This implicit assumption is deeply connected to the un-physicality of this definition as we will see in section 1.3.6.*

Remark 76. *In contrast to Definition 70, Definition 74 is not defined for functions parametrized in $[0,\infty)$ but rather only $[0,1]$, this choice is natural as our intended application is for the Schrödinger problem on which μ_0 and μ_1 are fixed. Curves satisfying $\gamma(0)=x$ and $\gamma(1)=y$ are called bridges between x,y or $(x,y,1)-$bridges.*

Definition 74 depends on the Borel σ-algebra generated by d^∞, we will show that even if this is a natural way to lift d from M to $\Omega_{x,y}$ it presents an (a-priori) seeming un-physicality that we aim to resolve in section 1.3.6. Although the Markov property and it's consequences have been studied extensively, our particular interest towards the study of the Schrödinger problem in general non-smooth Lorentzian spaces is the content of the following proposition:

Proposition 77. *(The Causal Markov property is also preserved)*
Suppose the reference measure $R \in \mathcal{P}(\Omega_{x,y})$ is Markov in terms of causal paths (Definition 74), then the solution to (RDSch) with respect to R is also Markov in terms of causal paths.

We postpone the proof of Proposition 77 and delay it for section 1.3.6. The idea is to follow [Leonard2014] and apply the disintegration theorem in the correct space (see section 1.3.10). We aim to study Markovian bridges as object themselves rather than constructing them as conditioned Markov processes. The theory of Markov Bridges as conditioned Markov Processes can be found in [Fitzsimmon-Pitman-Yor] and some of it's results are presented in section 1.3.8. Let us first study the constructions 1 and 2 in terms of Markovianity.

Theorem 78. *(Markov and Levy-like construction 1)*
The Levy-like bridge construction (Construction 1 from section 1.3.1) satisfies the following property: let $s, r, t \in \mathcal{D}$ and $n(s), n(r), n(t)$ their corresponding dyadic level, then if $\max\{n(t), n(s)\} < n(r)$ then

$$P(e_t \in A, e_s \in B | e_r) = P(e_t \in A | e_r) P(e_s \in B | e_r)$$

But in general, the measure associated to the Construction 1 from section 1.3.1 may not be Markov in the sense of Definition 74.

Proof. In Construction 1 from section 1.3.1 , observe that if $s \in \mathcal{D}$ then the law of X_s depends only on $X_{\lfloor 2^{n-1} s \rfloor}$ and $X_{\lfloor 2^{n-1} s+1 \rfloor}$, proceeding inductively we know X_s only depends on X_0 and X_1 which are considered in the $\sigma-$algebra and (78) follows.
For the latter, observe that if $s = 1/4, r = 3/8, t = 1/2$ then independently of X_r we know $X_s \sim \sigma_{x, X_{1/2}} \in \mathcal{P}(\mathrm{MID}(x, X_{1/2}))$ which depends on $X_{1/2}$ independently of X_r so (1.48) may not hold. ∎

Remark 79. *Although Constructions 1 and 2 may not be Markov, Theorem 78 indicates that they are almost Markov. An idea to change the constructions to make them Markov is presented in section 1.5.2.*

Theorem 71 is both intuitive and surprising: It is intuitive as the existence of velocities is incompatible with the classical understanding of the Markov property but surprising as it is not clear what the analogue of the standard objects should be anymore. Towards this end, Dudley proposed the study of a process in phase-space (the space and it's velocities) which arises naturally as a generalization of the Brownian motion in $\mathbb{M}^3$. We will study this process (referred to as Dudley's process), it's generalizations and properties. The fact that Dudley's process "lives" in phase-space rather than in the original space forces us to study a different version of the Schrödinger problem, "the Relativistic Kinetic Schrödinger Problem" RKSchP whose Euclidean analogue has been studied in [Chiarini-Conforti-Greco].

On the canonical relativistic Brownian Motion on Lorentzian Manifolds

On Riemannian manifolds Brownian motion can be defined as the unique diffusion process generated by the Laplace-Beltrami operator (Δ_M) via it's horizontal lift $\Delta_{O(M)}$ (see [Hsu, Chapters 3,4]). In the case of Lorentzian manifolds there is no clear analogue as the Laplace-Beltrami operator is hyperbolic and not elliptic.
The approach of [Dudley1966] in Minkowski Space and the extended in [Franchi-LeJan2007] is to study a stochastic process in phase space $(X_t, \dot{X}_t)$, a subset of the tangent bundle on which velocities

are constrained to be timelike in order to stay within the light cone, i.e. velocities are restricted to not surpass the speed of light.
This approach is similar to the use of the vertical Laplacian as infinitesimal generator on Sub-Riemannian manifolds.

There are many equivalent ways to define a diffusion process for an elliptic operator. We say that a stochastic process X_t corresponds to the diffusion of an elliptic operator L if X_t is the solution to the martingale problem for L, i.e. X_t is such that for every continuous bounded function f the process

$$f(X_t) - f(X_0) - \int_0^t Lf(X_s)ds \tag{1.50}$$

is a local martingale. The idea of [Dudley1966] and [Franchi-LeJan2007] is to solve a s.d.e. on the orthonormal frame after restricting velocities. In this section we study this approach and it's relation to Definition 74.

Construction of Dudley's relativistic diffusion

Following [Franchi-LeJan2007] we define Dudley's process in $\mathbb{M}^n$ via it's infinitesimal generator.

Definition 80. *(Dudley's process in phase-space of Minkowski spacetime)*
Let $(X_t^\sigma, \xi_t^\sigma)_{t\in[0,\infty)} \in \mathbb{M}^n \times \mathbb{H}^n$ be the solution of the martingale problem associated to the generator

$$L^\sigma f(x,p) = \sum_{k=0}^{d} p_i \frac{\partial f}{\partial x_i}(x,p) + \frac{\sigma^2}{2}\Delta_{(p)}^{\mathbb{H}^d} f(x,p) \tag{1.51}$$

where $\Delta_{(p)}^{\mathbb{H}^d}$ denotes the hyperbolic Laplacian (on $\mathbb{H}^d$). $(X_t^\sigma, \xi_t^\sigma)_{t\in[0,\infty)}$ is called Dudley's process.

Existence of the process was originally showed by Dudley in [Dudley1966, Section 6] and explained by Hakim in [Hakim, 1]. The fact that the process described above corresponds to the solution to the martingale problem in this definition is [Franchi-LeJan2012, Theorem VII.6]. The operator on (1.51) is hypoelliptic on $\mathbb{R}^{1,n} \times \mathbb{H}^n$ and so by Hörmander's theorem admits a smooth transition kernel.
Further, ξ_t is a Riemannian Brownian motion (in Mallavin-Ells-Elworthy sense) in hyperbolic space $\mathbb{H}^n$ and $X_t = \int_0^t \xi_s ds$, i.e.

$$(X_t, \xi_t) = \left(\int_0^t \xi_t dt, \xi_t \right). \tag{1.52}$$

Even though ξ_t is a hyperbolic Brownian motion, it is not evident how many features of ξ_t are passed on to (X_t, ξ_t). The construction of [Dudley1966] showed that the associated starting probabilities are $\mathcal{L}$-invariant and strongly Markov (as in Definition 70).
As a hypo-elliptic diffusion, Dudley's diffusion admits a kernel $p((x,w),(y,v),t)$ and by applying a result of Tutubalin on the characterization of infinitely divisible laws in the radial case, Dudley was able to obtain a explicit formula for the semigroup on one nappe of the hyperboloid (denoted $\mathcal{U}$):

$$P_t = \frac{1}{(4\pi m t)^{3/2}} \left(\frac{\rho}{\sinh(\rho)} \right) e^{-mt - \rho^2/(4mt)} d\mu_{\mathbb{H}}$$

where $\mu_{\mathbb{H}}$ is $\mathbb{H}^3$-invariant and given by

$$d\mu_{\mathbb{H}^3} = 4\pi(\sinh(\rho))^2 d\rho d\Omega$$

and $d\Omega$ is the normalized surface area on the sphere and ρ is the Riemannian distance in $\mathbb{H}^3$ (again see [Dudley1966, Section 10] for details).
Given $(x,v),(y,w) \in \mathbb{M}^n \times \mathbb{H}^n$ with $x \ll y$ it is natural to condition Dudley's process to hit (y,w) after "time" 1 being started at (x,v). We use quotations to express that the external parameter for Dudley's process is not the coordinate time but the arc-length parameter of X_t. This conditioning is well-defined via the strong Markov property proved in [Dudley1966, Section 6], so the law of $(X_t,\xi_t)|((X_0,\xi_0) = (x,v_1),(X_1,v_1) = (y,v_2))$ is a probability measure on $\Omega_{(x,v),(y,w)}$. Because Dudley's process emulates the behaviour of Brownian motion, it's conditioning is a natural measure to use on Bridge space. As explained at the beginning in section 1.3.1, the law of Brownian bridges is the canonical reference measure for the Dynamical Schrödinger problem on Riemannian manifolds, this motivates the idea of using the conditioned law of Dudley's process as a reference measure for a Schrödinger problem. We should make sure to define the correct ambient space for this problem before we relate it to RDSch. We will perform this conditioning carefully after we set up the correct framework (see (1.63)). This process will lead to Definition 81. In the next chapter (Section 1.4.3) we explain how to use Large deviation properties of such processes to obtain optimal transport plans as weak limits of entropically regularized optimal transport plans.

1.3.4 Phase-space and the Schrödinger Problem

In Minkowski space $\mathbb{M}^n$, we define relativistic phase-space as a subset of the tangent bundle, namely $\mathcal{PS} = \mathbb{M}^n \times \mathcal{U}$ where $\mathcal{U}$ is the upper nappe (sheet) of the hyperboloid $\mathbb{H}^n$.
We aim to define the Schrödinger problem in phase space of $\mathbb{M}^n$. In the case of $\mathbb{M}^3$ phase-space is a 7-dimensional sub-manifold of TM (8-dimensional). In [Chiarini-Conforti-Greco] the Schrödinger problem on euclidean phase-space with partial information was introduced, we recall the definition and one of their results before moving towards the case of special relativity.

Schrödinger problem on Euclidean phase-space with partial marginals

Given r a Borel probability measure on $C([0,1],\mathbb{R}\times\mathbb{R})$ assume that r is concentrated on paths of the form $t \to (x(t),v(t))$ where $t \to x(t)$ is almost everywhere differentiable and

$$\begin{cases} dx(t) = v(t)dt & (1.53) \\ dv(t) = dW_t & (1.54) \end{cases}$$

where dW_t refers to Ito-integration with respect to 1-dimensional Brownian motion W_t. Let us analyze the analogue of RDSch with respect to this measure r, we aim to find

$$\inf_{\mathbb{P}\in\mathcal{P}(C([0,1],\mathbb{R}\times\mathbb{R}))} \mathrm{Ent}(\mathbb{P}|r) \tag{1.55}$$

over all measures satisfying $\mathrm{Proj}_x \#(e_0\#\mathbb{P}) = \mu_0, \mathrm{Proj}_x(e_1\#\mathbb{P}) = \mu_1$.
By definition of entropy it is enough to consider measures concentrated on drifted Brownian motions

$$\begin{cases} dx = vdt & (1.56) \\ dv = adt + dw. & (1.57) \end{cases}$$

In the same document [Chiarini-Conforti-Greco], the authors show using Girsanov's theorem and an argument a la Benamou-Brenier that the problem is equivalent to the minimization

$$\inf_{a\in L^2(\mu_t\times dt)} \int_0^1 \int ||a||^2 d\mu_t(x,v)dt \tag{1.58}$$

where μ_t satisfies the Fokker-Planck equation

$$\frac{\partial \mu}{\partial t} + v\cdot\nabla_x\mu_t + \nabla_v(av) - \frac{1}{2}\Delta_v\mu_t = 0. \tag{1.59}$$

Equation (1.59) describes the evolution of the associated procees in phase-space $\mathbb{R}\times\mathbb{R}$.
Two main difficulties arise when trying to adapt the same techniques to the Schrödinger problem on phase space (or kinetic Schrödinger Problem) on globally hyperbolic chrono-regular Lorentzian length-spaces.

1. The absence of a well-adapted Girsanov theorem.
2. The unphysicality of the external parameter.

We address these problems in the following sections. In the following section we define and analogue of this idea in relativistic phase space for the case of special relativity and in section 1.3.4 we explain the approach one could follow for general curved geometries.

The kinetic relativistic Minkowskian Schrödinger Problem

Let $\mathbb{M}^3$ be Minkowski space and denote by $\mathcal{U}$ one nappe (sheet) of the hyperboloid. We refer to $\mathbb{M}^n\times\mathcal{U}$ as the Phase space for special relativity and denote it $\mathcal{PS}$. Notice that this simple splitting of the (co-)tangent bundle is a direct consequence of homogeneity of Minkowski space and is not usually available in curved geometries. In general curved geometries, phase space is not so easily described and we refer to [Franchi-LeJan2012] for a complete exposition.

Definition 81. *(Dynamical Schrödinger problem on phase-space of Minkowski space).*
Given $\tilde{\mu}_0,\tilde{\mu}_1\in\mathcal{P}(\mathcal{PS})$ with and $R\in\mathcal{P}(C([0,1],\mathcal{PS})$ we define the kinetic relativistic Schrödinger problem as

$$\inf_{\Pi\in\Gamma_{PS}(\tilde{\mu}_0,\tilde{\mu}_1)} \mathrm{Ent}(\Pi|R) \tag{RKSchP}$$

where $\Gamma_{PS}(\tilde{\mu}_0,\tilde{\mu}_1)$ is the set of Borel probability measures on $\mathcal{PS}$ whose marginals at endpoints are $\tilde{\mu}_0,\tilde{\mu}_1$ i.e. $e_0\#\Pi=\tilde{\mu}_0, e_1\#\Pi=\tilde{\mu}_1$ and

$$\mathrm{spt}(\Pi)\subseteq\{(\gamma,v) : v(t)=\dot{\gamma}(t) \text{ a.e. on } t\in[0,1]\} \tag{1.60}$$

Condition (1.60) ensure the probability measure is concentrated on differentiable curves (on space-time) whose velocity is in the hyperboloid.

Definition 82. *(Partial dynamical Schrödinger problem on phase-space of Minkowski space).*
Given $\mu_0,\mu_1\in\mathcal{P}(\mathbb{M}^n)$ and $R\in\mathcal{P}(C([0,1],\mathcal{PS})$ we define the external dynamical Schrödinger problem as

$$\inf_{\Pi\in\Gamma_x(\mu_0,\mu_1)} \mathrm{Ent}(\Pi|R) \tag{1.61}$$

where $\Gamma_x(\mu_0,\mu_1)$ is the set of Borel probability measures on $\mathcal{PS}$ whose $\mathbb{M}^n$-projected marginals at endpoints are μ_0,μ_1 i.e. $e_1\#\,\mathrm{Proj}_x\#\Pi=\mu_0, e_1\#\,\mathrm{Proj}_x\#\Pi=\mu_1$ and satisfy (1.60).

Notice the difference between the problems is the specification of initial data. In the first problem RKSchP initial data is prescribed as probability measures on phase-space meaning that momentum is specified at beginning and end. In contrast, (1.61) specifies only the space-time data. It is clear then that if $\mu_0 = \mathrm{Proj}_x \tilde{\mu_0}$ and $\mu_1 = \mathrm{Proj}_x \tilde{\mu_1}$ then any solution with (μ_0, μ_1, r) data for (1.61) would yield a solution for RKSchP. Note that by Lemma 39 the same is true when comparing with RKSchP. In the Kinetic version of the problem the causality condition is implicitly required through equation (1.60), in contrast to the dynamic Relativistic Schrödinger Problem RDSch where the requirement is explicit. We briefly review the analogue of Proposition 44. Existence: Note that for $\mathcal{P}$ we are using the Borel σ-algebra and given the product structure of $\mathcal{PS}$ we are implicitly using

$$d((\gamma_1, v_1), (\gamma_2, v_2)) = \sup_{t\in[0,1]} ||\gamma_1(t) - \gamma_2(t)||_2 + d_{\mathbb{H}^n}(v_1(t), v_2(t)) \tag{1.62}$$

where $||\cdot||_2$ is the Euclidean norm on $\mathbb{R}^{n+1}$ and $d_{\mathbb{H}^n}$ is the hyperbolic distance. This implies that the constraint (1.60) is again closed and existence is concluded similarly to that in Proposition 44. If a solution exists, condition (1.60) is also convex so strict convexity of entropy yields uniqueness. Fix $x \ll y, x, y \in \mathbb{M}^3$ and $v \in \mathbb{H}^3$, by [Dudley1966, Theorem 6.2] there exists a $\mathcal{L}$-invariant strongly Markov process starting at (x, v). By the strong Markov property, we condition Dudley's process (X_t, V_t) to hit (y, w) at proper-time 1. Set $\mathcal{PS}_{(x,v),(y,w)}$ to be the set

$$\{\gamma : [0,1] \to M, \gamma \in \mathcal{C}^1([0,1], M), (\gamma(0), \gamma'(0^+)) = (x, v), (\gamma(1), \gamma'(1^-)) = (y, w)\}$$

where $\mathcal{C}^1([0,1], M)$ refers to differentiable causal curves from $[0,1]$ to M. Let $\mathcal{W}^{(x,v),(y,w)}$ be the law on $\mathcal{PS}_{(x,v),(y,w)}$ of the process

$$(X_t, Y_t) \mid (X_0, v_0) = (x, v), (X_1, V_1) = (y, w). \tag{1.63}$$

Finally consider $\nu_0, \nu_1 \in \mathcal{P}_c(\mathcal{PS})$ such that $\mathrm{Proj}_x \nu_0 \preceq \mathrm{Proj}_x \nu_1$ and define

$$\mathcal{W}^D(\cdot) = \int \mathcal{W}^{(x,v),(y,w)}(\cdot) d\nu_0(x, v) d\nu_1(y, w). \tag{1.64}$$

For simplicity we have chosen ν_0, ν_1 to be probability measures with compact support in order to have a reference probability measure instead of the general case of reference measures studied by [Leonard2014] as mentioned in section 1.1. Existence of solutions to (RKSchP) and (1.61) then follows from topological properties of the underlying space.

Notice that X_t is parametrized by arc-length and it describes the evolution of the particle parametrized by proper time. By definition, $t \to (X_t, V_t)$ is the martingale solution for the operator L^σ. This conditioning allows us to exploit the theory of the Markov Semigroups (as in [Bakry-Gentil-Ledoux]) and correspond to $((x, v), (y, w), 1)$-Bridges in the sense of section 1.3.8. This means that our conditioning correctly corresponds to our convention of affine parametrization of curves $\sigma : [0,1] \to M$.

Assumption 1. *There exists an invariant measure m_σ, for Dudley's process, in the semigroup sense: if P_t denotes the associated Markov semigroup, then for every $f \in Dom(L^\sigma)$ we have*

$$\int P_t f(x, v) dm_\sigma(x, v) = \int f(x, v) dm_\sigma(x, v).$$

where L^σ is from (1.51).

Although the invariant measure on $\mathbb{H}^n$ may not be reversible (see [Baudoin],[Baudoin-Gordina-Mariano]), in the Euclidean analogue of Kolmogorov's operator the associated measure is

$$dm_\sigma(x,v) = \frac{1}{\sqrt{2\pi\sigma}} e^{-||v||^2/2\sigma} dvdx.$$

as seen in [Baudoin].

Proposition 83. *(Born's formula for the Kinetic problem)*
Under Assumption 1, let $\mathcal{W}^D$ be the measure associated to Dudley's process as in (1.64) and set

$$R^\sigma(\cdot) := \int \mathcal{W}^{(x,v)}(\cdot) dm_\sigma(x,v). \tag{1.65}$$

where the condititioning is done via (1.63). Assume that μ_0 and μ_1 are absolutely continuous with respect to m_σ and there exist $f_0, g_1 : \mathcal{PS} \to [0,\infty)$

$$\begin{aligned} f_0(x,v)\mathbf{E}_{\mathcal{W}^D}[g_1(X_1,V_1) \mid (X_0,v_0) = (x,v)] &= \frac{d\mu_0}{dm_\sigma}(x,v) \\ g_1(y,w)\mathbf{E}_{\mathcal{W}^D}[f_0(X_0,V_0) \mid (X_1,V_1) = (y,w)] &= \frac{d\mu_1}{dm_\sigma}(y,w). \end{aligned} \tag{1.66}$$

Then any solution Π of (RKSchP) is an (f_0,g_1)-transform of m_σ, i.e. for every $t \in [0,1]$ if $\Pi_t := e_t \# P$ then

$$d\Pi_t = f_t g_t dm_\sigma,$$

where $f_t, g_t : \mathcal{PS} \to [0,\infty)$ are given by

$$f_t(z) = \mathbf{E}_{\mathcal{W}^D}[f_0(e_0)|e_t = z],\ g_t(z) = \mathbf{E}_R[f_1(e_1)|e_t = z],$$

where e_t is again the evaluation map.

Remark 84. *Note that R^σ is indeed an infinite measure and so care needs to be taken when conditioning and entropy should be understood as the second definition in section 1.1 (Definition 28).*

Proof. The result is a direct application of [Leonard2014, Theorem 3.4] as it's hypothesis is exactly Assumption 1 given that m_σ is a reversible invariant measure for $\mathcal{W}^D$. ∎

In the following propositions we use the notation of Full Markov triples developed for the Bakry-Emery study of curvature dimension conditions. See [Bakry-Gentil-Ledoux] for a detailed exposition or [J] for a more accessible version, for the specific of the associated semigroup for Dudley's process see [Baudoin-Gordina-Mariano]. Recall that given an operator L, one can define it's associated Carré Du Champ as a way to evaluate deviation from linear differentiation, that is for $f, g \in D(L)$

$$\Gamma(f,g) := \frac{1}{2}\left(L(fg) - fLg - gLf\right) \tag{1.67}$$

and we often write $\Gamma(f) := \Gamma(f,f)$. The operator $\Gamma(\cdot)$ is called the *Carré du Champ* associated to L.

Lemma 85. *(De Bruijn's identity)*
For a full-Markov triple (X, Γ, μ) (see [J, Section 1]) we have for $d\nu_t = P_t f d\mu$

$$\frac{d}{dt}\operatorname{Ent}(\nu_t \mid \mu) = -\int \frac{\Gamma(P_t f)}{P_t f} d\mu. \tag{1.68}$$

where $P_t f$ is the Markov semigroup associated to (X, Γ, μ) evaluated at $f \in D(L)$.

For a proof see [Bakry-Gentil-Ledoux, Proposition 5.2.2] or the more accessible reference [J]. The right-handside of (1.68) is nothing but the negative of Fisher's information.

Corollary 86. *(Relativistic phase-space analogue of (1.58) via Carré du Champ)*
Let $\mathcal{W}_D^\sigma \in \mathcal{P}(\mathcal{PS})$ be Dudley's measure as in 1.3.3, define R^σ via (1.65) then for every $f \in D(L^\sigma)$,

$$\operatorname{Ent}(\nu_1^f | m_\sigma) = \operatorname{Ent}(\nu_0^f | m_\sigma) - \frac{\sigma^2}{2}\int_0^1 \int \left|\left|\nabla^{\mathbb{H}^d} \log a_t(x,v)\right|\right|^2 dm_\sigma(x,v)dt \tag{1.69}$$

where

$$a_t(x,v) = \int f(z) de_t \# \mathcal{W}^{(x,v)}(z), \tag{1.70}$$

$||\cdot||^2$ refers to $g_v(\cdot,\cdot)$ for the Riemannian metric on $\mathbb{H}^d$, and

$$\nu_t^f(A) = \int_A \int_{\mathbb{H}^d} f(z) de_t \# \mathcal{W}^{(x,v)}(z) dm_\sigma(x,v). \tag{1.71}$$

Further, $P_t f := \dfrac{d\nu_t^f}{dm_\sigma}$ satisfies the Fökker-Planck equation with respect to L^σ i.e.

$$\partial_t P_t f = L^\sigma P_t f. \tag{1.72}$$

Proof. By definition of R^σ (equation (1.65)) for any Borel set $A \subseteq \mathbb{M}^d$,

$$e_t \# R^\sigma(A) = \int \mathcal{W}^{x,v}(e_t^{-1}(A)) dm_\sigma(x,v) = \int e_t \# \mathcal{W}^{(x,v)}(A) dm_\sigma(x,v). \tag{1.73}$$

Evaluating (1.67) with L_σ as in 1.3.3 we have for every $f \in Dom(L^\sigma)$

$$\Gamma(f) = \frac{\sigma^2}{2}||\nabla_v^{\mathbb{H}^d} f||^2. \tag{1.74}$$

Substituting 1.74 in De Bruijn's identity, (1.68)

$$\operatorname{Ent}(e_1 \# R \mid m_\sigma) - \operatorname{Ent}(e_0 \# R \mid m_\sigma) = -\frac{\sigma^2}{2}\int_0^1 \int \left|\left|\nabla^{\mathbb{H}^d} \log a_t(x,v)\right|\right|^2 dm_\sigma(x,v)dt$$

which yields the result. ∎

Remark 87. *Although Corollary 1.69 is not posed exactly as the Benamou-Brenier formula 1.4, the Fökker-Planck equation (1.72) together with the explicit description of Γ are the essential tools for an entropic Benamou-Brenier type formulation as shown in [Leonard2014, Proposition 4.1] or [Leonard, Section 4] which motivated Corollary 86.*

Proposition 88. *(Recovering Dudley's process)*
Denote by $R_{x,y}$ the proyection onto $\mathbb{M}^n$ of the measure $\mathcal{W}_{(x,v),(y,w)}$ associated to Dudley's process conditioned to proper time with fixed endpoints $x \ll y$ as described in 1.63, then there exists a collection of measures on mid-sets in $J^+(x) \cap J^-(y)$ such that the construction 2 yields $R_{x,y}$.

Proof. Observe that $\mathcal{W}_{(x,v),(y,w)}$ is proper-time parametrized and so for $t = 1/2$ set $\sigma(A) = r(\{\gamma : \gamma(1/2) \in A\})$ for $A \in \mathcal{B}(\mathrm{MID}(x,y))$. The parametrization ensures $\gamma(1/2)$ is concentrated in $\mathrm{MID}(x,y)$. It is then automatic for every dyadic time. The condition in 2 is satisfied by continuity of the causal curves enforced by $\mathcal{W}^{(x,v),(y,w)}$. ∎

Remark 89. *Section 1.3.3 dealt only in the case of no potential, but if we assume a drift on velocities (as in the case of the process in [Dunkel-Hanggi]) then Theorem 83 would have an invariant measure depending on this potential*

$$d\mu_V(x,v) \propto e^{-V(x)-|v|^2} dxdv$$

but the abstract framework of Markov Triples still works by setting

$$\hat{P}_t = e^{Vt} P_f t,$$

see [Bakry-Gentil-Ledoux, Section 1.15.6] which yields a straight-forward generalization of section 1.3.4 to the framework of [Dunkel-Hanggi] and others.

Extensions to curved geometries

In [Franchi-LeJan2007] a process in general smooth Lorentzian manifolds was presented and shown to be the only diffusion whose law is invariant under the group of isometries of the manifold. This process is a generalization of Dudley's process to curved geometries. In this section we briefly give intuition for this process, following [Franchi], if M is a smooth Lorentzian manifold of signature $(+,-,\ldots,-)$ let T^1M denote the positive half of it's pseudo-unit tangent bundle. Let $G(M)$ be the bundle of direct pseudo-orthonormal frames with first element in T^1M and with fibers modeled on the Lorentz-Mobius group G. Let π_1 denote the canonical projection from $G(M)$ onto the unit tangent bundle T^1M. The infinitesimal operator in terms of the Casimir operator $(\mathcal{L}_0)$ can be defined via

$$L_\sigma := \mathcal{L}_0 + \frac{\sigma^2}{2}\Delta_v$$

The process generated by L_σ is the generalization of Dudley's process to completely general Lorentzian manifolds. We note also that [Dunkel-Hanggi], [Chevalier-Debbasch] have other generalization of Dudley's process when one consider velocity fields. We do not study the Schrödinger problem with respect to such laws on general manifolds and leave the question open for future research. We detail this line of investigation in section 1.5.2 and it's connection with Bismut's hypo-elliptic Laplacian in section 1.5.2.

1.3.5 The Levy-like Bridge construction in general frameworks

In this section we analyze constructions 1 and 2 in the general framework of globally hyperbolic chrono-regular Lorentzian length spaces (Definition 20). We also study other techniques to create measures for $(x,y,1)$-bridges on $(C[0,1],M)$ or $(Cad[0,1],M)$ in which the set must be endowed

with the induced Skohorod topology ([Billingsley, Theorem 12.1]). A first naive approach, could be to take measures on $C([0,1],M)$ and ignore the physicality. Because the Borel σ-algebra generated by d_∞ does not depend on the separation function ℓ, the resulting paths may not be causal so one could restrict the measure to only those paths which are causal. We will not follow this approach although we observe that it could lead to interesting constructions.

Lemma 90. *Let (M,d,ℓ) be a ghcrlls (Definition 20), let $x \ll y$ where $x,y \in M$, then*

$$\mathrm{MID}(x,y) = \{x \leq z \leq y : \ell(x,z) = \ell(z,y)\}$$

is compact.

Proof. The proof is the same as Proposition 54 as ℓ^+ is assumed to be continuous. ∎

Assumption 2. *Given $x \ll y$ assume that $\{\sigma(a,b)\}_{\{a,b\in J^+(x)\cap J^-(y)\}}$ is a collection of measures such that*

1. *$\sigma(a,b) \in \mathcal{P}(\mathrm{MID}(a,b))$*
2. *For every i,*
$$\lim_{s\to t, s\in\mathcal{D}} d({}_nP_i, {}_nP_{i+1}) \to 0, \text{ uniformly}$$
where ${}_nP_i$ are chosen as in Construction 2 and $n\to\infty$ as $s\to t$ in $\mathcal{D}$.

Theorem 91. *In a ghcrlls as in Definition 20, assume the space satisfies the uniform control on distance (Definition 60), let $\{\sigma(a,b)\}_{\{a,b\in J^+(x)\cap J^-(y)\}}$ satisfy Assumption 2, then there exists a unique measure in $\mathcal{P}(\Omega_{x,y})$ having $\{\sigma(a,b)\}_{\{a,b\in J^+(x)\cap J^-(y)\}}$ as conditional prescribed measures at dyadic times.*

Proof. Due to Lemma 90 and definition 60, by global hyperbolicity diamonds are compact and nested so the proof of Proposition 62 as 1.35 is now replaced by the second part of Assumption 2. ∎

Remark 92. *Note that the importance of constructions 1 and 2 is that they can be taken straightforwardly to ghcrlls and still satisfy important properties directly connected with the Schrödinger problem that we will study in the next chapter (section 1.4).*

Although the general assumption 2 is sufficient to replicate the proof of the $c-$contracted mid-sets of Theorem 62, it is in general very hard to check. Indeed, most physically relevant processes depend on ℓ-paths or ℓ-curves and not only on d. We will see in section 1.4 a different assumption which will allow us to generalize the method to prescribe probability at dyadic intervals but not necessarily uniformly bounded away from the boundary. We know that another way to generate measures in $\{\gamma \in C([0,1],M) : \gamma(0)=x, \gamma(1)=y\}$ is to use the tightness criteria from [Billingsley, Theorem 7.3]: Given a sequence of measures μ_n in this space if for each $\epsilon>0$ we have

$$\lim_{\delta\to 0}\limsup_{n\to\infty} \mu_n\left(\left\{\gamma : \sup_{|t-s|<\delta} d(\gamma(t),\gamma(s)) \geq \epsilon\right\}\right) = 0. \tag{1.75}$$

then there exists a weak-limiting measure for the sequence in the space endowed with the Skohorod topology.
Under the uniform control on distances (Definition 60), one can instead require that for every $\epsilon>0$

$$\lim_{\delta\to 0}\limsup_{n\to\infty} \mu_n\left(\left\{\gamma : \sup_{t\in[0,1-\delta]} d(\gamma(t),\gamma(t+\delta)) \geq \epsilon\right\}\right) = 0.$$

Furthermore, if we consider instead $(\mathrm{Cad}[0,1], M)$, by [Billingsley, Theorem 13.6] together with Kolmogorov consistency and right-continuity to prescribe measures with finite dimensional distributions $\mu_{t_1,\dots,t_k}$ another sufficient condition is that there exist $\alpha > 1/2, \beta \geq 0$ and F continuous non-decreasing such that for $t_1 \geq t \geq t_2$

$$\mu_{t_1,t,t_2}(\{(u_1,u,u_2) : d(u_1,u) \wedge d(u,u_2) \geq \lambda)\}) \leq \frac{1}{\lambda^{4\beta}}(F(t_1) - F(t_2))^{2\alpha}. \tag{1.76}$$

We will focus on the continuous case and try to prove existence from further assumptions in the underlying space. We observe that under milder assumptions on the underlying space the bridges may fail to be d-continuous as the physical nature of (M, ℓ, d) is too general. In this case, equation (1.76) may be the correct tool to generate measures useful in large deviation principles.
Following the ideas of Construction 2, we can ask ourselves whether or not considering the uniform (Hausdorff) measure in (M, ℓ, d) as a probability measure on mid-sets replicates the convergence of linear interpolations in Brownian bridges. It is then natural to ask whether or not joining mid-sets at dyadic times via ℓ-paths yield a sequence of tight measures (in the sense of (1.75)).
Fix $x \ll y$ and consider $z \in \mathrm{MID}(x,y)$, denote by $\mathbb{P}_1$ the measure in $\mathcal{P}(\Omega_{x,y})$ given by

$$\mathbb{P}_1(A) = \int_{\mathrm{MID}(x,y)} \delta_{\gamma^\ell_{x,z,y}}(A) d\mathcal{U}(z)$$

where $\gamma^\ell_{x,z,y}$ denotes the concatenation of the ℓ-path from x to z and that one from z to y parametrized via proper time so that

$$\gamma^\ell_{x,z,y}(0) = x, \gamma^\ell_{x,z,y}(1/2) = z, \gamma^\ell_{x,z,y}(1) = y.$$

and $\mathcal{U}$ denotes the uniform measure on $\mathrm{MID}(a,b)$ (the normalized Hausdorff measure of non-trivial dimension).
Proceed inductively by defining $\mathbb{P}_2$ to assign probability according to independent uniform measures on mid-sets between x and z and z and y. We replicate the idea of the construction on section 1.3.1 with the difference that we know ask ourselves about the *uniform tightness* of this sequence of measures *as probabilities on the path-space*. In the following theorem we show that this construction is not uniformly tight in the sense of (1.75) even in the simplest Minkowskian case.

Theorem 93. *Consider $\mathbb{M}^2$ and set $\sigma_{a,b} = \mathcal{H}_{a,b}$ then the sqeuence of measures prescribed by considering ℓ-paths joining the points of construction in section 1.3.1 is not uniformly tight in the sense of (1.75) i.e. there exists $\epsilon > 0$ such that*

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}_n \left(\left\{ \sup_{t\in[0,1]} d(\gamma(t+\delta), \gamma(t)) > \epsilon \right\} \right) \neq 0. \tag{1.77}$$

Indeed we will show the stronger result that for every $\epsilon > 0$

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}_n \left(\left\{ \sup_{t\in[0,1]} d(\gamma(t+\delta), \gamma(t)) > \epsilon \right\} \right) = 1 \tag{1.78}$$

Lemma 94. *Let $a = (x_1,t_1), b = (x_2,t_2) \in \mathbb{M}^2$ and assume that $(x_1,t_1) \ll (x_2,t_2)$, let $z \in \mathrm{MID}(a,b)$ and*

$$R := \max\left\{ \frac{d(a,z)}{d(a,b)}, \frac{d(z,b)}{d(a,b)} \right\}. \tag{1.79}$$

Denote

$$C = \frac{(x_2 - x_1)(t_2 - t_1)}{d(a,b)^2}. \tag{1.80}$$

If z is distributed uniformly in $\mathrm{MID}(a,b)$ *and* $1/2 \leq r \leq \sqrt{1/2 + |C|}$

$$\mathrm{P}(R \geq r) = 1 + 2|C| - \sqrt{4C^2 + 4r^2 - 1}. \tag{1.81}$$

Proof. The proof is a direct computation, without loss of generality we only show the case where $C > 0$. In $\mathbb{M}^2$ we can explicitly describe $\mathrm{MID}(a,b)$:

$$\left\{ \left(\frac{x_1 + x_2}{2} + s(t_2 - t_1) + \frac{t_1 - t_2}{2}, \frac{t_1 + t_2}{2} + s(x_2 - x_1) + \frac{x_1 - x_2}{2} \right) : s \in [0,1] \right\}. \tag{1.82}$$

Write $z = M_s$ if $s \in [0,1]$ is the parametrized description above, then the distances in (1.79) are given by

$$d(a, M_s) = \sqrt{(1/2 - s + s^2)\, d(a,b)^2 + (1 - 2s)(x_2 - x_1)(t_1 - t_2)}, \tag{1.83}$$

$$d(M_s, b) = \sqrt{(1/2 - s + s^2)\, d(a,b)^2 + (1 - 2s)(x_2 - x_1)(t_2 - t_1)}. \tag{1.84}$$

The only difference between (1.83) and (1.84) is the sign change on the last term. Note also that $s \in [0,1]$ so we can explicitly compute the maximum: in our case where $C > 0$, $d(M_s, b) \geq d(a, M_s)$ if $s \in [0, 1/2]$ and $d(M_s, b) \leq d(a, M_s)$ if $s \in [1/2, 1]$.
Hence,

$$\mathrm{P}(R \geq r) = \mathrm{P}\left(\frac{d(M_s, b)}{d(a,b)} \geq r, s \in [0, 1/2] \right) + \mathrm{P}\left(\frac{d(a, M_s)}{d(a,b)} \geq r, s \in [1/2, 1] \right).$$

Computing this probabilities is finding the set of solutions of the following inequalities:

$$\sqrt{1/2 - s + s^2 + (1 - 2s)C} \geq r, s \in [0, 1/2]$$
$$\sqrt{1/2 - s + s^2 - (1 - 2s)C} \geq r, s \in [1/2, 1].$$

The set of solutions for this inequalities is given by $[0, 1/2 + C - \sqrt{C^2 + r^2 - 1/4}]$ whose length is $1/2 + C - \sqrt{C^2 + r^2 - 1/4}$. Both inequalities yield the same length and so finally, we obtain (1.81):

$$\mathrm{P}\left(R \geq r\right) = 2(1/2 + C - \sqrt{C^2 + r^2 - 1/4}) = 1 + 2C - \sqrt{4C^2 + 4R^2 - 1}. \tag{1.85}$$

The case $C < 0$ is completely analogous, except the solutions to the respective inequalities give $1 - C - \sqrt{4r^2 + 4C^2 - 1}$. Putting both cases together we obtain that if $1/2 \leq R \leq \sqrt{1/2 + |C|}$ then

$$\mathrm{P}\left(R \geq r\right) = 1 + 2|C| - \sqrt{4R^2 + 4C^2 - 1}.$$

∎

Lemma 95. *(Scales increase to infinity almost surely in $\mathbb{M}^2$)*
Consider $a = (x_1, t_1), b = (x_2, t_2)$ with $a \ll b \in \mathbb{M}^2$. Choose uniformly $z_1 \in \mathrm{MID}(a,b)$. Set

$$R_1 := \max\left\{ \frac{d(a,z)}{d(a,b)}, \frac{d(z,b)}{d(a,b)} \right\}. \tag{1.86}$$

Proceed inductively, let $z_n \in \mathrm{MID}(a, z_{n-1})$ *and write*

$$R_n := \max\left\{\frac{d(a,z_n)}{d(a,z_{n-1})}, \frac{d(z_n,z_{n-1})}{d(a,z_{n-1})}\right\}, \tag{1.87}$$

Then

$$2^n \prod_{k=1}^{n} R_k \nearrow \infty \text{ almost surely.} \tag{1.88}$$

Proof. Observe that

$$2^n \prod_{k=1}^{n} R_k = \prod_{k=1}^{n} (2R_k). \tag{1.89}$$

As $2R_k \leq 1$ convergence of the second term in (1.89) is equivalent to the convergence of the following series

$$\sum_{k=1}^{\infty} (R_k - 1/2)\,.$$

By Kolmogorov's 3-series Theorem for dependent variables (see [Brown, Theorem 1]) we can show the series does not converge almost surely to a finite random variable if we show that there exists $A > 0$ such that

$$\sum_{k=1}^{\infty} \mathrm{P}(R_k - 1/2 \geq A) \not< \infty. \tag{1.90}$$

Given any $A \in (0, 1/\sqrt{2} - 1/2)$,

$$\sum_{k=1}^{\infty} \mathrm{P}(R_k - 1/2 \geq A) = \sum_{k=1}^{\infty} 1 + 2|C_k| - \sqrt{4(A+1/2)^2 + 4C_k^2} \tag{1.91}$$

$$\geq \sum_{k=1}^{\infty} (1 - \sqrt{4(A+1/2)^2}) = \infty. \tag{1.92}$$

where the last equality is due to $A \in (0, 1/\sqrt{2} - 1/2)$ which is an admissible value for every $R_k - 1/2$ as $R_k \in [1/2, \sqrt{1/2 + |C_k|}]$.
Therefore the series does not converge almost surely and note that

$$2^{n+1} \prod_{k=1}^{n+1} R_k = 2R_{n+1} \cdot 2^n \prod_{k=1}^{n} R_k,$$

so the sequence $2^n \prod_{k=1}^{n} R_k$ is non-decreasing (as $R_k \geq 1/2$) and does not converge uniformly which implies that it increases to infinity almost surely, as desired. ∎

Proof. of Theorem 93: By Lemma 95 we have

$$2^n \prod_{k=1}^{n} R_k \nearrow \infty \text{ almost surely.}$$

This implies also that for every $\epsilon, \delta > 0$ and $d(x,y) > 0$ fixed,

$$\lim_{n\to\infty} \mathrm{P}\left(2^n d(x,y) \prod_{k=1}^{n} R_k > \frac{\epsilon}{\delta}\right) = 1.$$

Finally note that

$$\begin{aligned}
&\limsup_{n\to\infty} \mathrm{P}\left(\sup_{t\in[0,1]} d(\sigma_n(t), \sigma_n(t+\delta)) > \frac{\epsilon}{\delta}\right) \\
=&\limsup_{n\to\infty} \mathrm{P}\left(2^n d(x,y) \max_{k\in\{0,\cdots,2^n-1\}} \{d(\sigma_n(k/2^n), \sigma_n((k+1)/2^n))\} > \epsilon\right) \\
\geq& \lim_{n\to\infty} \mathrm{P}\left(2^n d(x,y) \prod_{k=1}^{n} R_k \geq \frac{\epsilon}{\delta}\right) = 1.
\end{aligned}$$

From which we obtain

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}_n\left(\sup_{t\in[0,1]} d(\sigma(t+\delta), \sigma(t)) \geq \epsilon\right) = 1 \neq 0 \tag{1.93}$$

∎

One could also wonder whether the observation above is a result of the choice of compact sets. It is natural to think that the correct compact sets should be the ones associated to ℓ-length as $\{L_\ell(\cdot) < \tau\}$ is closed by upper continuity of ℓ. Nevertheless, the sequence is also not uniformly tight with respect to these sets as demonstrated by the next Theorem.

Theorem 96. *In $\mathbb{M}^{1,1}$ uniform measure on mid-sets family of ℓ-path joined dyadics is not tight with respect to the family of sets*

$$K_\tau = \{\gamma \in J(x,y) : L_\ell(\gamma) < \tau\}.$$

In other words, for every family of distributions in $\mathrm{MID}(a,b)$ *we have*

$$\lim_{\tau\to 0^+} \limsup_{n\to\infty} \mathbb{P}(L_\ell^n(\gamma) < \tau) \neq 0 \tag{1.94}$$

Proof. Indeed we will show that for any collection of distributions in mid sets, we have

$$\lim_{\tau\to 0} \limsup_{n\to\infty} \mathbb{P}(L_\ell^n(\gamma) < \tau) = 1. \tag{1.95}$$

Observe that if $z_1 \sim \mathcal{U}(x,z)$ and $z_2 \sim \mathcal{U}(z,y)$, let σ_2 denote the concatenation of the ℓ-paths $x \to z_1 \to z \to z_2 \to y$, then

$$\begin{aligned}
L_\ell(\sigma_2) &= \ell(x,z_1) + \ell(z_1,z) + \ell(z,z_2) + \ell(z_2,y) \\
&= 2\ell(z,z_1) + 2\ell(z_2,z).
\end{aligned}$$

Therefore by the Tower property,

$$\mathbf{E}[L_\ell(\sigma_2)] = 2\mathbf{E}[\mathbf{E}[\ell(Z_1,Z) + \ell(Z,Z_2)|Z]].$$

Observe that if $a := (x_1, t_1) \ll (x_2, t_2) =: b$ and $z \in \text{MID}(a, b), z = M_s$ as before, then it is easily checked that

$$\ell(a, M_s) = \left(\sqrt{s - s^2}\right) \ell(a, b)$$

Therefore,

$$\mathbf{E}[L_\ell(\sigma_2)] = 4 \cdot \mathbf{E}[L_\ell(\sigma_1)] \int_0^1 \sqrt{s - s^2} ds = \mathbf{E}[L_\ell(\sigma_1)] \frac{4\pi}{8} = 4 \left(\frac{\pi}{8}\right)^2.$$

Inductively,

$$\mathbf{E}[L_\ell(\sigma_n)] = 2 \cdot \frac{\pi}{8} \mathbf{E}[L_\ell(\sigma_{n-1})] = 2^n \left(\frac{\pi}{8}\right)^n = \left(\frac{\pi}{4}\right)^n, \tag{1.96}$$

using Markov's inequality for every fixed τ we have

$$\mathbb{P}\left(L_\ell(\sigma_n) > \tau\right) \leq \frac{\mathbf{E}[L_\ell(\sigma_n)]}{\tau} \to 0 \text{ as } n \to \infty. \tag{1.97}$$

Therefore, as desired we obtain

$$\limsup_{n \to \infty} \mathbb{P}(L_\ell(\sigma_n) < \tau) = 1. \tag{1.98}$$

■

These two observations show that the uniform measure on mid-sets is not automatically scaled to obtain a limiting curve (as one may think). They also show an interesting phenomena which reaffirms the ideas in section 1.3.1: In order to obtain a uniform limiting measure, the probability measure on $\text{MID}(a, b)$ needs to concentrate around the d-geodesic mid-point between a and b. It is clear how to obtain sufficient (or necessary) conditions to obtain uniform tightness. For example, by our construction above it is sufficient for the first case to have

$$\sum_{k=1}^{\infty} (R_k - 1/2) \text{ converging almost surely}$$

which is the same as

$$2^n \prod_{k=1}^{\infty} R_k \text{ converging almost surely}$$

and it is necessary for the second criteria that

$$\mathbf{E}[L_\ell(\sigma_n)] \not\to 0.$$

This conditions are met if we re-scale the measures further to concentrate around the d-geodesic mid-point.

1.3.6 Relativistic bridge spaces, topologies of timelike curves and Markovianity

In our definition of the relativistic dynamical Schrodinger problem RDSch and it's kinetic version in $\mathbb{M}^n$ there is an implicit assumption on the topology considered in the space of causal curves. Although the topology on the space itself is fixed, many choices can be made for the topology on $\mathcal{C}([0, 1], M)$. In this section we explore different topologies on $\mathcal{C}([0, 1], M)$ and the consequences that this choice has in RDSch and the Markov property.
We start by studying the classical physical topology on the x, y-Bridge space $\Omega_{x,y} = \{\gamma \in \mathcal{C}([0, 1], M) : \gamma(0) = x, \gamma(1) = y\}$.

Definition 97. *(Standard equivalence)*
Two curves are consider equivalent if there exists a continuous monotonic function such that $\gamma_1(f(u)) = \gamma_2(u)$ *i.e. they are reparametrizations of each other.*

Under this equivalence we can focus on considering only causal curves from $[0, 1]$ to M.

Definition 98. *(C^0 topology on $\Omega_{x,y}$)*
A neighborhood of γ *in* $\Omega_{x,y}$ *consists an all continuous functions in* $\Omega_{x,y}$ *whose points in* M *lie in a neighborhood* W *of the points on* γ *in* M*.*

This definition can be found in [Hawking-Ellis, Chapter 6]. Equivalently, one can describe the C^0 topology on the equivalence classes of causal curves and say that $[\gamma_n] \xrightarrow{C^0} [\gamma]$ if the sequence of future and past endpoints converge and for every open set U s.t. γ, one has $[\gamma_n] \subseteq U$ for sufficiently large n.
According to [Hawking-Ellis, Section 6.2], the topology on curves is the one whose convergence is determined by limit curves. That is, every point on a limit curve is a subsequential limit point of the sequence. The concept of this topology yields the following fundamental lemma:

Lemma 99. *(Hawking-Ellis)*
In the context of smooth Lorentzian manifolds under the C^0 *topology, every sequence of future inextendible curves with a limit point* p *has a future inextendible limit curve passing through* p*.*

For a proof see [Hawking-Ellis, Lemma 6.2.1]. This lemma is key to the development of the theory as it is the main step to show characterizations of the achronal boundary. In [Hawking-Ellis] it is shown that Ricci curvature bounds together with the generic chronology conditions and geodesic completeness yield strong causality, characterizations of closed achronal sets and the Theorem of Seifert-Geroch on global hyperbolicity. In this section we interpret Lemma 99 as a compactness criterion on $\Omega_{x,y}$ and define 2 topologies on the same space which are weaker but also inherit a compactness criteria.

Proposition 100. *(Seifert-Geroch)*
In the smooth spacetime case (M, d, ℓ) *is globally hyperbolic iff for all* $x \ll y$ *the set* $\Omega_{x,y}$ *is compact under the standard topology described above.*

The proof is based on Lemma 99 and can be found as [Hawking-Ellis, Proposition 6.6.2].

Remark 101. *The standard topology is a* C^0 *topology and length is not continuous but upper-semicontinuous, if instead we had chosen a* C^1 *topology length would be continuous but* $\Omega^1_{x,y}$ *would not be compact.*

A temporal function and the associated polish Bridge space (Miller)

Consider (M, d, ℓ) a smooth spacetime (from section 1.1). Following [Eckstein-Miller], we make the following definition.

Definition 102. *(Time functions)*
We say that a function $\mathcal{T} : M \to \mathbb{R}$
is a generalized time function if it is increasing along any future-directed causal curve,
is a time function if it is a generalized time function but is also continuous,
is a temporal function if it is smooth and it's gradient is past-directed.

Any time or temporal function whose level sets are Cauchy hypersurfaces [Wald, Section 8.3] is called Cauchy.

Definition 103. *(Time function parametrization)*
Given $I \subseteq \mathbb{R}$, and a time function $\mathcal{T} : M \to \mathbb{R}$, define

$$\mathcal{C}_{\mathcal{T}}^{I} = \{\gamma \in \mathcal{C}(I, M) : \exists\, c_\gamma > 0, \mathcal{T}(\gamma(t)) - \mathcal{T}(\gamma(s)) = c_\gamma(t - s)\} \tag{1.99}$$

where $\mathcal{C}(I, M)$ is the set of future directed causal curves from I to M.
We endow $\mathcal{C}_{\mathcal{T}}^{I}$ with the compact-open topology (the sub-base $B(K, U) = \{f : f(K) \subseteq U\}$ where K is compact and U is open).

The main contributions from [Miller] are summarized in the following results:

Theorem 104. *(Temporal re-parametrizations and the compact-open topology)*
Let (M, d, ℓ) be a smooth spacetime and let $\mathcal{T} : \mathbb{R} \to M$ be a time function (in the sense of Definition 102), then $\mathcal{C}_{\mathcal{T}}^{I}$ is a Polish space.

Theorem 105. *(Causality and $\mathcal{T}$)*
In a globally hyperbolic smooth spacetime (M, d, ℓ), let $\mathcal{T} : M \to \mathbb{R}$ be a Cauchy temporal function, let $I \subseteq \mathbb{R}$ be an interval and for $t \in I$ let $\mu_t \in \mathcal{P}(M)$ with $\mathrm{spt}(\mu_t) \subseteq \mathcal{T}^{-1}(t)$, then the following are equivalent

1. *$\mu_s \preceq \mu_t$, for all $s \leq t$*
2. *There exists $\sigma \in \mathcal{P}(\mathcal{C}_{\mathcal{T}}^{I})$ s.t. $e_t \# \sigma = \mu_t$ for all $t \in I$.*

Theorem 106. *(Causality and conditioning)*
Let (M, d, ℓ) and $\mathcal{T}$ as above, define

$$\mathcal{I}_{\mathcal{T}} = \{\gamma \in \mathcal{C}_{\mathcal{T}}^{\mathbb{R}} : \mathcal{T} \circ \gamma = Id_{\mathbb{R}}\}$$

then the following statements are equivalent

1. *$\mu_s \preceq \mu_t$ and $\mathrm{spt}(\mu_t) \subseteq \mathcal{T}^{-1}(t)$*
2. *There exists $v \in \mathcal{P}(\mathcal{I}_{\mathcal{T}})$ s.t. $(e_t \mid \mathcal{I}_{\mathcal{T}}) \# v = \mu_t$.*

For the proofs see [Miller, Theorems 1,2]. The main use of $I_{\mathcal{T}}$ is due to it's bijection with the set of future inextendible causal curves.
In the following section (1.3.9) we will see how to use this results in terms of Markovianity of measures and their connection to the general Schrödinger problem. An immediate consequence of Theorem 105 is the fact that the Schrödinger problem in $\mathcal{P}(C_{\mathcal{T}}^{I})$ is solved by gluing solutions for all intermediate times.

Proposition 107. *(Push forward in $\mathcal{P}(\mathcal{C}_{\mathcal{T}}^{I})$)*
Let $R \in \mathcal{P}(C_{\mathcal{T}}^{I})$ and assume that $\mathrm{spt}(e_t \# R) \subseteq \mathcal{T}^{-1}(t)$ and $e_s \# R \preceq e_t \# R$ if $s \leq t$ then the solution to

$$\min_{\substack{\Pi \in \mathcal{P}(C_{\mathcal{T}}^{I}) \\ e_0 \# \Pi = \mu_0 \\ e_1 \# \Pi = \mu_1}} \mathrm{Ent}(\Pi \mid R)$$

is obtained by solving $e_t \# \Pi = \mu_t$ which always has a solution where

$$\mu_t \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}(C_{\mathcal{T}^I})} \mathrm{Ent}(\mu \mid e_t \# R). \tag{1.100}$$

Proof. Note that if μ_t is a solution in (1.100) then if entropy is finite, we have by absolute continuity

$$\mathrm{spt}(\mu_t) \subseteq \mathrm{spt}(e_t \# R) \subseteq \mathcal{T}^{-1}(t)$$

and also $\mu_s \preceq \mu_t$ and so by Theorem 105 there exists $\Pi \in \mathcal{P}(\mathcal{C}_{\mathcal{T}}^I)$ such that $e_t \# \Pi = R$. By optimality, for every μ

$$\mathrm{Ent}(e_t \# \Pi \mid e_t \# R) \leq \mathrm{Ent}(\mu \mid e_t \# R)$$

and by (39)

$$\mathrm{Ent}(e_t \# \Pi \mid e_t \# R) \leq \mathrm{Ent}(\Pi \mid R).$$

∎

Proposition 107 shows that if we define the Schrödinger problem in the space of paths parametrized by a Cauchy time function we are also able to recover the stander techniques from the classical setting from [Leonard2014].

Pushing forward the Hausdorff measure via timelike curves

Suppose that we do not want to assume the existence of a temporal Cauchy function, in this section we focus on giving a definition of a topology (via convergence) on the space of causal curves from $[0,1]$ to M different to the C^0 topology of definition 98. The objective is to define a topology adapted to the underlying physical structure of spacetimes with enough properties to be useful in terms of the Relativistic Schrödinger problem.

Definition 108. *($\mathcal{H}^1$-convergence)*
Let (M, d, ℓ) be a ghcrlls (Definition 20), we define convergence on timelike paths of the form $\gamma : [0,1] \to M$ which are right-continuous, denoted $\Omega^D_{x,y}$. We say that $\{\gamma_n\} \in \Omega^D_{x,y}$ converges to $\gamma \in \Omega^D_{x,y}$ in $\mathcal{H}^1$ sense, and write

$$\gamma_n \xrightarrow{\mathcal{H}} \gamma \text{ iff } \gamma_n \# \mathcal{H}^1 \rightharpoonup \gamma \# \mathcal{H}^1$$

where $\rightharpoonup$ refers to weak convergence for probability measures on M and $\mathcal{H}^1$ is the 1-dimensional Hausdorff measure on $[0,1]$.

In other words, $\gamma_n \xrightarrow{\mathcal{H}} \gamma$ if for every $f \in C_b((M,d),[0,1])$ we have

$$\lim_{n\to\infty} \int f(\gamma_n(t)) d\mathcal{H}^1(t) \to \int f(\gamma(t)) d\mathcal{H}^1(t).$$

Remark 109. *(Uniqueness on $\Omega_{x,y}$)*
Observe that $\gamma_1 \# \mathcal{H}^1 = \gamma_2 \# \mathcal{H}^1$ implies $\gamma_1 = \gamma_2$ $\mathcal{H}^1$-a.s. which implies $\gamma_1 = \gamma_2$ when they are continuous.

Proposition 110. *($\mathcal{H}^1$ is weaker than point-wise)*
In the ghcrlls context as above, assume that $\mathcal{H}^1$-a.e $\gamma_n(t) \to \gamma(t)$ pointwise as $n \to \infty$, with respect to d, then $\gamma_n \xrightarrow{\mathcal{H}} \gamma$.

Proof. The result is an application of dominated convergence as for every $f \in C_b(M,[0,1])$ we have

$$\lim_{n\to\infty}\int f d\gamma_n \# \mathcal{H}^1 = \lim_{n\to\infty}\int f(\gamma(t))d\mathcal{H}^1(t) = \int f(\gamma(t))d\mathcal{H}^1(t) = \int f d\gamma \# \mathcal{H}^1.$$

∎

Proposition 111. *(Tightness is inherited)*
Every sequence $\{\gamma_n\}$ of d-continuous paths in $\Omega_{x,y}$ admits a measure $\mu \in \mathcal{P}(M)$ as subsequential limit and $\mu = \gamma\#\mathcal{H}$ for a path $\gamma : [0,1] \to M$ which is causal and continuous except on countably many points.

Proof. The sequence of measures $\{\gamma_n \# \mathcal{H}^1\} \subseteq \mathcal{P}(M)$ are all concentrated in the compact set $J(x,y)$, by Prokhorov's theorem this tight sequence of measures is weakly-sequentially compact. Let n_k be the convergent subsequence, by diagonalization there exists a subsequence n_{k_m} of n_k which converges at all rational points in $[0,1]$. Define for every rational $t_k \in [0,1]$

$$\gamma(t_k) = \lim_{m\to\infty} \gamma_{n_{k_m}}(t_k). \tag{1.101}$$

γ is d-rectifiable and hence extends to a curve which is continuous except at countably many points. Let $\epsilon > 0$, at every point $s \in (0,1)$ of continuity of γ there exist p_ϵ, q_ϵ rational points in $[0,1]$ with $J^+(\gamma(p_\epsilon)) \cap J^-(\gamma(q_\epsilon)) \subseteq B_\epsilon(x(s))$. By convergence, for m large enough $J^+(\gamma_{n_{k_m}}(p_\epsilon)) \cap J^-(\gamma_{n_{k_m}}(q_\epsilon)) \subseteq B_\epsilon(\gamma(s))$ implying that $\mu = \gamma\#\mathcal{H}$. ∎

1.3.7 A transference-plan construction with the Hausdorff measure

Similar to section 1.3.6, motivated by the theory of mass transportation we define a mode of convergence for causal curves from $[0,1]$ to M in terms of weak convergence of transference plans concentrated on the image of the curves.

Definition 112. *(Convergence with transference plans)*
Let (M,d,ℓ) be a ghcrlls as in Definition 20, we say that a sequence of causal curves $\{\gamma_n\}$ converges to γ in terms of concentrated transference plans, denoted $\mathcal{H}\otimes$ via

$$\gamma_n \xrightarrow{\mathcal{H}\otimes} \gamma \text{ iff } (Id\times\gamma_n)\#\mathcal{H} \rightharpoonup (Id\times\gamma)\#\mathcal{H} \tag{1.102}$$

where again $\rightharpoonup$ denotes weak convergence and $\mathcal{H}$ is the 1-dimensional Hausdorff measure.

Proposition 113. *($\mathcal{H}^1\otimes$ and pointwise convergence)*
In the context of ghcrlls (M,d,ℓ) as above if $\{\gamma_n\}_n$ is a sequence such that as $n\to\infty$ we have $\gamma_n \to \gamma$ d-pointwise $\mathcal{H}^1$-a.e. then $\gamma_n \xrightarrow{\mathcal{H}\otimes} \gamma$.

Proof. Let $f \in C_b([0,1]\times M)$ then

$$\lim_{n\to\infty}\int f(x,y)d(Id\times\gamma_n)\#\mathcal{H} = \lim_{n\to\infty}\int_0^1 f(t,\gamma_n(t))d\mathcal{H}^1(t) = \int_0^1 f(t,\gamma(t))d\mathcal{H}^t(t)$$

again by Dominated convergence and continuity of f. ∎

Proposition 114. *(Relation between topologies)*
If $\gamma_n \xrightarrow{\mathcal{H}\otimes} \gamma$ then $\gamma_n \xrightarrow{\mathcal{H}} \gamma$.

Proof. By definition every $f \in C_b(M,[0,1])$ can be associated with $(t,f(t)) \in C_b(([0,1]\times[0,1]))$ from which the convergence of integrals is concluded. ∎

Proposition 115. *(Tightness)*
Every sequence of continuous causal curves $\{\gamma_n\}$ is sub-sequentially $\otimes\mathcal{H}$ convergent to a measure μ on $[0,1]\times M$. Furhtermore, there exists a d-rectifiable causal path γ such that $(Id\times\gamma)\#\mathcal{H} = \mu$.

The proof is the same as the tightness criteria in the previous section and hence omitted.

1.3.8 Bridges from Strong Markov processes

With our understanding of the different topologies on the space of causal curves and with Proposition 100 in mind we aim to study probability measures on the space of causal curves with fixed endpoints. These measures are referred to as Bridge measures (as we did in Proposition 44). Bridge measures have been long studied and our goal is to study different alternatives for Definition 74.
Let us start by recalling the classical approach to Markov Bridges. In this section we follow [Fitzsimmon-Pitman-Yor].
Let $(\Omega,\mathcal{F},\{\mathcal{F}_s\}_{s\in\mathbb{R}},\mathbb{P})$ be a filtered probability space, any process $\{X_t\}_{t\in[0,\infty)}$ such that for every stopping time $\tau:\Omega\to\mathbb{R}$

$$\mathbb{P}(X_{\tau+t}\in A|\mathcal{F}_\tau) = \mathbb{P}(X_{[t,\infty]}\in A|X_\tau) \tag{1.103}$$

is called a strong Markov Process. The (x,y,t)-Bridge was defined in [Fitzsimmon-Pitman-Yor] on Lusian spaces (homeomorphic to a Borel subset of a compact metric space) we look at a slightly less general version:

Definition 116. *($(x,y,t)-$ Bridge)*
Let X_t be a right-continuous strongly Markov Process (as in (1.103)) taking values on a compact metric space E, given $x,y\in E$ and $t\in\mathbb{R}$, and assume that $\{\mathcal{F}_t\}$ is the (uncompleted) natural filtration from X. Assume the existence of a Borel semigroup P_t, a dual variable $\hat{X}$ with associated semigroup $\hat{P}_t$ with respect to $L^2(m)$ of an invariant measure m, i.e.,

$$\int f(P_t g)dm = \int (\hat{P}_t f)g dm.$$

for all Borel functions f,g. Under these conditions there exists a well-defined kernel for the semigroup $p_t(x,y)$ satisfying Chapman-Kolmogorov equation. Set for $0\le s\le t$

$$H_s = p_{t-s}(X_s,y)$$

then the pre-measure $Q^t_{x,y}$ given by

$$Q(A) = \frac{1}{p_t(x,y)}\int_A H_t d\mathbb{P}_x$$

can be extended to a measure $\mathbb{P}^t_{x,y}$ on the sigma-algebra $\mathcal{G}$ generated by $\cup_{0\le s\le t}\mathcal{F}_s$. This probability measure is called the law of the (x,y,t)-Bridge from X and the canonical process associated to $\mathbb{P}^t_{x,y}$ on $\mathcal{G}$ is called the (x,y,t)-bridge of X.

Theorem 117. *(The strong Markov Property is inherited)*
Let be a $\{X^t_{x,y}(s)\}_{0\le s\le t}$ be the (x,y,t)-Bridge (as in Definition 116) from a strongly Markov process $\{X_t\}$ then $\{X^t_{x,y}(s)\}_{0\le s\le t}$ is also strongly Markov.

For a proof see [Fitzsimmon-Pitman-Yor, Proposition 1] where it is also shown that the $(x, y, t)-$ Markov Bridge has a transition kernel given by

$$p^{y,t}((z_1, t_1), (z_2, t_2)) = \frac{p_{t_2-t_1}(x_1, x_2)p_{t-t_2}(x_2, y)}{p_{t-t_1}(x_1, y)} \tag{1.104}$$

and that $\mathbb{P}^t_{x,y}$ is a regular version of conditional probabilities $\mathbb{P}_x(\cdot \mid X_t = y)$.
The simple formula (1.104) is essential to the study of Large deviation principles for Bridge measures and was the key element in [Hsu1990] to obtain the large deviation principle of Brownian bridges of Riemannian manifolds.

1.3.9 The Markov property and filtrations

Our study on globally hyperbolic chrono-regular Lorentzian length-spaces (Definition 20) says that the underlying topology to be considered should be the chronological topology, i.e. the coarsest topology including all cones (Definition 19.2). Given a topology, there is no unique way to measure distance between curves, if the topology is induced by a metric $\tilde{d}$, then it is common to endow $\Omega_{x,y} = \{f \in C^{\tilde{d}}([0, 1], M) : f(0) = x, f(1) = y\}$ with the sup topology i.e. we recall (1.2)

$$d^\infty(\sigma, \tilde{\sigma}) = \sup_{s\in[0,1]} \tilde{d}(\sigma(s), \tilde{\sigma}(s)).$$

In the case of globally hyperbolic chrono-regular Lorentzian length-spaces, the only assumption we have is that the chronological topology is metrizable (although the definition depends on d by determining rectifiability of paths). If we aim to study properties of the underlying space M, we should restrict ourselves to infer only topological properties from any metrization of the topology. Properties like completeness are inherent to the distance and not the topology as two metrics giving the same topology are not equivalent in the sense of metrics and a complete and an incomplete metric can give the same topology.
For the study of RDSch, we need to study Borel probability measures on $\Omega_{x,y}$, but the Borel σ-algebra depends on our definition of topology (as it is the smallest σ-field containing open sets). This observation leads us to study *bridge spaces* $\Omega_{x,y}$ endowed with *different topologies.*

Definition 118. *(Causal bridge spaces and topologies)*
Given a globally hyperbolic chrono-regular Lorentzian space (M, d, ℓ), let $x \ll y$, we have defined

$$\Omega_{x,y} = \{\sigma \in C([0, 1], M) : \sigma(0) = x, \sigma(1) = y, \sigma \text{ is causal}\}$$
$$\Omega^D_{x,y} = \{\sigma \in \mathrm{Cad}([0, 1], M) : \sigma(0) = x, \sigma(1) = y, \sigma \text{ is causal}\}$$

where $C([0, 1], M)$ ($\mathrm{Cad}([0, 1], M)$) correspond to the continuous (resp. cad-lag) functions from $[0, 1]$ onto M with the chronological topology.
Given any topology τ on $\Omega_{x,y}$ (resp $\Omega^D_{x,y}$) we call the topological spaces

$$C^\tau_{x,y} := (\Omega_{x,y}, \tau) \tag{1.105}$$
$$\mathrm{Cad}^\tau_{x,y} := (\Omega^D_{x,y}, \tau). \tag{1.106}$$

the continuous (resp. cad-lag) Bridge space from x to y endowed with τ.
We write $\Omega_{A,B} = \{\sigma \in C([0, 1], M) : \sigma(0) \in A, \sigma(1) \in B, \sigma \text{ is causal}\}$ if $J^+(A) \subseteq B$ and $J^-(B) \subseteq A$.

Although the most general case of study corresponds to cad-lag functions $\mathrm{Cad}^\tau_{x,y}$ we will focus mostly on continuous ones $C^\tau_{x,y}$. We discuss this choice, it's consequences and further explorations on section 1.5.2

Proposition 119. *(Passing causality)*
In the framework of ghcrlls $(M, d\,\ell)$, if ℓ^+ is τ-upper-semi-continuous in the following sense: if $\{\sigma_n\} \in C^\tau_{x,y}$

$$\sigma_n \xrightarrow{\tau} \sigma \Rightarrow \ell^+(\sigma(s),\sigma(t)) \geq \lim_{n\to\infty} \ell^+(\sigma_n(s),\sigma_n(t)) \text{ for every } s,t \in [0,1], s \leq t. \tag{1.107}$$

then every τ-limit of causal curves inherits causality.

Proof. The proof of is immediate from the definition of $C^\tau_{x,y}$ as the inequality (1.107) gives $\ell(\sigma(s),\sigma(t)) \geq 0$. ∎

The condition is not enough for $C^\tau_{x,y}$ to be a closed set (needed so that we can study it's properties as a topological space). We need to ensure that limits of continuous causal curves stay continuous and causal. In the case of $\mathrm{Cad}^\tau_{x,y}$ it is easier to be closed as τ-limits of cadlag curves are more likely to be cadlag.

Definition 120. *(Uniformity condition)*
Let τ be a topology on $\Omega_{x,y}$ we say τ satisfies a uniformity condition if the topology τ satisfies

$$\sigma_n \xrightarrow{\tau} \sigma \Rightarrow \sigma_n \xrightarrow{d_\infty} \sigma. \tag{1.108}$$

where again d_∞ is as in Definition 1.2.

Proposition 121. *If τ satisfies the uniformity condition 120 and ℓ^+ is upper semi-continuous with respect to $\mathcal{I}$, then $\Omega^\tau_{x,y}$ is closed.*

Proof. By proposition 119 every τ-limit is also causal. Uniform condition 120 and closedness of $\Omega_{x,y}$ under d^∞ norm yield the closedness of $C^\tau_{x,y}$. ∎

Clearly, if we describe $\mathcal{I}$ (the chronological topology from Definition 19.2) by a specified metric d and τ denotes the supremum norm (1.2), then τ satisfies Definition 120. On the other hand, propositions 110 and 114 say these topologies are weaker than pointwise convergence and so this technique is not available for closedness of $C^\tau_{x,y}$ with these topologies.
In general, $C^\tau_{x,y}$ is only a topological space but we can study $\mathcal{B}^\tau$ the Borel-σ with respect to τ. Given any $T \subseteq [0,1]$ we denote by $\mathcal{B}^\tau_T$ the Borel σ-algebra associated to the restriction topology (of τ to curves with domain T). In $\Omega^\tau_{x,y}$ we can always define the evaluation map via

$$e_t : \Omega^\tau_{x,y} \to M, \;\; e_t(\omega) = \omega(t), \tag{1.109}$$

which we have used (for Definition 74 and in section 1.3.8). In the general context, e_t may not be adapted with the filtration generated by $\{\mathcal{B}^\tau_{[0,s]}\}_{\{0\leq s\leq t\}}$. The usual definition of the Markov property for a measure requires conditionability with respect to the canonical filtration (i.e. the minimal σ-algebra on which the evaluation map is measurable) and factorization with respect to that σ-algebra.

Definition 122. *(Ball σ-algebra)*
Let (X, d) be a metric space, the ball σ-algebra denoted $\mathcal{B}_d$ is the σ-algebra generated by open balls in (X, d).

Denote by τ_d the topology generated by d in X, in general $\mathcal{B}_d \neq \mathcal{B}^{\tau_d}$, with $\mathcal{B}_d$ being smaller. In the case where (X, d) is separable, $\mathcal{B}_d = \mathcal{B}^{\tau_d}$ but most path spaces are not separable so the distinction is relevant for our purpose.

Proposition 123. *(The Borel σ-algebra and the evaluation map)*
If (M, d, ℓ) is a complete, metric spacetime and τ the Skorokhod topology [Billingsley, Chapter 12] on $\Omega_{x,y}$ (resp. $\Omega^D_{x,y}$) then

$$\mathcal{B}^\tau = \sigma(\{e_t\}_{t\in[0,1]}). \tag{1.110}$$

In the case of continuous curves $(\Omega_{x,y}, \tau)$ the topology is equivalent to that generated by d^∞ from (1.2).

For a proof see [Billingsley, Theorem 12.5]. The relevance of Proposition 123 is that one can recover the ball σ-algebra via the evaluation map indicating what the correct topology and sigma-algebra should be in the general case where we aim not to use the evaluation map.

The Schrödinger problem and different topologies

Similar to RDSch we now use the topology on $\Omega^\tau_{x,y}$ to describe a dynamical Schrödinger problem.

Definition 124. *(τ-Schrödinger Problem)*
Let (M, d, ℓ) be a ghcrlls let $R \in \mathcal{P}^\tau(\Omega)$ and assume $\mu_0 \preceq \mu_1$, define the τ-Dynamical Schrödinger problem for a reference τ-Borel measure R to be the minimization program

$$\min_{\substack{\Pi\in\mathcal{P}^\tau(\Omega)\\ e_0\#\Pi=\mu_0\\ e_1\#\Pi=\mu_1}} \mathrm{Ent}(\Pi|R) \tag{$\tau-$RDSCH}$$

Analogously, in the case where τ is metrizable by some metric d_τ, we can also define $\tau-$RDSCH over probabilities with respect to the $\sigma-$algebra $\mathcal{B}_{d_\tau}$ instead. This subtlety may seem insignificant at first but the difference between σ-algebras is not trivial.
By convexity of entropy (immediate from Lemma 29), the τ-Dynamical Schrödinger problem admits a unique solution in the case where $\mathcal{P}^\tau$ is also Polish. Such is the case of $C(p, q)$, which is compact as shown in [Wald, Theorem 8.3.9], it is second countable regular Hausdorff compact and hence metrizable. Existence of solutions for $\tau-$RDSCH depends on lower semi-continuity of entropy with respect to τ. If $(\Omega_{x,y}, \tau)$ is a Polish space then ($\tau-$RDSCH) almost corresponds to the static Schrödinger problem in such space (whose existence and uniqueness is known [Leonard2014], [Leonard2001] as in section 1.1) in the sense that almost every statement can be shown mutandis mutatis. This is the case for example for the topology in [Miller] by Theorem 104, yielding a different dynamical problem than (RDSch).
We notice that if the space of paths studied for ($\tau-$RDSCH) is Polish, the topology only plays a role through the τ-Borel-measurability of evaluation maps. The underlying features of ($\tau-$RDSCH) depend on the Polishness of the topology τ.
We aim to study other definitions of the Markov Property to avoid the unphysicality of the parameter $t \in [0, 1]$ parametrizing causal curves. The Markov property is one of the most studied concepts in

mathematics, we will focus on extensions of the causal Markov Property (74) which are adapted for our Schrödinger problem. A reasonable extension of the definition of Markovianity *for our purposes* should satisfy the following:

1. Encapsulate "past and future are independent given the present".
2. Take into account a"physical" topology on causal curves.
3. Be inherited by solutions of the Schrödinger Problem associated to the Borel measures of that topology.

We study these vague properties in the following sections.

Any definition using the evaluation map reduces to Léonard's

Standard assumptions on stochastic analysis are to consider only filtrations $\{\mathcal{F}_t\}$ for which the evaluation map is adapted. It is common to call a measure in path space the law of a stochastic process with respect to $\mathcal{F}_t$ if it's canonical process e_t is $\mathcal{F}_t$ adapted. Under this assumption Léonard noted that if one considers a different filtration and define an $\{\mathcal{F}_t\}$-Markov property then the law will be automatically Markov as in definition 74. In the following section we make no such assumption aiming to understand only the image of causal curves and not use the external parameter t as a time variable. Motivated by Proposition 123 let us formulate a slightly more general version on which we modify the interval.

Theorem 125. *(Skorohod)*
Let $C|_{[a,b]}$ (respectively $\mathrm{Cad}|_{[a,b]}$) denote the continuous (cadlag) curves from $[a,b]$ to M. If T is a dense subset of $[a,b]$, if τ is the Skohorod topology then

$$\sigma(e_t : t \in T) = \mathcal{B}^{\tau}_{[a,b]}.$$

where $\mathcal{B}^{\tau}_{[a,b]}$ denotes the ball σ-algebra for the Skorohod topology.

This version with it's proof can be found in [Billingsley, Theorem 12.5], for notational purposes define also $\mathcal{B}^{\tau}_t := \cap_{s>0}\mathcal{B}^{\tau}_{[t,t+s]}$ the right-completion.

Definition 126. *(τ-Markov property)*
We say that a τ-Borel measure $\nu \in \mathcal{P}(C^{\tau}_{x,y})$ is τ-Markov if it is conditionable with respect to the $\mathcal{B}^{\tau}_t$ filtration and for every $t \in [0,1]$ and $A \in \mathcal{B}^{\tau}_{[t,1]}$

$$\nu(A|\mathcal{B}^{\tau}_{[0,t]}) = \nu(A|\mathcal{B}^{\tau}_t) \tag{1.111}$$

Definition 126 is unsatisfactory as the σ-algebras involved $(B^{\tau}_{[0,t]}, \mathcal{B}^{\tau}_t)$ are in general too big. The definition coincides with definition 74 only because of Theorem 125. The "size" of these topologies lead us to analyze the physical topologies and a Markov definition in every case.

Definition 127. *(p-Future, past and present of a causal curve)*
Let $[\gamma]$ be an equivalence class (with the Standard equivalence Definition 97)), let $p \in J^+(x) \cap J^-(y)$ we define the section of $[\gamma]$ in the future of p as a map

$$\mathbf{F}_p : (\Omega_{x,y}/\sim) \to \Omega$$

to be the mapping

$$\mathbf{F}_p([\gamma]) = \mathrm{Im}(\gamma) \cap J^+(p) \tag{1.112}$$

similarly define $\mathbf{P}_p$ *the past of* $[\gamma]$ *via*

$$\mathbf{P}_p([\gamma]) = \mathrm{Im}(\gamma) \cap J^-(p). \tag{1.113}$$

Finally, define the p-present $\mathbf{Pr}_p := \mathbf{F}_p \cap \mathbf{P}_p$.

Definition 128. *(p-Markov C^0 topology)*
We say that ν *is p-Markov in* $\mathbf{F}_p$-$\mathbf{P}_p$ *sense if for every* $p \in J^+(x) \cap J^-(y)$ *the measure* ν *is conditionable and*

$$\nu(\mathbf{F}_p \in A \mid \mathbf{P}_p) = \nu(\mathbf{F}_p \in A \mid \mathbf{Pr}_p) \tag{1.114}$$

Proposition 129. *($\mathcal{B}^\tau_{x,y} - \mathcal{B}^\tau_{p,y}$-measurability of $\mathbf{F}_p$)*
For every $p \in J^+(x) \cap J^-(y)$ *the maps* $\mathbf{F}_p, \mathbf{P}_p$ *and* $\mathbf{Pr}_p$ *are* $\mathcal{B}^\tau_{x,y} - \mathcal{B}^\tau_{p,y}$*-measurable.*

Proof. Enough to show that $\mathbf{F}_p$ is actually continuous, for any γ consider U open set such that $\mathcal{F}_p([\gamma]) \subseteq U$ then $M \setminus (J^+(p) \setminus U)$ is closed and $[\gamma]$ is compact so by the T_3-property of the topology, there exist B_{ϵ_i} where $i = 1, \dots, n$ where $B_{\epsilon_i} \cap M \setminus (J^+(p) \setminus U) = \emptyset$ and if

$$V := \bigcup_{i=1}^{n} B_{\epsilon_i} \tag{1.115}$$

then $\gamma \in V$ and $F_p(V) \subseteq U$ showing continuity of $\mathbf{F}_p$. ∎

Measurability of $\mathbf{F}_p$ is necessary for the previous definition to make sense.
Let $\sigma : [0,1]$ be **any** $[0,1]$ parametrization of $[\gamma]$, assume that $p \in [\gamma]$ then there exists t^* such that $\sigma(t^*) = p$ and observe that

$$\mathbf{F}_p([\gamma]) = \sigma\mid_{[t^*,1]}$$
$$\mathbf{P}_p([\gamma]) = \sigma\mid_{[0,t^*]}$$

and so (1.114) is exactly the one in Definition 74 (as it is required for every t and hence includes all parametrizations) which means that there is nothing new to learn from this "apparently physical" property.

Definition 130. *(Transference Markov on Cadlag curves)*
Let $\nu \in \mathcal{P}^\tau(\mathrm{Cad}^\tau_{x,y})$ *we define the* τ*-Markov property with respect to* $\otimes\mathcal{H}$ *if it is conditionable and for all t*

$$\nu(e_{[t,1]} \in A \mid e_{[0,t]} = \nu(e_{[t,1]} \in A \mid e_t) \tag{1.116}$$

Note that this definition is almost exactly Definition 74 with the slight-difference of considering only τ-Borel measures and henceforth we need to know the evaluation map is measurable there. Note that this definition is essentially Definition 74 as soon as the evaluation map is measurable.

Definition 131. *($\mathcal{T}$-Markov)*
A measure $\nu \in \mathcal{P}(C^I_{\mathcal{T}})$ *is called Markov with respect to the topology described in Theorem* (104) *if it is conditionable with respect to* e_t *and*

$$\nu(e_{[t,1]} \in A \mid e_t) = \nu(e_{[t,1]} \in A \mid e_{[0,t]}) \tag{1.117}$$

Again the subtle difference with respect to Definition 74 is the set on which ν belongs to $\mathcal{P}(C_{\mathcal{T}}^I)$. The measurability (in fact continuity) of the evaluation map with respect to the associated Borel σ-algebra is shown in [Miller].

The definition of 1.111 allows us to incorporate other information of the "past" and the "future" coming from τ. This means that τ-Markov is an analogue of the Markov property adapted to the causality on a globally hyperbolic chrono-regular Lorentzian length space through the topology of (causal) curves.

Note that we are setting as part of the definition that regular conditional probabilities must exist, by saying that ν *must be conditionable*.

Observe that by proposition 123 in the case of the topology τ being the uniform topology induced by a complete, separable metric, the τ Markov-Property is exactly the unphysical Markov Property (1.48).

Motivated again by Proposition 123 we generalize a different version on which we make the assumptions necessary for the τ-Markov property to be inherited in the Schrödinger problem.

Assumption 3. *(Realizability of the τ-Borel topology)*
Given τ a topology on $\Omega_{x,y}$ assume the evaluation maps $e_t, e_{[0,t]}, e_{[t,1]}$ are $\mathcal{B}^\tau - \mathcal{B}(M)$ measurable for every $t \in [0,1]$.

Remark 132. *As shown in the previous section, Assumption 3 holds in all the topologies of interest so far. The question of whether or not there exist physically relevant topologies for $C^\tau_{x,y}$ or $\mathrm{Cad}^\tau_{x,y}$ for which Assumption 3 does not hold remains open. When assumption 3 holds, Leonard's version of the classical definition of Markovianity (Definition 74) should be used.*

1.3.10 The consequence of Markovianity

In this section we follow the theory developed on [Leonard2014] and generalize them to our context. The main motivation of this section is that the fundamental lemma for intermediate times on [Leonard2014] is already general enough. The following lemma can be found in [Leonard2014] (A.8).

Lemma 133. *Let $\phi : X \to Y$ be a measurable map between polish spaces X, Y and let P, R be two Borel probability measures then we have*

$$\mathrm{Ent}(P \mid R) = \mathrm{Ent}(\phi\#P \mid \phi\#R) + \int_Z \mathrm{Ent}(P(\cdot \mid \phi = z) \mid R(\cdot \mid \phi = z))d(\phi\#P)(z) \tag{1.118}$$

where Z is the range of ϕ.

Proof. Apply directly the disintegration Theorem (eg. [Bogachev, Theorem 10.5.6]) which applies as ϕ is measurable, P, R are Borel and the spaces are Polish. ∎

Lemma 134. *(Intermediate optimality)*
Assume that $\Omega_{x,y}$ is Polish, under Assumption 3, assume further that

$$\sigma(\{e_s : o \le s \le t\}) = \mathcal{B}^\tau_{[0,t]},\ \sigma(\{e_s : t \le s \le 1\}) = \mathcal{B}^\tau_{[t,1]}$$

the solution to the entropic problem

$$\min_{P \in G_t(\mu, Q_1, Q_2)} \mathrm{Ent}(P \mid R) \tag{1.119}$$

where $G_t(\mu, Q_1, Q_2)$ *is the set of* τ*-Borel probability measures for which*

$$f_t \# P = \mu,\ g_t \# P = Q_1,\ h_t \# P = Q_1 \tag{1.120}$$

is given by the measure

$$P^*(\cdot) = \int Q_1 \otimes Q_2(A) d\mu(z) \tag{1.121}$$

For a proof see [Leonard2014, Proposition 2.10].

Proposition 135. *(τ-Markovianity is inherited)*
Under assumption 3 and the hypothesis of the previous Lemma, if Π *is a solution of* $\tau-RDSCH$*, where* R *is Markov as in definition 74, then* Π *is also Markov.*

Proof. If Π is a solution of $\tau-$RDSCH then at every $t \in [0,1]$ we apply Lemma 134 which yields (1.111) as in the proof of Léonard in [Leonard2014]. ∎

1.3.11 Discussion of topologies and Markovianity

In this section we explored the consequences of different topological approaches to the definition of Causal Markovianity. We first learnt the universality of Definition 74 used by C. Leonard in section 1.3.9. We also showed via Theorem 125 that it corresponds to the definition of R. Dudley (Definition 70). Consequently, it is reasonable to use this description to formulate the causal Markov property and therefore obtain a non-existence (Corollary 72) result for Lorentz-invariant Markov Processes. Nevertheless, we realize that we can write a definition of the Markov property even if the evaluation maps are not measurable (Definition 1.111). This generalization is useful for the topologies of curves in path space available in physics literature. In section 1.3.9 we considered a general version of the Schrödinger problem which considers these topologies only to realize that the general techniques of [Leonard2012] can still apply in most cases, demonstrating the significance of these tools in a more general setting. We have described this technique in physically relevant frameworks in section 1.3.9 but the question of constructing a hierarchy and characterization for these generalized Markov properties is left for future work as described in sections 1.5.2 and 1.5.2.

1.4 Large deviations

We arrive to the study of Large Deviation Principles and it's connections with RDSch. Throughout this work we have hinted (as it's well known in the theory of the Schrödinger Problem (see [Leonard2014], [Leonard2012], [Tamanini], [Gigli-Tamanini]) that in order to recover optimal transference plans through entropic regularizations, the reference measure for the entropic problem must satisfy a large deviation principle on which the rate function is given by the Lagrangian of the transport cost. In this section we make all these notions precise together with analyzing (if any) the large deviation principles satisfied by the process involved in sections 1.1 - 1.3.
In essence a collection of measures satisfies a Large Deviation Principle when their logarithms have a specific order of convergence, intuitively we say X_n satisfies a Large deviation principle with rate function I when

$$\Pr(X_n \in A) \asymp \exp\left(-n \min_{x \in A} I(x)\right). \tag{1.122}$$

The informal equation 1.122 is interpreted by thinking that as $n \to \infty$ the probability of X_n belonging to A becomes exponentially small where the rate of this exponential is given by the minumum of I. We proceed to describe rigorously Large Deviation Principles and we will do it in an abstract Hausdorff space X as the keen reader will guess we aim to apply Large deviations for the space-time (M) in RDSch but also for the phase-space version ($\mathcal{PS}$) for RKSchP and the path spaces $C^{\tau}_{x,y}$,$\mathrm{Cad}^{\tau}_{x,y}$, Ω.

Definition 136. *(LDP in a Hausdorff space)*
Let X be a Hausdorff space, we say that a collection of measures $\{\mu_\epsilon\}_{\epsilon>0}$ satisfies a weak large deviation principle with good rate function $I : X \to [0,\infty]$ at rate $\{h_\epsilon\}$ if and only if

1. *For every closed set F and every open set G,*

$$\limsup_{\epsilon\to 0} h_\epsilon \log(\mu_\epsilon(F)) \leq -\inf_{x\in F} I(x), \tag{1.123}$$

$$\liminf_{\epsilon\to 0} h_\epsilon \log(\mu_\epsilon(G)) \geq -\inf_{x\in G} I(x). \tag{1.124}$$

2. *$\{x \in X : I(x) \leq c\}$ is compact for every $c \in \mathbb{R}$*

In this case we say $\{\mu_\epsilon\}$ satisfies $LDP(\mu_\epsilon, h_\epsilon, I)$.

Condition 136.2 is usually referred to as I being a *good* rate function. Goodness of rate functions is often referred to as being coercive in the analysis literature.
Intuituively, Definition 136 should remind the reader of Portmanteau's theorem in which the characterization of weak convergence changes signs depending whether sets are close or open, this is a consequence of the fact that considering the two-sided limit to exist is too restrictive (see [Varadhan]) and justifies the term "weak". Sometimes it is convenient to consider the parametrization $\epsilon = 1/n$ and consider the limits in (1.123) and (1.124) as $n \to \infty$, in that case we write $\{\mu_n\}$ satisfies $LDP(\mu_n, r_n, I)$ and we make no distinction.
One of the fundamental results in the theory of Large Deviation Principles (LDP from now on) is Varadhan's Lemma whose content allows us to compute the limit of logarithms of integrals of exponentials.

Remark 137. *Before we state the general theorems from LDP let's build intuition for the limiting procedure. On $C[0,1]$ let $\mathbb{P}$ denote the Wiener measure, by definition we have,*

$$\mathbb{P}(\gamma(t_1) \in dz_1, \cdots, \gamma(t_n) \in dz_n) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{|z_{i+1}-z_i|^2}{t_{i+1}-t_i}\right\},$$

where $\propto$ means equal up to a normalizing constant.
If we were to only consider this limit, as the partition is refined we would get

$$-\frac{1}{2}\sum_{i=1}^{n} \frac{|z_{i+1}-z_i|^2}{t_{i+1}-t_i} = -\frac{1}{2}\sum_{i=1}^{n}(t_{i+1}-t_i)\left(\frac{|z_{i+1}-z_i|}{t_{i+1}-t_i}\right)^2 \xrightarrow{n\to\infty} -\frac{1}{2}\int_0^1 |\dot{\gamma}(t)|^2 dt$$

This is exactly the rate function for LDP associated to the slowed-down Brownian motion (Schilder's Theorem). The idea of LDP theory is to generalize this concept i.e. to study the convergence of the exponentially fast decrease in probability. It is of no surprise that if we can write a sequence of

measures in terms of exponential functions, we will recover LDP principles, this observation was made by Gibbs and we formulate it in Theorem 140.2.
Again informally, if L denotes a Lagrangian we aim to find $\mathbb{P}$ such that

$$\mathbb{P}(\gamma(t_1) \in dz_1, \cdots, \gamma(t_n) \in dz_n) \propto \exp\left\{\sum_{i=1}^{n}(t_{i+1}-t_i)L\left(\frac{z_{i+1}-z_i}{t_{i+1}-t_i}\right)\right\}. \tag{1.125}$$

to obtain (rigorously using a LDP) convergence to an expresion of the type

$$I(\sigma) = \int_0^1 L(\dot{\sigma}(t))dt.$$

when σ is absolutely continuous and ∞ otherwise.

1.4.1 Varadhan's Theorem

Theorem 138. *(Varadhan's Theorem)*
Suppose $LDP(\mu_n, r_n, I)$ holds on X where I is a good rate function and $f : X \to [-\infty, \infty]$ is continuous, if

$$\lim_{b\to\infty} \limsup_{n\to\infty} \frac{1}{r_n} \log \int_{\{f\geq b\}} e^{r_n f} d\mu_n = -\infty. \tag{1.126}$$

Then

$$\lim_{n\to\infty} \frac{1}{r_n} \log \int e^{r_n f} d\mu_n = \sup_{x: f(x)\wedge I(x)<\infty} \{f(x) - I(x)\}. \tag{1.127}$$

The proof can be found in [RassoulAgha-Seppäläinen, Section 3.2]. The reader should observe that the terms for which we are taking the limit on (1.127) are of the form of logarithms of integrals of exponentials, exactly like the ones appearing on Proposition 47.

Theorem 139. *(Varadhan's Theorem in the non-good case)*
Suppose $LDP(\mu_n, r_n, I)$ holds on X where I is a rate function (not necessarily good), X is regular Hausdorff and $f : X \to \mathbb{R}$ is continuous and bounded, then

$$\lim_{n\to\infty} \frac{1}{r_n} \log \int e^{r_n f} d\mu_n = \sup_{x\in X} \{f(x) - I(x)\}. \tag{1.128}$$

In the contest of Theorem 139, the bound condition 1.126 always holds by boundness of the function f. The proof relies in using boundness of f to generate balls on which f is (almost) bounded by integers [Kallenberg, Theorem 24.10].

1.4.2 The contraction Principle and Gibb's measures

In this section we summarize results from Large deviation principles that we will use in the rest of the chapter, these results are standard and can be found in [Varadhan], [RassoulAgha-Seppäläinen] or any standard reference in LDP.

Theorem 140. *Let X be a Hausdorff space and assume that $\{\mu_{\epsilon_n}\}$ satisfies a large deviation principle with rate ϵ_n and good rate function I, then*

1. *(Contraction Principle)*
   Assume that $f : X \to X'$ is measurable and X' is a different Hausdorff space, then the sequence of measures $\{f\#\mu_{\epsilon_n}\}_n$ satisfies a large deviation principle with rate ϵ_n and rate function given by
$$I'(x') = \inf\{I(x) : f(x) = x'\} = \inf_{x \in f^{-1}(x')} I(x). \tag{1.129}$$

2. *(Gibb's formula)*
   Let $F : X \to \mathbb{R}$ be bounded and continuous, then the measures
$$\mu_{\epsilon_n}^F(A) = \frac{1}{Z(F)} \int_A e^{1/(\epsilon_n F)} d\mu_{\epsilon_n} \tag{1.130}$$
   where
$$Z(F) = \int_X e^{1/(\epsilon_n F)} d\mu_{\epsilon_n}$$
   satisfy a LDP with rate function
$$I^F(x) = I(x) - F(x) + \sup_{y \in X}(F(y) - I(y)) \tag{1.131}$$

3. *(Conditional Large Deviation Principle)*
   Let $f : X \to X'$ and $x' \in X'$ are such there exist regular conditional measures $\mu_{\epsilon_n}(\cdot \mid f = x')$, if $\{\mu_\epsilon\}$ satisfies a LDP with rate function I then $\mu_{\epsilon_n}(\cdot \mid f = x')$ satisfy a LDP with rate function
$$I^{f,x'}(x) = \begin{cases} I(x) - I'(x) & \text{if } f(x) = x' \\ \infty & o.c. \end{cases} \tag{1.132}$$

Again, the proof is standard in the theory of deviation principles and can be found in [Varadhan] or [RassoulAgha-Seppäläinen].
The contraction principle Theorem 140.1 will be used together with (39), while Gibb's measures Theorem 140.2 are used to generate measures with specified rate functions. Note that if one starts with a constant sequence of measures, Theorem 140.2 gives a way to generate a sequence of measures of specified rate $I(x) := \sup_{x' \in X} F(x') - F(x)$.

Theorem 141. *(Chaganti's conditional LDP)*
Assume that X_1, X_2 are polish spaces (endowed with their Borel σ-algebras) and that $\mu_\epsilon^1 \in \mathcal{P}(X_1)$ satisfies LDP at rate ϵ with good rate function I. Let μ_ϵ be a sequence of probability measures on $X_1 \times X_2$ with
$$\mu_\epsilon(B_1 \times B_2) = \int_{B_1} \nu_\epsilon(y, B_2) d\mu_\epsilon^1(y) \tag{1.133}$$
where $\{\nu_\epsilon(x_1, \cdot)\} \in \mathcal{P}(X_2)$ satisfies LDP with rate $I^{x_1}(\cdot)$ and

1. *For every $x_1 \in X_1$ the rate function $I^{x_1}(\cdot)$ is good on X_2.*
2. *If $x_\epsilon^1 \to x_1$ in X_1 then $\nu_\epsilon(x_\epsilon^1, \cdot)$ satisfies LDP with rate I^{x_1}.*
3. *$(x, z) \to I^x(z)$ is lower semi-continuous in $X_1 \times X_2$.*

Let $\mathbf{I} : X_1 \times X_2$ *be given by*

$$\mathbf{I}(x_1, x_2) = I(x_1) + I^{x_1}(x_2). \tag{1.134}$$

If $\mathbf{I}(x_1, x_2)$ *is good on* $X_1 \times X_2$*, then* μ_ϵ *satisfies LDP at rate* ϵ *with rate function* $\mathbf{I}$*. In the case where* $\mathbf{I}$ *is not good, one obtains the upper bound of LDP for all compact sets and the lower bound still holds for every open set (i.e. the weak principle).*

For a proof see [Chaganti, Theorem 2.3]. The main idea of the proof is to use $F_\epsilon(\cdot) = \frac{1}{\epsilon}\nu_\epsilon(\cdot, B)$ and apply Varadahn's Theorem (Theorem 138) for which the continuity hypothesis allows us to take the limit.

Theorem 142. *(Dawson-Gartner finite dimensional principle)*
If $\{X_n\}$ *is exponentially tight and for each* $t_1, t_2, \dots t_m$ *then* $(X_n(t_1), \dots, X_n(t_m))$ *satisfies a LDP in* X^m *with rate function* $I_{t_1,t_2,\dots,t_m}(x)$ *then* $\{X_n\}$ *satisfies a LDP on* $D_X[0,\infty)$ *with good rate function*

$$I(\gamma) = \sup_{\{t_i\} \subseteq \Delta_\gamma^c} I_{t_1,\dots,t_m}(\gamma(t_1), \dots \gamma(t_m))$$

where Δ_γ^c *is the complement of set of discontinuities of* $t \to \gamma(t)$*.*

For a proof on the product topology see [Kallenberg, Theorem 24.12]. We combine [Dembo-Zeitouni, Theorem 4.6.1] and the strategy in [Dembo-Zeitouni, Section 5.1] to obtain our particular formulation. Note that in Theorem 142 if we restrict to continuous functions on $[0, 1]$ with the sup-norm there are no discontinuities and we obtain the supremum over all possible partitions. This formulation is also well-suited for the Cad-lag version when endowed with the Skorohod topology.

1.4.3 Large Deviation Principles for collections relevant to the Schrödinger Problem

In this section we study the large deviation principles for the collections of measures we have used along this document. We explain a technique to prescribe large deviation rates and use it towards our goal of studying entropic convergence in ghchrlls. It is important to note that it is not the only way to obtain small time asymptotics as we explain in section 1.5.2 another way (due to Ben Arous et. al.) to obtain a sequence of measures with a prescribed rate working in very general frameworks related to the theory of partial differential equations satisfying the condition of Hörmander. Due to the absence of the heat semigroup we choose to avoid techniques involving elliptic operators.

LDPs for the intrinsic constructions of section 1.3

In the seminal work of Hsu [Hsu1990], it was shown that in Riemannian manifolds, bridge measures associated to Brownian motion satisfy LDP with rate function

$$J^{x,y}(\gamma) = \frac{1}{2}\left(\int_0^1 |\dot{\gamma}_s|_g^2 ds - d(x,y)^2\right) + \iota_{\Omega_{ac}},$$

where Ω_{ac} is the set of absolutely continuous curves in Ω and ι is the convex indicator (see (1.148)). Following the ideas of [Hsu1990], we expect rate functions of Bridge measures to be of the form

$$I^{x,y}(\gamma) = L(\gamma) - c(x,y) \tag{1.135}$$

or similar where $L : \Omega \to (-\infty, \infty]$ represents a Lagrangian/action/minimizing principle. We also know from the proof of Proposition 46 that any solution of a dynamic Schrodinger problem shares its bridges with it's reference measure, that is, if P is optimal in the Polish space case for Dynamical Schrödinger then

$$P(\cdot) = \int R^{x,y} d\pi(x,y) \tag{1.136}$$

where π is a solution for the static problem and R is the reference measure.
Using these observations set

$$\mathbf{I}^{x,y}(\gamma) = \ell(x,y) - L_\ell(\gamma) + \iota_{\Omega_{x,y}} \tag{1.137}$$

where L_ℓ stands for the ℓ-length defined in section 1.1. Observe that the choice is fully justified in accordance to the strongly causal smooth spacetime case as the action of curves coincides with ℓ-length [Kunziger-Saemann, Proposition 2.32].

Proposition 143. *(A priori prescription of rate function for Construction 2)*
Consider $x \ll y$ in (M, d, ℓ) ghcrlls (Definition 20) and for any $a, b \in J^+(x) \cap J^-(y)$ define

$$\sigma^\epsilon_{a,b}(A) = C^{-1}_{a,b} \int_A \exp\left\{+\frac{1}{\epsilon}(\ell(a,z) + \ell(z,b))\right\} d\mathcal{H}_{a,b}(z) \tag{1.138}$$

where A is a Borel subset of $\mathrm{MID}(a,b)$, $C_{a,b}$ is the normalizing constant;

$$C_{a,b} = \int_{\mathrm{MID}(a,b)} \exp\left\{+\frac{1}{\epsilon}(\ell(a,z) + \ell(z,b))\right\} d\mathcal{H}_{a,b}(z),$$

and $\mathcal{H}$ is the non-trivial Hausdorff measure on each $\mathrm{MID}(a,b)$ from [McCann-Saemann].
Assume that there exists $\epsilon' > 0$ such that if $\epsilon < \epsilon'$, μ_ϵ the probability measure on $\Omega_{x,y}$ from construction 2 associated to $\{\sigma(a,b)^\epsilon\}$ is well-defined, then μ_ϵ satisfies LDP on $\Omega_{x,y}$ with rate function $\mathbf{I}^{x,y} : \Omega_{x,y} \to [0, \ell(x,y)]$ given by

$$\mathbf{I}^{x,y}(\sigma) = \ell(x,y) - L_\ell(\sigma). \tag{1.139}$$

Proof. By continuity of ℓ^+ (Lemma 17), $\ell(a,\cdot), \ell(\cdot,b)$ are continuous on $\mathrm{MID}(a,b)$ and so by definition of $\sigma^\epsilon_{a,b}$ we can apply Gibb's principle and obtain that $\{\sigma^\epsilon_{a,b}\}_{\epsilon>0}$ satisfies LDP at scale ϵ with rate function

$$I(z) = \sup_{w \in \mathrm{MID}(a,b)} \{\ell(a,w) + \ell(w,b)\} - \ell(a,z) - \ell(z,b) = \ell(a,b) - \ell(a,z) - \ell(z,b),$$

where the equality occurs because the space is assumed to be curve-connected and hence there always exists an ℓ-path between a and b which intersects $\mathrm{MID}(a,b)$ at a point $z^* \in \mathrm{MID}(a,b)$ where $\ell(a,z^*) = \ell(z^*,b) = \frac{1}{2}\ell(a,b)$.
If $Z_\epsilon \sim \sigma^\epsilon_{x,y}$ then we have shown LDP at rate ϵ with rate function

$$I(z) = \ell(x,y) - \ell(x,z) - \ell(z,y).$$

Recall that in construction 2 the value of $X^\epsilon(1/4)$ and $X^\epsilon(3/4)$ are given conditionally on $X^\epsilon(1/2)$. That is,

$$(X^\epsilon(1/4) \mid X^\epsilon(1/2) = z) \sim \sigma^\epsilon_{x,z}$$
$$(X^\epsilon(3/4) \mid X^\epsilon(1/2) = z) \sim \sigma^\epsilon_{z,y}$$

By Gibb's principle again we know that these measures satisfy LDP at rate ϵ with rate functions $I_1(x_1) = \ell(x,z) - \ell(x,x_1) - \ell(x_1,z)$ and $I_2(x_2) = \ell(z,y) - \ell(z,x_2) - \ell(x_2,y)$ respectively. Applying Theorem 141 we learn that $(X^\epsilon(1/4), X^\epsilon(1/2), X^\epsilon(3/4))$ satisfy LDP at rate ϵ with rate function

$$\begin{aligned} I(x_1,z,x_2) = \ell(x,y) &- \ell(x,z) - \ell(z,y) \\ &+ \ell(x,z) - \ell(x,x_1) - \ell(x_1,z) \\ &+ \ell(z,y) - \ell(z,x_2) - \ell(x_2,z). \end{aligned}$$

where we abuse the notation by using the same letter I.
Proceeding inductively, by Theorem 141, if $(t_1, t_2, \cdots, t_n) \in \mathcal{D}$ the joint measure of $(X_\epsilon(t_1), \cdots, X_\epsilon(t_n))$ satisfies a LDP with rate function

$$I(x_1, \ldots, x_n) = \ell(x,y) - \sum_{i=1}^{n-1} \ell(x_i, x_{i+1})$$

We apply Theorem 142 (exponential tightness is inherited as in [Dembo-Zeitouni, Theorem 5.1.2]), so the laws of the sequence $\{X^\epsilon\}$ as probability measures in $\Omega_{x,y}$ satisfy a LDP with rate function

$$\begin{aligned} I(\gamma) &= \sup_{\{t_k\}} I(\gamma(t_1), \ldots, \gamma(t_k)) \\ &= \ell(x,y) + \sup_{\{t_k\}} \left\{ -\sum_{k=1}^{n} \ell(\gamma(t_{k-1}), \gamma(t_k)) \right\} \\ &= \ell(x,y) - \inf_{\{t_k\}} \left\{ \sum_{i=1}^{n} \ell(\gamma(t_{k-1}), \gamma(t_k)) \right\} = \ell(x,y) - L_\ell(\sigma) \end{aligned}$$

where the last equation is just the definition of ℓ-length (5) as the supremum is taken over all partitions $\{t_k\}$ of $[0,1]$. ∎

Despite the intrinsically interesting properties of the construction, it is clear that it is not the only way to obtain a sequence of measures with prescribed rate function. We can start with any measure $\mu \in \mathcal{P}(\Omega_{x,y})$ and apply Gibbs principle directly, this is the content of the following Lemma.

Remark 144. *The assumption on Proposition 143 can be replaced by the following: there exists $\epsilon' > 0$ such that if $\epsilon < \epsilon'$ then there exist $\alpha(\epsilon), \beta(\epsilon) > 0$ such that*

$$\mathbf{E}[d(X_s, X_t)^{\beta(\epsilon)}] \leq K|t-s|^{1+\alpha(\epsilon)}$$

where $t \in \mathcal{D}$ and $X_t \sim \sigma^\epsilon$ in their respective mid-set.

Even though this assumption is expected to hold, a strategy for it's proof is not clear even in the simplest cases.

Lemma 145. *(A posteriori-prescription towards LDP)*
Let $\{\sigma(a,b)\}$ be any collection of measures for constructions 1 or 2 (or 1.3.5 in the general case). Consider μ the associated Bridge measure according to Proposition 62, define for $F : \Omega_{x,y} \to \mathbb{R}$ continuous and bounded (where d^∞ has been chosen for the topology on $\Omega_{x,y}$ and define μ_ϵ^F via Gibb's principle. Then $\{\mu_\epsilon^F\}$ satisfies a Large deviation principle with rate function

$$I(\sigma) = -F(\sigma) + \sup_{\gamma \in \Omega_{x,y}} F(\gamma). \tag{1.140}$$

In particular in the case where $\sigma \to L_\ell(\sigma)$ is d^∞-continuous we obtain a sequence of measures $\mu_\epsilon \in \mathcal{P}(\Omega_{x,y})$ satisfying a LDP with rate ϵ and rate function

$$\mathbf{I}^{x,y}(\sigma) = \ell(x,y) - L_\ell(\sigma). \tag{1.141}$$

Proof. The hypothesis are written to satisfy those of Theorem 140.2 so the conclusion follows from it. The form of the rate function follows from the assumption of curve-connectedness for which the supremum is achieved by the ℓ-curve. ∎

LDP and Dudley's process

The general case for the small-time asymptotics of the kernel asociated to Dudley's process is not known. Although some progress has been made to establish an anlogue of [Hsu1990, Theorem 2.2] for Dudley's process, the results aren't yet satisfactory. The specific case of $n = 1$ for the Kolmogorov operator is well known and we present it in the next section (section 1.4.3).
In an attempt to study the general behaviour of the hypoelliptic operator associated to Dudley's process, in [Franchi-LeJan2012] the author considers a simpler version of the problem $(\int_0^t \omega_s^2 ds, \omega_s)$ where ω_s is a 1-dimensional Brownian motion in $\mathbb{R}$. Even there, the author observes that there is no preferred scaling and only the specific formula of the square in the second coordinate allows some estimates contrary to the case of Brownian bridges where the time re-scaling $t \to \epsilon t$ is canonical. Dudley's operator is intimately connected to the hypo-elliptic Laplacian of Bismut [Bismut] where progress has been made in the case of general dimensions, nevertheless the approach of Bismut is of completely different nature and out of the scope of this document. We briefly explain the approach and discuss it in section 1.5.2.

Matsumoto Ikeda LDP for Dudley's one-dimensional diffusion

Let us denote by A, the 1-dimensional analogue of the operator in section 1.3.3, that is: In $\mathbb{R}^2$ set

$$A = \frac{1}{2}\frac{\partial^2}{\partial p^2} + p\frac{\partial}{\partial x} \tag{1.142}$$

In this $n = 1$ case, Dudley's process can be written in the form

$$\left(x_0 + \int_0^t p_s ds, p_0 + p_s\right) \tag{1.143}$$

where p_s is a Brownian motion on the line and x_0, p_0 are fixed. In [Matsumoto-Ikeda], the authors computed the explicit formula for the kernel associated to (x_t, p_t), that is:

$$p_t((x_1,p_1),(x_2,p_2)) = \frac{\sqrt{3}}{\pi t^2}\exp\left(-\frac{(p_2-p_1)^2}{2t} - \frac{6}{t^3}\left(x_2 - x_1 - \frac{(p_2+p_1)}{2}t\right)^2\right) \tag{1.144}$$

Note that one can now get explicit formulation for the heat kernel with respect to RKSchP by using equation (1.104) for the kernel (1.144).
In the $n = 1$ case, the authors decided to balance the "degeneracy" of the operator by adding the

noise to the position variable. This approach differs from [Dudley1966] and [Bismut] but generates a sequence of operators of the form:

$$A_\epsilon = \frac{1}{2}\frac{\partial^2}{\partial p^2} + \epsilon^2 \frac{1}{2}\frac{\partial^2}{\partial x^2} + p\frac{\partial}{\partial x} \tag{1.145}$$

for which the authors manage to prove that the Kernel $p^\epsilon(t,(x_1,p_1),(x_2,p_2))$ can be written explicitly:

$$\frac{\sqrt{3}}{\pi t\sqrt{t^2+12\epsilon^2}} \exp\left(-\frac{(p_2-p_1)^2}{2t} - \frac{6}{t(t^2+12\epsilon^2)}\left(x_2 - x_1 - \frac{(p_2+p_1)}{2}t\right)^2\right), \tag{1.146}$$

on which one can take the limit as $\epsilon \to 0$ and corresponds to a Hamiltonian of the form

$$H_\epsilon(v,q) = \frac{\epsilon^2}{2}|v|^2 + \frac{1}{2}|q|^2 + v\cdot q. \tag{1.147}$$

In the general case for Dudley's process no perturbations with controlled Hamiltonian are yet known but it is expected that the correct scaled diffusion will converge w.r.t. large deviation principles (as in (1.146)) to geodesic flow (as (1.147)). This has been claimed to be proven in a general case in [Bismut] which we discuss futher in section 1.5.2.

Use of Varadhan's Lemma in entropic regularizations

The following proposition, based in the work [Leonard2014] explains how one can infer limiting optimal transport problems form the Schrödinger problem if the set of measures satisfy a large deviation principle. Denote by ι_A the convex indicator of A,

$$\iota_A = \begin{cases} 0 & \text{if } x \in A \\ \infty & \text{otherwise} \end{cases} \tag{1.148}$$

For the next theorem we say that $f : X \to [0,\infty]$ is coercive if $\{f \le a\}$ is compact for every $a > \inf f$. The following general result is due to Léonard [Leonard2012].

Theorem 146. *(Convergence of entropic minimizers to (positive) cost minimizers in polish space X)*
Let (X,d) be a Polish space and furnish $\Omega = C([0,1],X)$ with the sup-topology (1.2).
Assume $\{R^x_\epsilon\}$ satisfies a large deviation principle with rate $1/\epsilon$ with a lower semi-continuous coercive rate function

$$C^x = C + \iota_{\{X_0=x\}} : \Omega \to [0,\infty] \tag{1.149}$$

then as $k \to \infty$, there exists a sequence of measures μ_1^k weakly converging to μ_1 such that

$$\lim_{\epsilon\to 0} \min_{\Pi\in\Gamma(\mu_0,\mu_1^k)} \epsilon \operatorname{Ent}(\Pi \mid R) \to \min_{\Pi\in\Gamma(\mu_0,\mu_1)} \int_\Omega C(\gamma) d\Pi(\gamma) \tag{1.150}$$

And further, any limit point of the sequence of dynamical solutions $\{\Pi_k^*\}$ is a solution of the dynamical optimal transport problem with cost C (i.e. right hand-side of (1.150)).

For a proof see [Leonard2012, Theorem 3.6]. To get back to our setting notice that the underlying topology on (M, d, ℓ) is Hausdorff and the proof of Proposition 44 shows that the causality of R is inherited by every $\Pi \in \Gamma(\mu_0, \mu_1)$ so it costs nothing to add the causality restriction (as every probability with finite entropy will satisfy it). Hence, Theorem 146 will hold as soon as we identify the correct cost C and check it's coerciveness (for every C^x). We first look at a slightly different cost version of Proposition adapted to our constructions from sections 1.1-1.3.

Lemma 147. *$(x, y$-Coerciveness)*
Let (M, d, l) be a ghcrlls then for every $x \ll y$ and every $a \in \mathbb{R}$ $\{\sigma \in \Omega : \mathbf{I}^{x,y}(\sigma) \leq a\}$ is d^∞-compact.

Proof. Because of our sign convention, $\mathbf{I}^{x,y} \geq 0$ for every curve as $L_\ell(\gamma) \in [0, \ell(\sigma(0), \sigma(1))]$. Further as $\ell(x, y)$ is a constant, to show coerciveness it is enough to show $\{\gamma \in \Omega_{x,y} : -L_\ell(\gamma) \leq c\}$ is compact. By global hyperbolicity (see [Braun, Lemma B.4]) it is enough to show the set is closed, but closedness follows from L_ℓ being upper-semi-continuous with respect to d^∞ as shown in [Kunziger-Saemann, Proposition 3.17]. ∎

Intuitively, if $R_\epsilon^{x,y}(A) \to \delta_{\sigma(x,y)}(A)$ as $\epsilon \to 0$ for every A, by dominated convergence (as $R_\epsilon^{x,y}(A) \leq 1$),

$$P^\epsilon(\cdot) = \int R_\epsilon^{x,y}(\cdot) d\pi(x, y) \xrightarrow{\epsilon \to 0} \int \delta_{\sigma_{x,y}}(\cdot) d\pi(x, y) \tag{1.151}$$

where $\sigma_{x,y}$ is a geodesic from x to y. Equation (1.151) tell us that for fixed π the π-mixture of the R_ϵ bridges converges to the π-mixture of ℓ-maximizing curves. As we know from the relation between (RDSch) and (RSch), solutions to the static problem with the R_ϵ reference measure, depend on ϵ and so the idea of above needs to be rigorously proved.

Lemma 148. *Assume that $R \in \mathcal{P}(\Omega)$ and that $\{R_\epsilon^{x,y}\}$ satisfies LDP with rate ϵ and rate function $I^{x,y}$ satisfying*

$$I^{x,y}(\gamma) = 0 \Leftrightarrow L_\ell(\gamma) = \ell(x, y) \tag{1.152}$$

further assume uniqueness (up to reparametrization) of L_ℓ-maximizing curves then

$$R_\epsilon^{x,y} \rightharpoonup \delta_{\sigma(x,y)}$$

where $\sigma_{x,y}$ is a ℓ-maximizing.

Proof. By Portmanteau's theorem, it is *enough to show that* for every open set A w.r.t. $(\Omega_{x,y}, d^\infty)$, we have

$$\liminf_{\epsilon \to 0} R_\epsilon^{x,y}(A) \geq \delta_{\sigma(x,y)}(A). \tag{1.153}$$

By definition of the LDP satisfied by $R_\epsilon^{x,y}$, for every A open and every F closed in $(\Omega_{x,y}, d^\infty)$ we have

$$\begin{aligned} \liminf_{\epsilon \to 0} \epsilon \log(R_\epsilon^{x,y}(A)) &\geq -\inf_{\gamma \in A} I^{x,y}(\gamma) \\ \limsup_{\epsilon \to 0} \epsilon \log(R_\epsilon^{x,y}(F)) &\leq -\inf_{\gamma \in F} I^{x,y}(\gamma) \end{aligned} \tag{1.154}$$

Fix an open set A, by definition of the limit inferior, there exists ϵ_0 s.t. for every $\epsilon < \epsilon_0$ we have

$$\epsilon \log(R_\epsilon^{x,y}(A)) \geq -\inf_{\gamma \in A} I^{x,y}(A),$$

from which taking exponentials we get that for every $\epsilon < \epsilon_0$,

$$R_\epsilon^{x,y}(A) \geq \exp\left(-\frac{1}{\epsilon}\inf_{\gamma\in A} I^{x,y}(\gamma)\right). \tag{1.155}$$

We argue by cases: if $\sigma_{x,y} \in A$ then by the minimizing property of $\sigma_{x,y}$ and (1.152)

$$\inf_{\gamma\in A} I^{x,y}(\gamma) = I(\sigma_{x,y}) = 0$$

Plugging this value in (1.155) we obtain

$$R_\epsilon^{x,y}(A) \geq 1 = \delta_{\sigma(x,y)}(A)$$

where the last equality holds becausae we are in the case $\sigma_{x,y} \in A$, yielding in this case (1.153).
Now if $\sigma_{x,y} \notin A$,
Assume that $0 < C_A < \infty$ where

$$C_A = \inf_{\gamma\in A} I^{x,y}(\gamma).$$

By (1.155),

$$\liminf_{\epsilon\to 0} R_\epsilon^{x,y}(A) \geq \liminf_{\epsilon\to 0} \exp\left(-\frac{1}{\epsilon}C_A\right) = 0 = \delta_{\sigma(x,y)}(A),$$

where the last equality holds as we are in the case $\sigma_{x,y} \notin A$.
The case where $C_A = 0$, then the bound (1.153) is trivial:

$$\liminf_{\epsilon\to 0} R_\epsilon^{x,y}(A) \geq 1 \geq 0 = \delta_{\sigma(x,y)}(A).$$

The case $C_A = \infty$ is impossible as A is open and curves in $\Omega_{x,y}$ are ℓ-rectifiable so $\gamma \in A$ implies $I(\gamma) < \infty$. Therefore (1.153) holds in both cases ($\sigma_{x,y} \in A, \sigma_{x,y} \notin A$) concluding the proof. ∎

Theorem 149. *(Approximation via dynamic entropic regularizations)*
Let (M, d, ℓ) a ghcrlls (as in Definition 20). Set $\Omega := C([0,1], M)$ endowed with the supremum norm d_∞ from (1.2). If $x \ll y$ denote by $\Omega_{x,y}$ the elements in Ω which are causal and have start-point at x and endpoint at y. Let $\{R_\epsilon^{x,y}\}_{\epsilon>0} \in \mathcal{P}(\Omega_{x,y})$ be constructed according to construction 2 and satisfy a LDP at rate ϵ and good rate function given by $\mathbf{I}^{x,y}$ as in (1.137).
Let $\mathbb{Q} \in \mathcal{P}(\Omega)$, set $\pi = (e_0, e_1)\#\mathbb{Q}$ be such that $\pi(M^2_{\ll}) = 1$ and

$$\int \ell(x,y)d\pi \in (-\infty, \infty).$$

Define $C : \Omega \to [0,\infty]$ via

$$C(\gamma) = \begin{cases} \ell(\gamma(0),\gamma(1)) - L_\ell(\gamma) & \text{if } \gamma \in \mathcal{C} \\ \infty & \text{otherwise.} \end{cases} \tag{1.156}$$

then there exists a sequence $\{\mathbb{Q}_\epsilon\} \in \mathcal{P}(\Omega), \mathbb{Q}_\epsilon \rightharpoonup \mathbb{Q}$ w.r.t. weak convergence for measures on Ω such that

$$\lim_{\epsilon\to 0} \epsilon \operatorname{Ent}(\mathbb{Q}_\epsilon | R_\epsilon^\pi) = \int \ell(x,y)d\pi(x,y) - \int_\Omega L_\ell d\mathbb{Q}. \tag{1.157}$$

where R_ϵ^π is the π-mixture of the bridges of $R_\epsilon^{x,y}$.
Furthermore, for any sequence $\{\tilde{\mathbb{Q}}_\epsilon\}$ with $\tilde{\mathbb{Q}}_\epsilon \rightharpoonup \mathbb{Q}$ we have

$$\liminf_{n\to\infty} \mathrm{Ent}(\tilde{\mathbb{Q}}_\epsilon \mid R_\epsilon^{x,y}) \geq \int C d\mathbb{Q}$$

Remark 150. *Theorem 149 is not completely satisfactory as generalization of [Leonard2012, Theorem 3.7]. In the original work the author introduces an artificial time asymmetry that facilitates the proof by considering R^{μ_0} instead of R^π and so fixing only one of the projections of R. This trick allows the author to find the solution to the static problem with reference R by evaluating a Schrodinger problem with respect to R^{μ_0}. Namely, the x-projection is enough to obtain coercive properties of C^x,*

$$C^x = C + \infty \cdot \mathbf{1}_{\{e_0(\gamma)\neq x\}}. \tag{1.158}$$

The structure of causality theory (global hyperbolicity) impedes this coercivity because of the lack of compactness of points proper-time close to x. Global hyperbolicity yields compactness of $J^+(x)\cap J^-(y)$ so coercivity (necessary for the rate function to be good and so to have a full Varadahan-Laplace principle) is available for a cost involving past and future points ($C^{x,y}$) and not only x (via C^x) which explains why Theorem 149 depends on the M^2 distributions rather that only on one of the marginals (in contrast with [Leonard2012, Theorem 3.7]).

Proof. We follow *closely* the proof of [Leonard2012, Proposition 3.4] but we have to *modify it*. We need to modify the strategy to use the fact that $J^+(x) \cap J^-(y)$ is compact and so $C^{x,y}$ (defined below) is coercive but the one-sided cost C^x is not. A reader familiar with the proof of [Leonard2014, Theorem 3.7] may be able to skip most of this proof. Denote by $C_b(\Omega)$ the real-valued bounded continuous functions on Ω endowed with the norm

$$||f|| = \sup_{\gamma\in\Omega} |f|. \tag{1.159}$$

By $(C_b(\Omega))'$ we denote the topological dual of $C_b(\Omega)$ and we use $\langle\cdot,\cdot\rangle$ to describe the dual pairing. Let $\pi \in \mathcal{P}(M^2)$ such that $\pi(M^2_\leq) = 1$ and assume that $P \in \mathcal{P}(\Omega)$ satisfies $(e_0, e_1)\#P = \pi$ and $P^{x,y} \in \mathcal{P}(\Omega_{x,y})$ whenever $x \leq y$. For any $R \in \mathcal{P}(\Omega)$, define the π-mixture of it's bridges via

$$R^\pi(\cdot) = \int_{M^2} R^{x,y}(\cdot) d\pi(x, y). \tag{1.160}$$

We will first show that for every $Q \in (C_b(\Omega))'$

$$\begin{aligned} \mathrm{Ent}(Q \mid R^\pi) + \iota_{Q\in\mathcal{P}(\Omega):(e_0,e_1)\#Q=\pi} \\ = \sup_{f\in C_b(\Omega)} \left\{ \int f dQ - \int_{M^2} \log\left(\langle e^f, R^{x,y}\rangle\right) d\pi(x, y) \right\}. \end{aligned} \tag{1.161}$$

where ι_A denotes the convex indicator from Definition (1.148). We show (1.161) by considering $\Theta : C_b(\Omega) \to \mathbb{R}$ given by

$$\Theta(f) = \int_{M^2} \log\left(\langle e^f, R^{x,y}\rangle\right) d\pi(x, y) \tag{1.162}$$

and computing it's Fenchel-transform (with respect to the $\langle\cdot,\cdot\rangle$ pairing). For $Q \in (C_b(\Omega))'$

$$\begin{aligned}\Theta^*(Q) &= \sup_{f\in C_b(\Omega)} \{\langle f, Q\rangle - \Theta(f)\} \\ &= \sup_{f\in C_b(\Omega)} \left\{\langle f, Q\rangle - \int_{M^2} \log\left(\langle e^f, R^{x,y}\rangle\right) d\pi(x,y)\right\}.\end{aligned} \tag{1.163}$$

Let us show that $\{Q \in (C_b(\Omega))' : \Theta^*(Q) < \infty\} \subseteq \{Q \in M_b^+(\Omega) : (e_0, e_1)\#Q = \pi\}$.
Let $Q \in (C_b(\Omega))'$ s.t. $\Theta^*(Q) < \infty$ following [Leonard2012, Lemma 5.2] to show $Q \in M_b(\Omega)$ it is necessary and sufficient to show that for every decreasing sequence $f_n \downarrow 0$ then $\lim_{n\to\infty}\langle f_n, Q\rangle = 0$.
Still following the approach of [Leonard2014, Lemma 5.2], to show $Q \geq 0$ assume $f \geq 0$, if $a \leq 0$ then $\Theta(af) \leq 0$ from which

$$\begin{aligned}\Theta^*(Q) &\geq \sup_{a\leq 0}\{a\langle f, Q\rangle - \Theta(af)\} \\ &\geq \sup_{a\leq 0}\{a\langle f, Q\rangle\} = \iota_{\langle f,Q\rangle<0}.\end{aligned}$$

Hence $\Theta^*(Q) < \infty$ implies $\langle f, Q\rangle \geq 0$ for every $f \geq 0$ yielding $Q \geq 0$. If (f_n) is a sequence of positive functions decreasing to 0 by dominated convergence for every $a \geq 0$,

$$\lim_{n\to\infty}\Theta(af_n) = \lim_{n\to\infty}\int_{M^2}\log\left(\langle e^{af_n}, R^{x,y}\rangle\right) = 0.$$

By definition of Θ^* as a supremum,

$$\begin{aligned}\Theta^*(Q) &\geq \sup_{a\geq 0}\limsup_{n\to\infty}\{a\langle f_n, Q\rangle - \Theta(af_n)\} \\ &\geq \sup_{a\geq 0}\limsup_{n\to\infty}\langle f_n, Q\rangle\end{aligned}$$

Which again shows that if $\Theta^*(Q) < \infty$ then $\limsup_{n\to\infty}\langle f_n, Q\rangle = 0$.
Now we know that $\Theta^*(Q) < \infty \Rightarrow Q \in M_b^+(\Omega)$. We continue to show that $(e_0, e_1)\#Q = \pi$, let for $f \in C_b(\Omega)$, and $\phi \in C_b(M^2)$ via $\phi = f \circ (e_0, e_1)$ as composition of continuous functions as (e_0, e_1) is continuous in d^∞ therefore

$$\begin{aligned}&\sup_{\phi\in C_b(M^2)}\left\{\int\phi dQ - \int_{M^2}\log\left(\langle e^f, R^{x,y}\rangle\right)d\pi(x,y)\right\} \\ &= \sup_{\phi\in C_b(M^2)}\int_{M^2}\left\{\int_{M^2}\phi d(e_0,e_1)\#Q - \int_{M^2}\log\left(\int_\Omega e^{\phi(x,y)}dR^{x,y}(\gamma)\right)d\pi(x,y)\right\} \\ &= \sup_{\phi\in C_b(M^2)}\left\{\int_{M^2}\phi(x,y)d((e_0,e_1)\#Q - \pi)(x,y)\right\} \\ &= \iota_{(e_0,e_1)\#Q\neq\pi},\end{aligned}$$

from which we obtain that $\Theta^*(Q) < \infty$ yields $(e_0, e_1)\#Q = \pi$ as desired. With this result we desintegrate Q,

$$Q(\cdot) = \int_{M^2} Q^{x,y}d\pi(x,y). \tag{1.164}$$

By definition of Θ^*,

$$\begin{aligned}\Theta^*(Q) &= \sup_{f\in C_b(\Omega)} \left\{ \int_{M^2} \langle f, Q^{x,y}\rangle - \log\left(\langle e^f, R^{x,y}\rangle\right) d\pi(x,y) \right\} \\ &\leq \int_{M^2} \sup_{f\in C_b(\Omega)} \left\{ \langle f, Q^{x,y}\rangle - \log\langle e^f, R^{x,y}\right\} d\pi(x,y) \\ &= \int \mathrm{Ent}(Q^{x,y} \mid R^{x,y}) d\pi(x,y) \\ &= \mathrm{Ent}(Q \mid R^\pi),\end{aligned}$$

where the first equation is due to Fatou's lemma, the second one to Gibb's principle together with (133) and the last one as $\pi = (e_0, e_1)\#Q = (e_0, e_1)\#R^\pi$. To obtain the reverse inequality, note by Jensen's inequality,

$$\int \log\langle e^f, R^{x,y}\rangle d\pi \leq \log \int \langle e^f, R^{x,y}\rangle d\pi.$$

which gives $\Theta^*(Q) \geq \mathrm{Ent}(Q \mid R^\pi)$.
Towards the proof of Theorem 149, we define

$$C^{x,y}(\gamma) = \begin{cases} \ell(x,y) - L_\ell(\gamma) & \text{if } \gamma \in \Omega_{x,y} \text{ is causal} \\ \infty & \text{otherwise.} \end{cases} \tag{1.165}$$

We also define the associated Λ-operator which we expect as Γ-limit:

$$\begin{aligned}\Lambda(f) :&= \int_{M^2} \sup_{\gamma\in\Omega} \{f(\gamma) - C^{x,y}(\gamma)\} d\pi(x,y) \\ &= \int_{M^2} \sup_{\gamma\in\Omega_{x,y}} \{f(\gamma) - C^{x,y}(\gamma)\} d\pi(x,y)\end{aligned}$$

where the equality holds because if γ is not causal or $\gamma(0) \neq x$ or $\gamma(1) \neq y$ then $C^{x,y}(\gamma) = \infty$. We now show that $\{\Lambda^* < \infty\} \subseteq M_b^+(\Omega)$. Consider as before $f_n \downarrow 0$ and by in [Leonard2012, Lemma 5.3] $(\sup_\Omega\{f_n - C^{x,y}\})_n$ is decreasing and

$$\lim_{n\to\infty} \sup_\Omega\{f_n - C^{x,y}\} = 0.$$

Observe that

$$|\sup_\Omega\{f_n - C^{x,y}\}| \leq \sup_\Omega |f| \tag{1.166}$$

The bound (1.166) allows us to use dominated convergence and so $\Lambda(af_n) \to 0$. Because $C^{x,y}$ is coercive and $[0,\infty]$-valued, $\sup_\Omega\{f_n - C^{x,y}\}$ is decreasing by [Leonard2012, Lemma 5.3] and $J = C^{x,y}$ satisfies

$$\lim_{n\to\infty} \sup_\Omega\{f_n - J\} = \sup_\Omega\{f - J\}. \tag{1.167}$$

We set

$$\Lambda_\epsilon(f) := \epsilon \int_{M^2} \log\langle e^{1/\epsilon f}, R_\epsilon^{x,y}\rangle d\pi(x,y) \tag{1.168}$$

By [Leonard2012, Corollary 6.4] to show Γ-convergence of Λ_ϵ^* to Λ^* it is enough to show:

1. $\lim_{\epsilon\to 0}\Lambda_\epsilon(f) = \Lambda(f)$ for every $f \in C_b(\Omega)$
2. $\sup_{0<\epsilon<1} |\Lambda_\epsilon(f)| \leq ||f||, |\Lambda(f)| \leq ||f||$.
3. $\Lambda_\epsilon, \Lambda$ are convex.

Because $C^{x,y}$ is coercive, $\mathbf{I}^{x,y}$ is a good rate function so Varadahn's lemma (1.127) applies for $f \in C_b(\omega)$ and we have by dominated convergence

$$\lim_{\epsilon\to 0}\Lambda_\epsilon(f) = \Lambda(f), \tag{1.169}$$

proving the first item. The second item follows from the point-wise bound of f and $R_\epsilon^{x,y}$ being a probability measure and convexity is a consequence of Hölder's inequality. Again using [Leonard2012, Corollary 6.4] their convex conjugates Γ-converge. But we have computed the convex conjugates and the Γ-convergence is exactly the statements on Theorem 149 finalizing the proof if we note that

$$\begin{aligned}\Lambda^*(Q) &= \sup_{f\in C_b(\Omega)}\left\{\int_{M^2}\langle f, Q^{x,y}\rangle - \sup_{\gamma\in\Omega_{x,y}}\{f(\gamma) - C^{x,y}\}d\pi(x,y)\right\}\\ &= \int \ell(x,y)d\pi(x,y) - \int_\Omega L_\ell(\gamma)dQ(\gamma)\end{aligned}$$

where the first equality holds because of (1.165) in the finite case as $(e_0, e_1)\#Q = \pi$ and both sides are ∞ in any other case. The last equality is due to [Leonard2012, Lemma 5.5] as L_ℓ is upper semi-continuous making C in (1.156) lower semi-continuous as a sum of lower semi-continuous functions satisfying that for every $x \ll y$ we have

$$\inf_{\gamma\in\Omega} C^{x,y}(\Omega) = \inf_{\gamma\in\Omega_{x,y}}\{\ell(x,y) - L_\ell(\gamma)\} = 0$$

due to the space being path-connected and Lemma 14 which is the condition π-a.e. of the hypothesis of [Leonard2012, Lemma 5.5] because $\pi(M^2_\ll) = 1$. ∎

Corollary 151. *(An approximation result of entropy maximizers)*
Let $\mu_0, \mu_1 \in \mathcal{P}(M)$ such that

$$C_\ell(\mu_0,\mu_1) := \sup_{\pi\in\Gamma_\leq(\mu_0,\mu_1)}\int_{M^2_\leq}\ell(x,y)d\pi(x,y) < \infty. \tag{1.170}$$

Assume that for every sequence $\{\pi_n\} \in \mathcal{P}(M^2_\ll)$ there is a sequence of probability measures $\{\mathbb{Q}_n\}_n \in \mathcal{P}(\Omega)$ with $(e_0, e_1)\#\mathbb{Q}_n = \pi_n$ such that

$$\lim_{n\to\infty}\int L_\ell(\sigma)d\mathbb{Q}_n(\sigma) = 0. \tag{H}$$

Then there exists a (double sequence) $\{{\mathbb{Q}_\epsilon}^n\}_{n,\epsilon} \in \mathcal{P}(\Omega)$ such that

$$\limsup_{n\to\infty}\left\{\lim_{\epsilon\to 0^+}\epsilon\,\mathrm{Ent}\left({\mathbb{Q}_\epsilon}^n \mid R_\epsilon^\pi\right)\right\} = C_\ell(\mu_0,\mu_1). \tag{1.171}$$

Proof. For every fixed $\pi \in \Gamma_{\leq}(\mu_0, \mu_1)$ by Theorem 149 for every fixed $Q \in \mathcal{P}(\Omega)$ there exists a sequence Q_ϵ with $Q_\epsilon \rightharpoonup Q$ as $\epsilon \to 0$,

$$\lim_{\epsilon \to 0} \epsilon \operatorname{Ent}(Q_\epsilon \mid R_\epsilon^\pi) = \int \ell(x,y) d\pi - \int L_\ell(\gamma) dQ(\gamma).$$

Assume π_n is a maximizing sequence in 1.170, let $\mathbb{Q}_\epsilon^n$ be the sequence associated to $\mathbb{Q}^n$ (of hypothesis H) from Theorem 149, then

$$\lim_{\epsilon \to 0} \operatorname{Ent}(\mathbb{Q}_\epsilon^n \mid R_\epsilon^{\pi_n}) = \int \ell(x,y) d\pi_n - \int L_\ell(\gamma) d\mathbb{Q}_n(\gamma).$$

Considering the limit superior on both sides yields (1.171). ∎

As a last observation note that by joint lower-semicontinuity of entropy, we know that if $\mathbb{Q}_\epsilon \rightharpoonup \mathbb{Q}$ then

$$\liminf_{\epsilon \to 0^+} \operatorname{Ent}(\mathbb{Q}_\epsilon \mid R_\epsilon^\pi) \geq \operatorname{Ent}(\mathbb{Q} \mid \delta_\sigma^\pi),$$

which together with Corollary 151 yields

$$\int_\Omega C d\mathbb{Q} \geq \operatorname{Ent}(\mathbb{Q} \mid \delta_\sigma^\pi).$$

Open Question 1. *Is it possible to improve Theorem 149 to obtain the actual limit as solution of a true Schrödinger problems with respect to varying target measures μ_1^k as in [Leonard2012]? A natural approach would be to enforce compactness by considering a sequence of compact rectangles $K_n^1 \times K_n^2$ approximating in some way $\operatorname{spt}(\mu_0 \times \mu_1)$ and not only $\Omega_{x,y}$ but it is not clear if this procedure yields any convergence properties.*

Observe that in the context of section 1.3.6 one needs to ensure that an analogue of Theorem 1.150 holds which includes (in it's proof) the use of the duality of $(C_b(\Omega), d^\infty)$ with the space of measures which is unavailable in the general case.
In the case of $\mathbb{M}^1$ for RKSchP we can apply directly Theorem 1.150 and obtain a solution to the optimal transport associated to the rate function of section 1.4.3 which corresponds to geodesic flow according to 1.147.
In the case of the kinetic Schrödinger problem in $\mathbb{M}^n$ associated to Dudley's process we do not know the rate function (if any) to try to apply Theorem 1.150 unless the results of Dr. Bismut can be adapted our framework (which we discuss in 1.5.2).

1.5 Conclusions and further work

1.5.1 Conclusions

We have generalized the Schrödinger problem to adapt to recent investigations of the theory of optimal transportation on spaces endowed with causality features. This work adapts the problem to physically relevant settings were we have studied existence, uniqueness and other properties. The absence of a canonical heat semigroup in the Lorentzian and pre-Lorentzian case was resolved through the construction of bridge measures which lacked enough properties to be considered a satisfactory generalization of Brownian Bridges. In the known case of Minkowski space, the analogue

of Brownian bridges was the conditioning of Dudley's process whose intrinsic existence in phase-space was addressed by formulating the relativistic kinetic Schrodinger problem. Although there seems to be no straight-forward generalization of the Brownian Bridge, we dealt separately with each of it's fundamental features: Markovianity and it's large deviation principle. This separation made the theory of each feature interesting by itself. In the last chapter we showed a strategy to almost recover the Lorentzian costs in the entropic limit, through the theory of Large deviation for different bridge measures. The **main impediment** to applying the theory of [Leonard2014] in our context is not the non-compactness of sub-level sets of ℓ but rather the interplay of the signs and suprema/infima; one can't seem to apply directly the theory of Γ-convergence as one is dealing with maximizers but if one flips signs to deal with minimizers instead, the now non-positive cost is not ensured to be the limit of Λ_n.

1.5.2 Future work and open problems

A physical digression and a second approach

We have studied the Schrödinger problem from the perspective of the abstract minimization of entropy for space-time measures (RSch and RDSch). In the Euclidean version of the problem, the Schrödinger potentials (ϕ, ψ) are intrinsically related to the Schrödinger equation, for example see [Tamanini, Section 5.5] where it is shown that a transformation of Schrödinger potential solves

$$i\frac{\partial \psi_t}{\partial t} + \frac{1}{2}\Delta\psi_t - \frac{\Delta\sqrt{\rho_t}}{\rho_t}\psi_t = 0. \tag{1.172}$$

This corresponds in physics literature to the study of the Schrödinger equation with Bohm's potential. This indicates that an alternative approach to the study of the Schrödinger equation in space-times would be working backwards, i.e. to define first the analogue of (1.172) and aim to recover connections with RSch and RDSch.

Open Question 2. *Although some work [Mauri-Giona] exists on the space-time version of Bohm's potential and the Schrödinger equation is classical, it is not yet clear to the author if connections described in [Tamanini, Section 5.5] still hold (even) in the Minkowskian case.*

Bismut Laplacian, Molchanov's work non-Markovian Semi-groups

The seminal work of [Dudley1966] and subsequent work of [Franchi-LeJan2007], [Dunkel-Hanggi], [Chevalier-Debbasch] (and others see references therein) study the diffusion process generated by a hypo-elliptic operator (see section 1.3.3). In the last years, Bismut has studied a general version of this operator (see [Bismut]). The so-called Bismut's hypo-elliptic Laplacian is an interpolation between geodesic flow and the Laplacian by a noise parameter. This parameter is exactly σ (or $1/\sigma$ in our notation) on 1.3.3. It seems that the theory of Bismut's hypo-elliptic Laplacian shows the probabilistic convergence of the operators on general dimension n and not only in $n = 1, 2$ cases known to Kolmogorov, Matsumoto and Ikeda (see section 1.4.3). It is unknown to the author whether this general framework indeed shows the Large Deviation Principle as required in section 1.4 for the general case in $\mathbb{M}^n$ for the Kinetic Schrödinger problem or (maybe) even in the case of curved geometries (section 1.3.4).

Open Question 3. *Whether one can take the results from [Bismut] and apply them directly to the kinetic Schrödinger problem is a very promising line of investigation.*

The Carnot-Caratheodory distance and the time-like curvature dimension condition

In the study of Theorem 86 we encountered the generalized Markov semigroup for Kolmogorov operator. By [Baudoin-Gordina-Mariano] this semigroup is related to the Carnot-Caratheodory distance on sub-riemannian manifolds. This idea hints towards studying Dirichlet-like energies on phase-space, whether or not this is a reasonable approach to improve on the theory of relativistic kinetic Schrödinger problems is an interesting problem. Observe that the generalized Bakry-Emery condition satisfied by the Kolmogorov operator in $\mathbb{M}^n$ is an alternative way of describing an intrinsically causal curvature dimension condition. Although it's apparent that it is not to be compared to the $CDT^e(K,N)$ condition of [McCann2019], [Cavalletti-Mondino] [McCann2023] as in the Riemannian case, it would be extremely relevant to relate both notions or to that of [Rupert-Woolgar].

Limit of theory: Girsanov's theorem

In the study of the Euclidean Schrödinger problem, it is often common to use a version of Girsanov's theorem (see [Leonard2014], [Chiarini-Conforti-Greco]). The idea is that via Girsanov's theorem, we can write entropy in a simple way using

1. Doob's decomposition of semi-martingales
2. Riesz-representation theorem.

Although the second one is readily available on Hausdorff spaces the first one is not. The goal of using Girsanov's theorem is that the exponential martingale cancels out:

$$\begin{aligned}\int \log\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{P} &= \int_0^1 \int \log\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\Big|_{\mathcal{F}_t}\right) d\mathcal{P}_t dt = \int_0^1 \int (Z_t + \frac{1}{2}[Z]_t) d\mathcal{P}_t dt \\ &= \int_0^1 \int \frac{1}{2}[Z]_t d\mathcal{P}_t dt\end{aligned}$$

which corresponds to the functional to minimize/maximize in the analogue of Benamou-Brenier's formula (1.4) for entropy. Girsanov's theorem requires an underlying Hilbertian structure so it is likely that the proof of Léonard can be replicated in an abstract Wiener space but not in the general Lorentzian case.

Modifying the bridge constructions

The constructions 1, 2 and their generalizations in section 1.3.5 lack a Markov property as shown in Proposition 78. Nevertheless, we expect that one can modify the constructions by further randomizations on conditional steps which would preserve control on large deviation principles. This idea is intuitive and exciting although we expect the construction to be much more elaborate than the the ones presented in this document (in terms of showing existence and studying kernels).

Topology-Markov and operators (hypo-ellipticity)

Section 1.3.6 is a mathematical approach to avoid the non-physicality of the Markov property associated to measures on curves with external parameter. The theory of Markov processes is vast in literature and it's connection to elliptic operators and semigroups has been widely established.

Open Question 4. *The question whether the condition of 1.111 relates to semigroup theory remains open.*

Occupation measure and the result of G. Arous and G. Guionnett

A well-known general principle to find small time asymptotics and large deviation principles arising from operators in Hörmander form was further studied by Arous and Guionnett in [Arous-Guionnet]. We briefly explain the technique as the approach to obtain measures in ghcrlls spaces is clearly not unique, a generalization of the hypo-ellipticity technique now described would also be natural. In the context when M is a m-dimensional differentiable manifold, assume $\mathbb{P}_x$ is the law of a process solving the martingale problem started at $x \in M$, for an operator L in Hörmander form

$$L = X_0 - \sum_{k=1}^{d} X_k^* \circ X_k$$

and satisfy the strong condition of Hörmander, i.e. for every $x \in M$,

$$\text{Lie}(X_1(x), \dots, X_n(x)) = T_x M \tag{1.173}$$

Let L be a differential operator satisfying the strong Hörmander condition (1.173), the empirical process L_t is the random variable that assigns to $\sigma \in C([0,\infty), M)$

$$L_t(\sigma) = \frac{1}{t} \int_0^t \delta_{e_s(\sigma)} ds \tag{1.174}$$

according to the measure associated to the diffusion process of L. In [Arous-Guionnet], it was shown that the laws of L_t satisfies a Large deviation principle as $t \to \infty$ with rate function:

$$J(\nu) = \sup_{u \in C_+^\infty(M,\mathbb{R})} \left\{ - \int \frac{Lu}{u} d\nu \right\}. \tag{1.175}$$

The work of [Arous-Guionnet] showed that an alternative representation of the rate function is

$$J(\nu) = \frac{1}{4} \sup_{\substack{\varphi \in C^\infty(M,\mathbb{R}) \\ |X(\varphi)|_{L^2(\nu)} \neq 0}} \left\{ \frac{(\int L\varphi d\nu)^2}{|X(\varphi)|^2_{L^2(\nu)}} \right\}$$

This idea gives us a direct way to form non-trivial and physically relevant measures satisfying a large deviation principles in spacetime. Observe that the construction of the empirical process depends only on differentiable manifolds and not Riemannian ones!. This observation makes the technique relevant towards the study of Large deviation principles on spacetime. Arous's construction is intimitely related to the metric space version of Sanov's theorem (see [Leonard2012]). We can focus on trying to replicate Hörmander's condition 1.173 in a more abstract framework.

Cadlag

Most of the work in chapter 1.3 was made on $C^\tau_{x,y}$ but many topologies don't make $C^\tau_{x,y}$ closed but make $\text{Cad}^\tau_{x,y}$ closed (as pointwise convergence of continuous functions may not be continuous but cad-lag limits are cad-lag). The development of more general bridge measures in abstract spacetimes through cad-lag processes is a natural next step towards understanding Markov-like processes with fixed underlying topology on causal curves. We expect (1.76) to be easy to verify in many cases.

Notation

ℓ - time separation function
d - distance (Polish topology in most cases)
$\ll, \leq$ - chronological and causal relations
L_ℓ - The ℓ-length (definition 5).
d^∞ - Supremum norm (1.2).
$\mathrm{Ent}(\mu|\nu)$ - Relative Entropy of μ with respect to ν 26.
$\mathcal{B}$ - Borel σ-algebra.
$\mathcal{B}^\tau$- Borel σ-algebra for the topology τ.
$\Gamma_\leq(\mu,\nu)$ - Causal couplings of μ_0 and μ_1 (1.9)
$M^+(X)$ - Positive Borel measures on X.
$M_b^+(X)$ - Positive bounded (finite) Borel measures on X.
$\Omega_{x,y}$ - Causal continuous bridge space from x to y endowed with the supremum topology.
$\Omega_{x,y}^D$ - Causal cad-lag bridge space from x to y endowed with the Skorohod topology.
e_t - Evaluation map $e_t(\omega) = \omega(t)$.
δ_x - Dirac delta at x.
$C(X,Y)$ - Continuous functions from X to Y
$\mathcal{C}(X,Y)$ - Continuous and causal functions from X to Y.
$\mathrm{MID}(x,y)$ - Mid-set 53.
$c-\mathrm{MID}(x,y)$ - Mid-set contracted by c 57
$||\cdot||_2$ - Euclidean norm.
nP_k - Notation for the inductive construction 1.21
$\mathcal{B}_t^s$ - Dudley's notation for $s-t$ Borel σ-algebra in $\mathbb{R}^3$
$\mathbb{M}^n$ - $n+1$-dimensional Minkowski space ($\mathbb{R}^{1,n}$)
$\mathbb{H}^n$ - Hyperboloid of dimension n.
$\mathcal{U}$ - Upper Sheet of the hyperboloid (dimension in the context)
$\mathcal{H}$ - Hausdorff measure.
$\mathcal{U}(S)$ - Uniform probability measure on the set S(normalization of the Hausdorff measure).
Δ - Laplace-Beltrami operator.
P_t Markov semigroup semigroup in a full Markov-Triple (see [J, Chapter 1])
Γ - Carré du Champ defined in 1.67
$\mathcal{T}$- Temporal function (or time function Definition 102).
$\xrightarrow{\mathcal{H}}$ - $\mathcal{H}^1$-convergence Definition1.3.6.
$\rightharpoonup$ Narrow convergence (against continuous bounded functions).
$\xrightarrow{\mathcal{H}\otimes}$ - Convergence with transference plans Definition 112.
$\mathcal{L}$ - Lorentz group.
$\mathcal{PS}$ Phase-Space defined in section 1.3.4

Chapter 2

The minimizing movement scheme for the aggregation equation on compact Riemannian manifolds

2.1 Introduction

We consider an aggregation equation on (M, g), a smooth, connected, compact Riemannian manifold without boundary, in which the evolution of the density μ_t of population (or particles) is described by

$$\partial_t \mu_t - \nabla_M \cdot ((\nabla_M (W * \mu_t))\mu_t) = 0 \tag{2.1}$$

where $W * \mu_t(x) = \int_M W(x,y) d\mu_t(y)$ is the convolution operator for a potential function $W : M \times M \to \mathbb{R}$. Further, we assume that the interaction depends only on intrinsic distance, i.e. $W(x,y) = h(d(x,y)^2)$ where $h : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and $d(x,y)$ denotes the Riemannian distance between x and y. We assume h is non-decreasing for big enough distances.
Aggregation models of the form (2.1) have been recently used in many different applications. In a wide sense, aggregation models describe the collective behaviour of groups of individual entities whose movement is determined by simple rules. These simple rules are modeled through a potential. For example, in [Mogilner-Edelstein.Keshet] they describe the swarming behaviour of biological entities for which the different characteristics of the operator W determine the shapes at front and end of the swarm. In [Ji-Egerstedt], the aggregation model is used to describe the behaviour of multiple agents exploring a region while mantaining a coordinated goal (such as policing). The euclidean version of model (2.1) has received a lot of attention in recent years, see [Bertozzi-Laurent-Rosado], [Carrillo-Figalli2011], [Bonaschi-Carrillo-DiFrancesco-Peletier] or [Carrillo-James-Lagoutiere-Vauchelet]. In [Carrillo-Figalli2011], the authors showed global in time existence of measure solutions and the possibility of finite time concentration. The technique is similar to our approach, we aim to compute the metric derivative of the interaction energy and use an optimality condition to pass some properties from the potential to the flow. The minimizing movement scheme is also considered

in [Carrillo-Figalli2011] with the main difference that the potentials uniquely determine the direction of movement everywhere. In our case, the presence of the cut locus impedes us from knowing the direction of movement. A model similar to ours has been considered in [Patacchini-Slepcev], where the authors also used a technique motivated by the minimizing movement scheme. The difference relies in using the euclidean distance of the ambient space into which the manifold is embedded. The idea of using a generalization of the projection onto the tangent space allows the authors to use the information from the Euclidean setting and allows them to conclude stability of the model.
The approach of using the euclidean distance of the ambient spaced (as in [Patacchini-Slepcev]) is referred to as an *extrinsic* model. In this document we consider a completely *intrinsic* approach: the dependence of the interaction potential on the agents position is only through the Riemannian distance on the manifold between the agents. Another approach can be found in [Fetecau-Park-Patacchini] and [Fetecau-Patacchini], there the authors solve the existence and uniqueness question for weak solutions of model (2.1) using the theory of partial differential equations and Lipschitz-coefficients theorems. This approach is very fruitful and motivated the work presented here. In contrast we aim to obtain similar conclusions by only looking at the measures involved. Although both approaches are completely intrinsic (in the sense that they depend only on Riemannian distance) the goal of this document is to adapt the theory of Wasserstein gradient flows as presented in [Ambrosio-Gigli-Savare].
In comparison with [Fetecau-Park-Patacchini] and [Fetecau-Patacchini] where the authors studied a similar problem, our notions of solutions differ slightly. Their notion of solutions is well adapted for an application of the Cauchy-Lipschitz theory, while in contrast our notion comes from a generalization of the notion of solutions from the theory of gradient flows in metric spaces from [Ambrosio-Gigli-Savare]. In [Fetecau-Patacchini] the authors assume Lipschitz regularity of the coefficients of the ODE associated to the potential W. In our assumptions **(W0)**-**(W1)** we allow slightly less regular potentials at the cost of obtaining only small time measure-valued solutions in contrast with the global result [Fetecau-Patacchini, Theorem 2.6]. Our assumption **(W2)** describes a situation on which after certain perceiving enough distance to the other agents, the individuals are compelled to get closer.
The main technical difficulty for finding solutions of (2.1) in the manifold setting is that the presence of the cut locus stops us from applying directly the theory of Wasserstein gradient flows as the interaction can not be shown to be globally λ-convex. We overcome this problem by forcing a regularizing Wasserstein term and noting that optimality conditions control the distance of transport for every timestep τ.
An introductory treatment of the Wasserstein gradient flow for the interaction energy on $\mathbb{R}^n$ can be found [Villani2003, Chapter 8].
One of the main problems of models of the form (2.1) in the manifold setting is the possible non-differentiability of potentials on the cut locus (the Riemannian distance fails to be differentiable there). The existence of the cut locus presents a significant challenge which impedes the use of the techniques from the euclidean case. In essence, when agents happen to be in the cut locus of each other, they don't have a preferred direction to move. This can be seen through the failure of differentiability of the terms involved in the optimization process. To overcome this difficulty we first prove that the speed of propagation of the minimizing movement scheme is finite (Proposition 182), which intuitively says the spread of the particles is slow enough to not instantly fall into the cut locus.
Our approach is specially interesting for applications, as it leads the way to different types of numerical algorithms using the algorithms for optimal transportation (see [Peyre-Cuturi]). This

could open a new line of investigation to compare the efficiency of algorithms derived from PDE-approximation methods against optimal transport based methods. The methods in [Benoit et al.] show the implementation of (upwind) discretizations for approximations to aggregation equations in $\mathbb{R}^p$. These methods rely on minimization schemes such as the JKO scheme from [Jordan-Kinderlehrer-Otto]. There is no analogue of such results for intrinsic equations on Riemannian manifolds. This document addresses the existence of measure valued solutions obtained by such discretizations. In this document we generalize the use of this discretizations to the manifold case. After discovering the limits of the theory in the Riemannian case (due to the presence of the cut locus) we can study the performance of such methods up to a small time determined in this document. Whether or not one can prove the same orders of convergence of the minimization schemes from [Benoit et al.] in the Riemannian case remains an open question and interesting line of investigation.
To summarize the paper, we prove that for small times, solutions of the aggregation equation (2.1) exist whenever the interaction is intrinsic and satisfies **(W0)**-**(W2)**. To this end we use the minimizing movement scheme in which the Euler-Lagrange conditions for optimality allow us to upgrade the regularity properties of Kantorovich potentials (in the support of the measures) via non-smooth analysis. This analysis shows that the minimizing movement scheme does not immediately move to the cut locus from which one can deduce several properties of the limiting measure.

2.2 Preliminaries and precise formulation of the problem

Let (M, g) be a smooth, connected, compact n-dimensional Riemannian manifold without boundary. Let $(x^1, x^2, \cdots, x^n)$ be local coordinates, we denote by T_pM the tangent space at $p \in M$ and let g_{ij} be the metric coordinates. For the canonical basis $\{\frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2}, \cdots, \frac{\partial}{\partial x^n}\}$ the gradient of a scalar function on M is given by

$$\langle \nabla_M f(p), v\rangle_p = df_p(v)$$

for every $v \in T_pM$, where df_p is the differential of f at $p \in M$ and $\langle \cdot, \cdot \rangle_p$ denotes the inner product in T_pM. Hence, by using the local coordinates,

$$\nabla_M f = g^{ij} \frac{\partial f}{\partial x^i} \frac{\partial}{\partial x^j}.$$

For a tangent vector field $F = F^i \frac{\partial}{\partial x^i}$, we define the divergence,

$$div(F) = \frac{1}{\sqrt{\det g}} \sum_{i=1}^{n} \frac{\partial}{\partial x^i} \left(\sqrt{\det g} F^i \right).$$

If $f : M \to \mathbb{R}$ is differentiable, we define the Hessian, $\operatorname{Hess} f$ of f at $p \in M$ as the linear operator $\operatorname{Hess} f : T_pM \to T_pM$ via the formula

$$\operatorname{Hess} f(Y) = \nabla_Y (\nabla_M f)$$

for $Y \in T_pM$ and ∇_Y denoting the covariant derivative along Y, see [Do Carmo]. The standard volume in local coordinates is

$$dvol = \sqrt{\det g} dx^1 \wedge dx^2 \wedge \cdots \wedge dx^n.$$

For given $p \in M$ the cut locus at p, denoted $Cut(p)$ denotes the set of points on M that can not be linked to p by any extendable geodesic. The Cut locus is the subset of $M \times M = \{(x,y) \in M \times M : y \in Cut(x)\}$. We denote by $t_{Cut(x_0)}$ the time to cut locus from x_0 and $t_{\inf}$ the minimum of these times (achieved by compactness).
For $p \in [1,\infty)$ denote by $\mathcal{P}_p(M)$ the set of probability measures with p-finite moment, and $\mathcal{P}_{ac}^p(M)$ the subset of $\mathcal{P}_p(M)$ of measures absolutely continuous with respect to *dvol*. For $\mu,\nu \in \mathcal{P}_p(M)$ we define the Wasserstein-p metric via

$$d_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{M\times M} d(x,y)^p d\pi(x,y) \right)^{1/p}, \tag{2.2}$$

where $\Pi(\mu,\nu)$ denotes the set of probability measures on $M \times M$ whose marginals are μ and ν respectively and again $d(x,y)$ denotes the Riemannian distance between x and y. Because M is assumed to be compact, $d(x,y) \leq diam(M)$ for every x,y and so $\mathcal{P}_p(M) = \mathcal{P}(M)$. Throughout this work we refer to weak convergence of measures to the convergence with respect to all continuous (hence bounded as M is compact) functions from M to $\mathbb{R}$. Recall:

Theorem 152. *(Optimal Transportation on Riemannian manifolds)*
In a smooth, connected Riemannian manifold, if μ,ν are compactly supported measures on M and μ is absolutely continuous with respect to Riemannian volume, then considering the cost function $d(x,y)^2/2$, there exists an optimal transport map T, transporting μ onto ν determined uniquely μ a.e. by

$$T(x) = \exp_x(-\nabla\phi(x)),$$

where ϕ is some $d^2/2$-concave function.

The proof can be found in [McCann, Theorem 9] and it is very standard in optimal transport literature, hence omitted. Finally, we recall that on a compact space the Wasserstein p-metrics are ordered, if $p_1 \geq p_2$ then $d_{p_2}(\mu,\nu) \leq d_{p_1}(\mu,\nu)$ whenever $\mu,\nu \in \mathcal{P}(M)$, see [Villani2003, Section 7.1.2]. We start setting up the problem by defining what we mean by a solution. We are interested in measure-valued solutions to the aggregation model.

Definition 153. *(Measure-valued solutions)*
Given $T \in (0,\infty)$ we say that $\{\mu_t\}_{t\in[0,T]}$ is a measure-valued solution to the aggregation model (2.1) with potential function $W : M \times M \to \mathbb{R}$ if for every test function, $\phi \in C_c^\infty([0,T) \times M)$ we have

$$\int_0^T \int_M \partial_t\phi(t,x) + \langle \nabla_M \phi(t,x), \nabla_M (W * \mu_t)(x) \rangle_x d\mu_t(x) dt = 0, \tag{2.3}$$

where C_c^∞ denotes smooth functions with compact support.

Remark 154. *In a more general setting one prescribes the value of solutaions at $t = 0$ to be a given measure $\nu \in \mathcal{P}(\Omega)$. One way to do this is to add to (2.3) the following term:*

$$\int_\Omega \phi(0,x) d\mu_0 - \int_\Omega \phi(0,x) d\nu(x). \tag{2.4}$$

Every construction in this work will have the same initial measure (denoted μ_0 and it's fixed throughout the work) and hence we do not need to include the $t = 0$ boundary term, as the expression (2.4) is always 0.
Note also that (2.3) does not include space boundary terms as we will always assume M is a manifold without boundary.

For self-containment, we recall the specific version of Arzela-Ascoli in topological spaces that we will use in this document.

Proposition 155. *(General version of Arzela-Ascoli)*
Let X be a topological space and Y a metric space, let H be an equicontinuous family of functions from X to Y such that for every $x \in X$, $H(x) := \{h(x) : h \in H\}$ is relatively compact in Y. Then H is relatively compact with respect to the compact topology.

For a proof see [Bourbaki, Corollary 1 to Theorem 2, section V].

2.2.1 Assumptions on the potential

Equation (2.1) may have no solutions if the potential function is not appropriate. Given that our goal is to study measure-valued solutions of the equation, we are going to use the energy E_W associated to a potential function $W : M \times M \to \mathbb{R}$ given by

$$E_W(\rho) = \frac{1}{2} \int_M \int_M W(x,y) d\rho(x) d\rho(y). \tag{2.5}$$

We first assume that the interaction depends only through the Riemannian distance between points so

(W0) Without loss of generality, we assume $h(0) = 0$.

(W1) $W(x,y) = h(d(x,y)^2)$ where $h : [0,\infty) \to \mathbb{R}$ is twice continuously differentiable on $(0, diam(M)^2)$, further assume $h'(0)$ exists (as the limit from the right). Namely, we assume $h \in C^1([0, diam(M)^2]) \cap C^2_{loc}((0, diam(M)^2)$.

(W2) There exists $0 < r_h < diam(M)^2$ such that h is non-decreasing on $[r_h, diam(M)^2)$.

Remark 156. *Note that by triangle inequality $x \to d(x,p)$ is a Lipschitz function with constant 1 for all $p \in M$. Because h' is continuous on $[0, diam(M)^2]$, which is compact, h is Lipschitz and hence, for every fixed y the interaction potential $W_y(x) = W(x,y) = h(d(x,y)^2)$ is also Lipschitz (as a function of x). Denoting by* Lip *the Lipschitz constant, for every $y \in M$, the potential W satisfies* $\mathrm{Lip}(W_y) \leq 2\,\mathrm{Lip}(h) diam(M)$. *For this reason we define $L := 2\,\mathrm{Lip}(h) diam(M)$.*
Further, Rademacher's Theorem ensures that for every $y \in M$ the function W_y is differentiable dvol-almost everywhere. In general this conclusion will not be enough as some of the measures involved may not be absolutely continuous.

2.2.2 Non-differentiability of the potential

In order to show that the limit of the minimizing movement scheme (2.6) solves the aggregation equation, we will need the potential function to be differentiable. So far, our interaction potential with assumptions **(W0)**-**(W1)** fails to be differentiable at the cut locus. We will use the finite speed of propagation (Proposition 182) to ensure the potential is differentiable in the support of the measures involved in the minimizing movement scheme.
Note that assumptions **(W0)**-**(W1)** guarantee:

- The energy E_W is proper ($\{\rho \in \mathcal{P}(M) : E_W(\rho) < \infty\} = \mathcal{P}(M) \neq \emptyset$).

- The energy E_W is lower semi-continuous with respect to weak (i.e. narrow) convergence.

Remark 157. *As a consequence of compactness of M, W is bounded from below by a constant that we will denote by k, that is*

$$\inf_{(x,y)\in M\times M} W(x,y) \geq k.$$

The approach is to use the so called "Minimizing Movement Scheme" from [Jordan-Kinderlehrer-Otto],[Ambrosio-Gigli-Savare],[De Giorgi],[Almgren-Taylor-Wang]. The scheme consists in taking a time step of size $\tau > 0$ to balance the contribution of the original energy (in our case E_W) and a term that penalizes moving away from the previous configuration. The minimizer in this scheme approximates a step in the direction of steepest descent, getting more accurate as τ approaches 0.

Main problem

Given an intrinsic potential $W : M \times M \to \mathbb{R}$ satisfying assumptions **(W0)**,**(W1)**,**(W2)** does there always exist a measure-valued solutios (Definition 2.3) to the aggregation equation (2.1) for all times $t \in [0,\infty)$? Can this solutions be numerically approximated by discretization schemes?
We will answer positively both questions (for small times) in Theorem 163 using the minimizing movement scheme that we now define.

Definition 158. *(Minimizing movement scheme on $(\mathcal{P}(M), d_2)$ for E_W)*
Let $\mu_0 \in \mathcal{P}(M)$ be fixed, for $\tau > 0$ if it is possible to define a sequence $\{\mu_k^\tau\}$ of probability measures such that

$$\mu_{k+1}^\tau \in \arg\min \left\{ E_W(\rho) + \frac{1}{2\tau} d_2^2(\rho, \mu_k^\tau) : \rho \in \mathcal{P}(M) \right\}; \tag{2.6}$$

we call $\{\mu_k^\tau\}$ a sequence of the minimizing movement scheme for E_W at level τ.

Proposition 159. *(Existence of minimizing movement scheme)*
*Let $\mu_0 \in \mathcal{P}(M)$ be fixed assume that W satisfies **(W0)-(W1)**, for $\tau > 0$ the minimizing movement scheme μ_k^τ is well-defined.*

Proof. Note that the functional in (2.6) is lower semi-continuous with the assumptions **(W0)**- **(W1)** as the distance to any given measure is lower semi-continuous by triangle inequality. Hence, the lower semi-continuous functional on a compact set achieves a minimum yielding existence of a sequence $\{\mu_k^\tau\}_{k\in\mathbb{N}}$ for every $\tau > 0$. ∎

Existence of minimizers of the scheme does not ensure the model (2.1) will be solved by any time interpolation, the rest of the work is dedicated to interpolating the measures in a continuous way and showing the limiting measure solves the aggregation equation.
Because our goal is to solve (2.3), we need time interpolation of the sequences in the minimizing movement scheme. We denote by $\underset{\tau}{\mathbf{PC}}(\{\mu_k^\tau\})(t)$ the piecewise constant interpolation such that $\underset{\tau}{\mathbf{PC}}(\{\mu_k^\tau\})(t) = \mu_k^\tau$ if $t \in [k\tau, (k+1)\tau)$. Finally, we define the geodesic interpolation by the following formula, for $t \in [k\tau, (k+1)\tau)$

$$\mathbf{Geo}_\tau(\{\mu_k^\tau\}) = \exp_x \left(\left(\frac{(k+1)\tau - t}{\tau} \right) \nabla \phi_{k,k+1}^c \right)_{\#} \mu_{k+1}^\tau, \tag{2.7}$$

where $\phi_{k,k+1}$ is the Kantorovich potential from μ_k^τ to μ_{k+1}^τ and $\phi_{k,k+1}^c(x)$ is the $c-$transform (or infimal convolution) of $\phi_{k,k+1}$ given by

$$\phi_{k,k+1}^c(x) = \inf_{z\in M}\left\{\frac{d(x,z)^2}{2} - \phi_{k,k+1}(z)\right\}.$$

As Kantorovich potentials with respect c always exist [McCann, Proposition 3] we will need to show their differentiability. We obtain this condition in Lemma 180 and so for $\mathbf{Geo}_\tau(\{\mu_k^\tau\})$ to be well-defined we need the geodesic map to be uniquely defined for every point. Therefore to use the geodesic interpolation there can be no points in $\mathrm{spt}(\mu_{k+1}^\tau)$ in the cut locus of $\mathrm{spt}(\mu_k^\tau)$. This condition will be ensured by Proposition 182 for the time-interval specified in Theorem 163.

Remark 160. *A function f is called $c-$concave if it is not $-\infty$ and is the $c-$transform of another function. The fundamental theorem of optimal transport says that optimal plans in (2.2) are supported in $c-$subdifferentials of $c-$concave functions, see [Ambrosio-Gigli-Savare2, Theorem 1.13].*
Given measures $\mu,\nu\in\mathcal{P}(M)$ it is not necessarily true that the Kantorovich potential $\phi_{\mu,\nu}$ from μ to ν is differentiable in $\mathrm{spt}(\mu)$. Some conditions are always necessary (e.g. absolute continuity of the source measure) to ensure differentiability. The optimality criterion (Euler-Lagrange) of Lemma 181 will yield differentiability as we will see in Proposition 180.

Square estimates on the Wasserstein norms

Proposition 161. *Let $\{\mu_k^\tau\}$ be a minimizing movement scheme for E_W as in (2.6), i.e. $\{\mu_k^\tau\}$ satisfies*

$$\mu_{k+1}^\tau \in \operatorname*{arg\,min}_{\mu\in\mathcal{P}^2(M)}\left\{E_W(\mu) + \frac{d_2(\mu,\mu_k)^2}{2\tau}\right\}.$$

Then there exists a constant $C>0$ independent of τ such that

$$\sum_{k=0}^{\infty}\frac{d_2(\mu_k^\tau,\mu_{k+1}^\tau)^2}{\tau} \le C. \tag{2.8}$$

Proof. The proof is standard and can be found in [Villani2003, Section 8.4.1], presented here for completeness. Note that the optimality condition of μ_{k+1}^τ implies

$$E_W(\mu_{k+1}^\tau) + \frac{d_2(\mu_k^\tau,\mu_{k+1}^\tau)^2}{2\tau} \le E_W(\mu_k^\tau).$$

Hence, given that E_W is proper, the sequence gives finite values for E_W and so

$$\frac{d_2(\mu_k^\tau,\mu_{k+1}^\tau)^2}{2\tau} \le E_W(\mu_k^\tau) - E_W(\mu_{k+1}^\tau).$$

By summing all the terms, we get a telescopic sum on the right hand side, and the fact that E_W is bounded from below by k (Remark 157) gives

$$\sum_{k=0}^{\infty}\frac{d_2(\mu_k^\tau,\mu_{k+1}^\tau)^2}{2\tau} \le E_W(\mu_0) - k,$$

where k is the lower bound of E_W obtained by compactness (Remark 157), hence putting $C := 2(E_W(\mu_0)-k)$ gives the claim as it is finite and independent of τ. ∎

2.2.3 Statement of small time existence of measure valued solutions

We aim to analyze measure-valued solutions to (2.1). The main technical difficulty is dealing with the fact that an interaction potential $W(x,y) = h(d(x,y)^2)$ may not be differentiable at the cut locus $Cut \subseteq M \times M$ (see section 2.2.2). The aggregation equation together with the minimizing movement scheme (2.6) will be shown to satisfy a finite-speed of propagation Proposition 182. This means that if we start with a probability measure concentrated away from the cut locus, we can apply the d_2-gradient flow method to generate solutions to the equations for small times.
As we are going to use the gradient of the interaction potential, we must ensure W remains differentiable. We are going to prove that if the initial measure μ_0 is concentrated away from the cut locus, solutions (for small time) exist in measure sense. The idea is that the minimizing movement scheme will not instantly move to the cut locus, it needs time to spread.

Definition 162. *(Distance to Cut)*
Let (M,g) be smooth, compact, connected Riemannian manifold, for $\mu \in \mathcal{P}(M)$ we define the distance to cut locus, δ_μ, as the distance between every pair on the support to the cut locus, i.e.

$$\delta_\mu := \inf_{\substack{x,y \in \mathrm{spt}(\mu) \\ (x',y') \in Cut}} \{d(x,x') + d(y,y')\}. \tag{2.9}$$

Theorem 163. *(Local existence of measure solutions to the aggregation equation)*
*Given $\mu_0 \in \mathcal{P}^2_{ac}(M)$ let δ_μ be the distance to cut as in Definition 162, if $\delta_\mu > 0$ and L denotes the Lipschitz constant of W (from **(W1)**); under **(W0)**-**(W2)**, for every $0 < T < \frac{\delta_\mu}{2L}$ there exists a sequence from the minimizing movement scheme for E_W at level $\tau > 0$ starting at μ_0 such that as $\tau \to 0$ the geodesic interpolation $\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)$ converges in d_2-metric to a path $\mu(t)$ which is a measure valued solution (in the sense of (2.3)) to the aggregation equation on M (2.1) up to time T.*

The proof of Theorem 163 is the main goal of this document and will occupy the rest of the article. We will ensure the convergence of the minimizing movement scheme using a general version of the Arzela-Ascoli (Proposition 155 that can be found in [Bourbaki, Corollary 1 to Theorem 2, section V].

2.2.4 Continuity and optimality

Definition 164. *(Absolute continuity in $(\mathcal{P}(M), d_2)$)*
We say that a curve $t \to \rho_t$ mapping (a,b) to $(\mathcal{P}(M), d_2)$ is absolutely continuous if there exists an integrable (w.r.t Lebesgue) function $g : (a,b) \to \mathbb{R}$ such that

$$d_2(\rho_t, \rho_s) \leq \int_s^t g(r)dr.$$

And we say it is p-absolutely continuous if g is L^p-integrable.

Definition 165. *(Norm of metric derivative)*
Let I be an interval and $\rho_t \in \mathcal{P}(M)$ for every $t \in I$, we call the metric derivative (or slope of metric derivative or speed) the function

$$|\rho_t'| := \lim_{h \to 0} \frac{d_2(\rho_{t+h}, \rho_t)}{h}$$

whenever it exists.

Lemma 166. *(Metric derivative for p-absolutely continuous curves)*
Let $\{\rho_t\}_{t\in[a,b]}$ be p-absolutely continuous. Then $|\rho_s'|$ exists Lebesgue a.e. and $t \to |\rho_t'|$ is also p-integrable in (a,b).

Proof. As presented in [Ambrosio-Gigli-Savare, Theorem 1.1.2], letting $\{y_n\}$ be dense in $\{\rho_s\}_{s\in(a,b)}$ one can check that

$$\liminf_{t\to s} \frac{d(\rho_s,\rho_t)}{|t-s|} \geq \sup_n \liminf_{t\to s} \frac{|d(y_n,\rho_s) - d(y_n,\rho_t)|}{|t-s|}$$

from which the result follows. ∎

Lemma 167. *($\frac{1}{2}$-Hölder continuity of geodesic interpolation uniformly in τ)*
For $T>0$ if $\mathbf{Geo}_\tau(\{\mu_k^\tau\}(t)$ denotes the geodesic interpolation (as in equation (2.7)) on the interval $[0,T]$, then $\{\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)\}_{t\in[0,T]}$ is $\frac{1}{2}$-Hölder uniformly continuous, i.e. there exists $\tilde{C}>0$ independent of τ such that

$$d_2(\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t), \mathbf{Geo}_\tau(\{\mu_k^\tau\})(s)) \leq \tilde{C}(t-s)^{1/2}. \tag{2.10}$$

Proof. Let $\tau>0$ be fixed, if $t_1,t_2 \in [k\tau,(k+1)\tau)$ and $t_1>t_2$ then the geodesic property of the exponential map yields:

$$d_2(\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t_1), \mathbf{Geo}_\tau(\{\mu_k^\tau\})(t_2)) = \left(\int |\frac{t_1-t_2}{\tau}\nabla\phi^c_{k,k+1}|^2 d\mu^\tau_{k+1}\right)^{1/2} \tag{2.11}$$

$$= \left(\frac{t_1-t_2}{\tau}\right) d_2(\mu_k^\tau,\mu_{k+1}^\tau). \tag{2.12}$$

Hence, by definition of the metric derivative we obtain

$$|\mathbf{Geo}_\tau(\{\mu_k^\tau\})'(t)| = \lim_{h\to 0} \frac{(h/\tau)d_2(\mu_k^\tau,\mu_{k+1}^\tau)}{h} = \frac{d_2(\mu_k^\tau,\mu_{k+1}^\tau)}{\tau}.$$

With this calculation in mind, we compute using Hölder's inequality and Proposition 161,

$$d_2(\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t), \mathbf{Geo}_\tau(\{\mu_k^\tau\})(s))$$
$$= \int_s^t |\mathbf{Geo}_\tau(\{\mu_k^\tau\})'(r)|dr \leq (t-s)^{1/2}\left(\sum_{k=1}^\infty \frac{d_2(\mu_k^\tau,\mu_{k+1}^\tau)^2}{\tau}\right)^{1/2} \leq C(t-s)^{1/2}$$

∎

Recall that by Corollary 169 we have ensured the existence of a limiting measure path (as $\tau\to 0$) of the geodesic interpolation of the minimizing movement scheme for E_W. We observe that uniform (on τ) absolute continuity (Lemma 167) implies absolute continuity of the limiting path.

Corollary 168. *(The limit shares the Hölder constant)*
Let $T>0$ and suppose that for every $t\in[0,T]$ we have that as $\tau\to 0$, $\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t) \xrightarrow{d_2} \mu(t)$, then $\mu(t)$ is 1/2-Hölder continuous in $[0,T]$ with constant C.

Proof. Given $\epsilon > 0$, there exists $\tau = \tau(\epsilon)$ such that $d_2(\mu(t), \mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)) < \frac{\epsilon}{2}$ and $d_2(\mu(s), \mathbf{Geo}_\tau(\{\mu_k^\tau\})(s)) < \frac{\epsilon}{2}$ from which applying the previous result (Lemma 167) and triangle inequality we obtain

$$d_2(\mu(t), \mu(s)) \leq \epsilon + C(t-s)^{1/2}.$$

Because ϵ is arbitrary and C does not depend on τ we get the result. ∎

Corollary 169. *(Existence of a Limiting path)*
Fix $T > 0$, suppose that $\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)$ is defined for all $\tau \in (0,1]$ and for all $t \in [0,T]$. Then there exists a subsequence τ_n, with $\tau_n \to 0$ and a curve $\mu : [0,T] \to (\mathcal{P}_2(M), d_2)$ such that

$$\sup_{t\in[0,T]} d_2(\mathbf{Geo}_\tau(\{\mu_k^{\tau_n}\})(t), \mu(t)) \to 0 \ \ (\text{as } n \to 0). \tag{2.13}$$

Proof. Note that Lemma 167 (proved in the next section) shows uniform equicontinuity of the family, as the Hölder constant does not depend on τ. For every $t \in [0,T]$, the family $\{\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)\}$ is tight by Prokhorov's theorem and because M is compact, it is also d_2 relatively compact as the d_2-Lipschitz constant from Lemma 167 is independent of τ, $H = \{\mathbf{Geo}_\tau(\{\mu_k^\tau\}) : [0,T] \to \mathcal{P}_2(M)\}_\tau$ satisfy the hypothesis of Proposition 155 and the result follows. ∎

Alternatively one can directly use [Ambrosio-Gigli-Savare, Proposition 3.3.1].

Remark 170. *By Corollary 169 we know that as long as we can define the geodesic interpolation up to time T, we obtain the existence of a limiting path $\mu(t)$. We have yet to show that this limiting path $\mu(t)$ satisfies (2.1), for which we will work with the Euler-Lagrange conditions of the minimizing movement scheme (2.6).*

Recall that if v_t is a Borel integrable velocity field and (μ_t, v_t) satisfies the continuity equation in the sense of distributions, then for every $f \in C_c^\infty(M)$

$$\frac{d}{dt}\int f(x) d\mu_t = -\int \langle \nabla f(x), v_t(x)\rangle_x d\mu_t(x).$$

See for example [Santambrogio, Proposition 4.2].

Lemma 171. *(Computation of the velocity field)*
Letting (M, g) be a smooth, connected, compact manifold without boundary, the velocity field of the geodesic interpolation of the minimizing movement scheme $\mathbf{Geo}_\tau(\{\mu_k^\tau\})$ is given by parallel transporting the gradient of the c-transform of it's Kantorovich potential on each interval.

Proof. Suppose that $\mu_t = \exp_x(tv(x))_\# \mu_0$ for some $\mu_0 \in \mathcal{P}_{ac}(M)$ and a differentiable map $v : M \to TM$, then by compactness of M and dominated convergence,

$$\begin{aligned}\frac{d}{dt}\int f(x) d\mu_t(x) &= \int \frac{d}{dt} f(\exp_x(tv(x))) d\mu_0(x) \\ &= \int \langle \nabla f(\exp_x(tv(x))), \Pi_{t,v(x)}(v(x))\rangle_{\exp_x(tv(x))} d\mu_0(x),\end{aligned}$$

where $\Pi_{t,v(x)}$ denotes parallel transport along the geodesic $t \to \exp_x(tv)$.
Now denote $T_t^\tau(x) = \exp_x\left((\frac{(k+1)\tau - t}{\tau})\nabla\phi_{k,k+1}^c(x)\right)$ and let v_t^τ be such that

$$\frac{\partial T_t}{\partial t}(x) = v_t^\tau(T_t(x)) = (d\exp_x)_{-t\frac{\nabla\phi_{k,k+1}^c}{\tau}}\left(-\frac{\nabla\phi_{k,k+1}^c}{\tau}\right).$$

Because the differential of the exponential map at 0 is the identity operator we get that by Taylor expansion for $t \in [k\tau, (k+1)\tau)$

$$v_t^\tau(x) = -\frac{\nabla\phi_{k,k+1}^c(x)}{\tau} + R_t^\tau(x),$$

where $R_t^\tau(x) \in T_xM$ and satisfies that as $k\tau \to t$ (equivalently $\tau \to 0$) we have $|R_t^\tau|_x \to 0$. ∎

Remark 172. *Replacing the differential of the exponential for parallel transport can only be done because the evaluation is at the direction of the geodesic. If evaluated at a different vector, the differential of the exponential and parallel transport do not coincide as maps (but the norm of their difference is bounded see [Criscitiello-Boumal, Proposition 2.8] in our case $s = \dot{s}$ in their notation).*

Definition 173. *(First variation of a functional in $\mathcal{P}(M)$)*
Let F be a functional $F : \mathcal{P}(M) \to \mathbb{R}$, let $\rho \in \mathcal{P}(M)$ be fixed and $\epsilon > 0$, for any $\tilde{\rho} \in \mathcal{P}_{ac} \cap L^\infty(M)$, define $\nu = \tilde{\rho} - \rho$, we say that $\frac{\delta F}{\delta \rho}(\rho)$ is the first variation of F evaluated at ρ if

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} F(\rho + \epsilon\nu) = \int \frac{\delta F}{\delta \rho}(\rho) d\nu.$$

Proposition 174. *(Optimality criteria)*
For a functional $F : \mathcal{P}(M) \to \mathbb{R}$ suppose that $\mu \in \underset{\nu\in\mathcal{P}(M)}{\arg\min} F(\nu)$. Assume that for every $\epsilon > 0$ and for every ρ absolutely continuous with $L^\infty(M)$ density

$$F((1-\epsilon)\mu + \epsilon\rho) < \infty$$

Let $\tilde{c} := essinf\left\{\frac{\delta F}{\delta \rho}(\mu)\right\}$. If $\frac{\delta F}{\delta \rho}(\mu)$ is continuous, then

$$\frac{\delta F}{\delta \rho}(\mu)(x) \geq \tilde{c} \ \forall x \in M, \tag{2.14}$$

$$\frac{\delta F}{\delta \rho}(\mu)(x) = \tilde{c} \ \forall x \in \mathrm{spt}(\mu). \tag{2.15}$$

The proof can be found [Santambrogio, Theorem 7.20].

Lemma 175. *(Computation of first variations)*
For each of the following cases let μ satisfy for each functional F the hypothesis of the last theorem, then

$$\begin{cases} \frac{\delta F}{\delta \rho}(\mu) = \phi_{\mu,\nu} \text{ if } F(\mu) = \frac{d_2^2(\mu,\nu)^2}{2}, \\ \\ \frac{\delta F}{\delta \rho}(\mu) = 2(W * \mu) \text{ if } F(\mu) = \int_M \int_M W(x,y) d\mu(x) d\mu(y), \end{cases}$$

where as before $\phi_{\mu,\nu}$ is the Kantorovich potential whose negative gradient pushes μ to ν optimally with respect to $d^2/2$.

Proof. The first computation can be found in [Santambrogio, Proposition 7.16], while for the second one, note that

$$\begin{aligned} F(\rho+\epsilon\nu) &= \int_M \int_M W(x,y)d(\rho+\epsilon\nu)(x)d(\rho+\epsilon\nu)(y) \\ &= F(\rho)+\epsilon^2 F(\nu)+\epsilon\left(\int_M\int_M W(x,y)d\rho(x)d\nu(y)+\int_M\int_M W(x,y)d\rho(y)d\nu(x)\right). \end{aligned}$$

where the result is obvious by dividing by ϵ and taking the limit.
Clearly if W is symmetric, as in the case of assumptions (2.2.1),

$$\frac{\delta F}{\delta \rho}(\mu) = 2\int_M W(x,y)d\mu(y) = 2(W*\mu)(x).$$

∎

The Kantorovich potentials are known to exist in general settings such as Polish spaces but the question of their regularity is usually more subtle (e.g. [Villani2009, Theorem 10.8]). We recall the concepts of semiconcavity/semiconvexity from non-smooth analysis for which we follow [Cordero-Erausquin-McCann-Schmuckenschläger]. We will show in Lemmatta 179 and 180 semiconcavity of both the Kantorovich potential and the convolution of the interaction which together yield differentiability of the infimal convolution as well. For these proofs, compactness of M seems essential.

Definition 176. *(Locally semi-concave)*
Let $U \subseteq M$ be open, we say $f: U \to \mathbb{R}$ is semi-concave at x_0 if there exists a neighborhood of x_0 and a constant $C \in \mathbb{R}$ such that for every $x \in U$ and $v \in T_xM$

$$\limsup_{r\to 0} \frac{f(\exp_x(rv))+f(\exp_x(-rv))-2f(x)}{r^2} \leq C, \tag{2.16}$$

where $\exp_x$ denotes the exponential at x.

Remark 177. *In [Cordero-Erausquin-McCann-Schmuckenschläger] it is shown that semi-concave functions admit non-empty superdifferentials, which implies that semi-concavity together with semi-convexity yields differentiability. It is also shown there that c-concave functions are semi-concave ([Cordero-Erausquin-McCann-Schmuckenschläger, Proposition 3.14]) and that $x \to d(x,y)^2$ is everywhere semi-concave but fails to be semi-convex at the cut locus ([Cordero-Erausquin-McCann-Schmuckenschläger, Proposition 2.5]). We refer to [Cordero-Erausquin-McCann-Schmuckenschläger] for details and proofs, specifically see [Cordero-Erausquin-McCann-Schmuckenschläger, Lemma 3.11]*

Lemma 178. *(Joint smoothness or Riemannian distance squared away from cut locus)*
In the context of our smooth, connected compact manifold (M,g), the square of Riemannian distance is smooth away from the cut locus, i.e. if $(x_0,y_0) \notin Cut$ then $(x,y) \to d(x,y)^2$ is smooth in a neighborhood of (x_0,y_0).

Proof. This lemma can be understood as a particular case to [McCann2020, Theorem 3.6 c)], so we follow the proof from there. Because Cut is closed, if $(x_0, y_0) \notin Cut$ there exists a ball around (x_0, y_0) not intersecting Cut. Let (x, y) be an element of such a ball, we aim to show distance squared is smooth at (x, y). Note that by the inverse function theorem the function $(x, v) \to (x, \exp_x(v))$ acts as a smooth diffeomorphism in such a neighborhood of $(x_0, \exp_x^{-1}(y_0))$. By symmetry, $(y, w) \to \exp_y(w)$ acts as a smooth diffemorphism in a neighborhood of $(y_0, \exp_y^{-1}(x_0))$. Given $z \in T_xM$ denote by $z_* = \langle z, \cdot \rangle_x \in T_x^*M$ (it's dual covector) then by differentiating (which we can do as the function is locally Lipschitz and semi-convex [McCann, Theorem 3.6 d)]) we get the formula for the gradient:

$$-\nabla_M(d(x,y)) = \left(\frac{v_*}{|v_*|_x}, \frac{w_*}{|w_*|_y} \right) \Bigg|_{(v,w)=(\exp_x^{-1}(y), \exp_y^{-1}(x))} \tag{2.17}$$

as $\exp_x^{-1}(y_0), \exp_y^{-1}(x_0)$ are the tangent vectors of the geodesic (that exists as $(x_0, y_0) \notin Cut$) we conclude from (2.17) that all components depend smoothly on (x, y) yielding the result (see [McCann, Theorem 3.6] for more details). ∎

Lemma 179. *(Semiconcavity of convolution)*
*Let $\mu \in \mathcal{P}(M)$ Borel, assume that $W(x, y)$ satisfies assumptions **(W0)-(W2)** with $r_h \leq t_{inf}$, i.e. h is non-decreasing on $[t_{inf}, diam(M)^2]$, then everywhere on M the function $x \to (W * \mu)(x)$ is semiconcave.*

Before we write the rigorous proof, let us outline the idea: the proof consists in looking at the convolution as a sum of integrals on different regions where the regularity of $x \to d(x, y)^2$ changes significantly.
In the first region, h and d^2 are of class C^2 and we can use compactness and differentiability of these functions to deduce semi-concavity. The second region requires a more subtle analysis, the non-decreasing property of h from assumption **(W2)** allows us to use the Chain rule for super-gradients [McCann, Lemma 5] from which (with some technical work) we can deduce semi-concavity.

Proof. Note that we can decompose the convolution:

$$W * \mu(x) = \int_{\{y : d(x,y) \leq \sqrt{r_h}\}} W(x, y) d\mu(y) + \int_{\{y : d(x,y) > \sqrt{r_h}\}} W(x, y) d\mu(y). \tag{2.18}$$

where r_h is as in **(W2)**.
For the first term, as in [Cordero-Erausquin-McCann-Schmuckenschläger, Proposition 2.5] if $d(x_0, y) < t_{Cut(x_0)}$ then $x \to d(x, y)^2$ is smooth at x_0 and so $h(d(x, y)^2)$ is semiconcave as a C^2 function on a compact set has a bounded Hessian. Notice that the bound may depend on y but $(x, y) \to d(x, y)^2$ is jointly smooth away from the cut locus by Lemma 178 so $y \to \mathrm{Hess}_x W(x, y)$ is continuous and because M is complete, $\{y : d(x, y) \leq \sqrt{r_h}\}$ is compact and hence the x-Hessian of the first term in (2.18) is uniformly upper bounded.
For the second term of (2.18), we note that in the region $\{y : d(x, y) > \sqrt{r_h}\}$ the assumption **(W2)** ensures the hypothesis of the chain rule for supergradients ([McCann, Lemma 5]) is satisfied for $x \to h(d(x, y)^2)$, as $x \to d(x, y)^2$ is everywhere semiconcave, h is twice differentiable there and $r \to h(r)$ is non-decreasing on $\{r > r_h\}$. The chain rule for supergradients together with [Cordero-Erausquin-McCann-Schmuckenschläger, Corollary 3.13] imply there exists $C > 0$ such that for every $(x, y) \in M$ and $u \in T_xM$ with $d(x, y) \geq r_h$,

$$\limsup_{r \to 0^+} \frac{W(\exp_x(ru), y) + W(\exp_x(-ru), y) - 2W(x, y)}{r^2} \leq C. \tag{2.19}$$

Let us use the following notation:

$$f_r(x,y,u) := \frac{W(\exp_x(ru),y) + W(\exp_x(-ru),y) - 2W(x,y)}{r^2}. \tag{2.20}$$

To conclude semi-concavity of the second term in (2.18), we will use reverse Fatou's Lemma. Because the functions $y \to f_r(x,y,u)$ are not necessarily positive, to apply reverse Fatou's lemma we need to show they are dominated by some $L^1(\mu)$ function.
Observe that (2.19) is not enough to apply reverse Fatou's lemma as a-priori the bound on the limit superior does not yield a uniform bound on $\{y \to f_r(x,y,u)\}_{r>0}$. To obtain this uniform bound, note that (2.19) holds for every (x,y) in the region of integration for the first term in (2.18), by [Cordero-Erausquin-McCann-Schmuckenschläger, Lemma 3.11] and the definition of semi-concavity we obtain that for every x_0 there exists $r^*_{x_0} > 0$ and a smooth function $V : B_{r^*_{x_0}}(x_0) \to \mathbb{R}$ such that $x \to W(x,y) + V(x)$ is geodesically C-concave at x_0 for every $y \in \{y : d(x_0,y) \geq r_h\}$. Notice that the condition on y is due because only in that region we can use the Chain rule for super-gradients ([McCann, Lemma 5]).
When $r < r^*_{x_0}$ C-concavity (as x_0 is the midpoint between $\exp_{x_0}(ru)$ and $\exp_{x_0}(-ru)$), yields a uniform bound for $y \to f_r(x_0,y,u)$. Hence, for all quadruples $(r,x_0,y,u) \in \{(r,x_0,y,u) : 0 < r < r^*_{x_0}, d(x_0,y) > r_h, u \in T_{x_0}M\}$, where $x_0 \in M$ is fixed, the functions $y \to f_r(x_0,y,u)$ satisfy a uniform upper bound $f_r(x,y,u)$ and therefore using reverse Fatou's Lemma:

$$\limsup_{r\to 0} \int_{\{y:d(x,y)^2>r_h\}} \frac{W(\exp_x(ru),y) + W(\exp_x(-ru),y) - 2W(x,y)}{r^2} d\mu(y)$$
$$\leq \int_{\{y:d(x,y)^2>r_h\}} C d\mu(y) \leq C,$$

which is semi-concavity of the convolution (with the same constant) as desired. ∎

Lemma 180. *(Differentiability of Kantorovich potentials in the whole support)*
*Let $\phi_{k,k+1}$ be the Kantorovich potential from μ^τ_k to μ^τ_{k+1} from Definition 2.6 where W satisfies **(W0)-(W2)**, then its c-transform $\phi^c_{k,k+1}$ is differentiable at x for every $x \in \mathrm{spt}(\mu^\tau_{k+1})$.*

Proof. By optimality (Proposition 174) we know that for every $x \in \mathrm{spt}(\mu^\tau_{k+1})$

$$\frac{\phi^c_{k,k+1}}{\tau}(x) + W * \mu^\tau_{k+1}(x) \geq \tilde{c} \text{ with equality on } \mathrm{spt}(\mu^\tau_{k+1}).$$

By definition of the infimal convolution $\phi^c_{k,k+1}$ is $c-$concave and hence semiconcave as in [Cordero-Erausquin-McCann-Schmuckenschläger, Proposition 3.14].
By assumptions **(W0)-(W2)** we apply Lemma 179 to conclude that $W * \mu^\tau_{k+1}(x)$ is semiconcave. Hence, everywhere in $\mathrm{spt}(\mu^\tau_{k+1})$, $\phi^c_{k,k+1} = \tau(\tilde{c} - W * \mu^\tau_{k+1})$ is also semiconvex meaning that $\phi^c_{k,k+1}$ is both semiconcave and semiconvex and hence continuously differentiable at $x \in \mathrm{spt}(\mu^\tau_{k+1})$. ∎

Lemma 181. *(First variations for the minimizing movement scheme)*
Let $\{\mu^\tau_k\}_{k\in\mathbb{N}}$ be the minimizing movement scheme with initial measure $\mu_0 \in \mathcal{P}^2_{ac}(M)$, let $\phi_{k,k+1}$ be the Kantorovich potential for which the exponential of it's negative gradient pushes μ^τ_k onto μ^τ_{k+1}, then on the support of μ^τ_{k+1}

$$-\frac{\nabla_M \phi^c_{k,k+1}}{\tau} = \nabla_M (W * \mu^\tau_{k+1}).$$

Proof. To agree with notation, let $F(\nu) = \dfrac{d_2^2(\nu, \mu_k^\tau)}{2\tau} + E(\nu)$. Theorem (174) says

$$\frac{\phi_{k,k+1}^c}{\tau} + W * \mu_{k+1}^\tau = \frac{\delta F}{\delta \rho}(\mu_{k+1}^\tau) = c \text{ on } \operatorname{spt}(\mu_{k+1}^\tau).$$

Using the previous lemma taking the gradient on both sides gives the result. ∎

2.3 Finite speed of propagation and proof of the main theorem 163

In this section we take a look at a consequence of the Euler-Lagrange condition that will ensure that the evolution of the measures is controlled. This proposition is key to ensure differentiability of the interaction potential needed to conclude convergence in the continuity equation.

Proposition 182. *(Finite speed of propagation in the minimizing movement scheme) Given $\tau > 0$ and $\mu_k^\tau \in \mathcal{P}(M)$, let $L > 0$ denote the Lipschitz constant of the potential from **(W0)**-**(W1)**, if*

$$\mu_{k+1}^\tau \in \operatorname*{arg\,min}_{\rho \in \mathcal{P}(M)} \left\{ \frac{1}{2} \int \int W(x,y) d\rho d\rho + \frac{1}{2\tau} d_2(\mu_k^\tau, \rho)^2 \right\}$$

we have

$$\operatorname{spt}(\mu_{k+1}^\tau) \subseteq \{x \in M : d(x, \operatorname{spt}(\mu_k^\tau)) \leq L\tau\}.$$

Proof. By Lemma 181 we know we can compute the gradient of $\phi_{k,k+1}^c$, which is defined μ_{k+1}^τ everywhere on $\operatorname{spt}(\mu_{k+1}^\tau)$ and hence the map $x \to \exp_x(\nabla \phi_{k,k+1}^c(x))$ is well defined and supported in the subdifferential of a $c-$concave map $(\phi_{k,k+1}^c)$, by the converse as [Ambrosio-Gigli-Savare2, Proposition 1.30] we get that this map is optimal and pushes μ_{k+1}^τ to μ_k^τ meaning that

$$d_2(\mu_{k+1}^\tau, \mu_k^\tau)^2 = \int_M |\nabla \phi_{k,k+1}^c|^2 d\mu_{k+1}^\tau. \tag{2.21}$$

Now by assumptions **(W0)**-**(W1)** we know that (Remark 156) $y \to W(x,y)$ is Lipschitz for every $x \in \operatorname{spt}(\mu_{k+1}^\tau)$, from which the convolution is Lipschitz with the same constant L, as μ_{k+1}^τ is a probability measure, i.e.

$$|W * \mu_{k+1}^\tau(x_1) - W * \mu_{k+1}^\tau(x_2)| \leq L d(x_1, x_2).$$

Consequently, because the norm of the gradient is bounded by the metric derivative ([Ambrosio-Gigli-Savare, Theorem 1.1.2]), we obtain that for every $x \in \operatorname{spt}(\mu_{k+1}^\tau)$

$$|\nabla_M (W * \mu_{k+1}^\tau)(x)|_x \leq L. \tag{2.22}$$

By Lemma 181 this means that

$$d(x, \exp_x(\nabla \phi_{k,k+1}^c(x))) = |\nabla \phi_{k,k+1}^c(x)| \leq L\tau.$$

Note that L is the global Lipschitz constant of the interaction potential and hence independent of τ. Consequently every point $x \in \operatorname{spt}(\mu_{k+1}^\tau)$ is transported at a distance of at most $L\tau$ from which triangle inequality yields the result. ∎

Corollary 183. *(Small time differentiability of the potential)*
Fix $\tau > 0$, let $\mu_0 \in \mathcal{P}_{ac}(M)$ with $\delta_{\mu_0} > 0$, where δ_{μ_0} is the distance to cut from Definition 162, then for all $t \in [0, \lfloor \frac{\delta_\mu}{2L\tau} \rfloor]$ the function $x \to W(x, y)$ is differentiable in the support of $\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)$ where μ_k^τ is the minimizing movement scheme at level τ defined in (2.6).

Proof. Observe that the finite speed of propagation (Proposition 182) immediately ensures that if $\mu_0 \in \mathcal{P}_{ac}(M)$ with $\delta_{\mu_0} > 0$ then

$$\mathrm{spt}(\mu_1^\tau) \subseteq \{x \in M : d(x, \mathrm{spt}(\mu_0)) \leq L\tau\}.$$

Consequently, as we have seen in Lemma 180 the map $x \to \exp_x(\nabla \phi_{k,k+1}^c(x))$ is well defined and is an optimal transport map (with cost $d^2/2$) which means that we can apply finite speed of propagation (Proposition 182) at every k and hence,

$$\mathrm{spt}(\mu_k^\tau) \subseteq \{x \in M : d(x, \mathrm{spt}(\mu_0)) \leq Lk\tau\}.$$

For $k \in \mathbb{N}$, consider $(x, y) \in \mathrm{spt}(\mu_0)^2, (x_k, y_k) \in \mathrm{spt}(\mu_k^\tau)^2$ and $(x', y') \in Cut$. By definition of δ_{μ_0},

$$d(x, x') + d(y, y') \geq \delta_{\mu_0}$$

Using the triangle inequality twice,

$$d(x_k, x') + d(x_k, x) + d(y_k, y') + d(y_k, y) \geq \delta_{\mu_0}. \tag{2.23}$$

By finite speed of propagation $2k$-times,

$$2kL\tau + d(x_k, x') + d(y_k, y') \geq \delta_{\mu_0}.$$

Now using definition 162, because $\delta_{\mu_k^\tau}$ is an infimum, (x_k, y_k) and (x', y') are arbitrary,

$$\delta_{\mu_k^\tau} \geq \delta_{\mu_0} - 2k\tau L. \tag{2.24}$$

Hence, as long as $\delta_{\mu_0} - 2k\tau L > 0$, the geodesic interpolation (2.7) guarantees that $x \to W(x, y)$ is differentiable in the support of the measures up to μ_k^τ.
Notice that $\delta_{\mu_0} - 2k\tau L > 0$ occurs exactly when $k < \delta_{\mu_0}/(2L\tau)$ as desired. ∎

Lemma 184. *(Contraction of Wasserstein distances for product measures)*
Let $\mu, \nu \in \mathcal{P}_1(M)$, denote by $\mu \otimes \mu$ and $\nu \otimes \nu$ the product measures on $M \times M$, then

$$d_1(\mu \otimes \mu, \nu \otimes \nu) \leq 2d_1(\mu, \nu)$$

Proof. Note that if $\pi \in \Pi(\mu, \nu)$ then $\pi \otimes \pi \in \Pi(\mu \otimes \mu, \nu \otimes \nu)$ and note that

$$\int_M d(x, y) d\pi(x, y) + \int_M d(\tilde{x}, \tilde{y}) d\pi(\tilde{x}, \tilde{y}) = \int_{M \times M} d_{M \times M}((x, \tilde{x}), (y, \tilde{y})) d\pi \otimes \pi(x, \tilde{x}, y, \tilde{y})$$

from which taking infimum on both sides yields the result. ∎

2.3.1 Evaluation of the limt

In this section we prove Theorem 163, the goal is to show that the limiting measure path from Corollary 169 satisfies the continuity equation. The idea is to use the d_2-convergence together with the finite speed of propagation to ensure $x \to W(x,y)$ is differentiable in the whole support of the measures at level τ, in order to differentiate inside of the convolution term in (2.1).

2.3.2 Proof of Theorem 163

Proof. Assumptions 2.2.1 on E_W ensure the hypothesis of Corollary 169 are satisfied in $(\mathcal{P}(M), d_2)$ which ensures the existence of a limiting path $\{\mu(t)\}_{t\in[0,\infty)}$ for the family of geodesic interpolations $\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)$.
This interpolation satisfies the continuity equation with v_t^τ given as in Lemma 171, by Proposition 174 we replace the vector field with $\nabla(W * \mu_{k+1}^\tau)$ and finally for $f \in C_c^\infty((0,T)\times M)$ we aim to compute

$$\lim_{\tau\to 0}\left(\underbrace{\int_0^T\int_M \partial_t f(x,t)d\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)dt}_{(i)} + \underbrace{\int_0^T\int_M \langle\nabla f(x,t), v_t^\tau\rangle_x d\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)dt}_{(ii)}\right)$$

The limit of (i): notice that $\partial_t f(x,t)$ is continuous on x so by $\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t) \xrightarrow{d_2} \mu(t)$ as $\tau\to 0$:

$$\int_M \partial_t f(x,t)d\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t) \xrightarrow{\tau\to 0} \int_M \partial_t f(x,t)d\mu(t) \qquad \forall t\in[0,T].$$

Because $[0,T]$ is compact, the Dominated Convergence Theorem yields

$$\int_0^T\int_M \partial_t f(x,t)d\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)dt \xrightarrow{\tau\to 0} \int_0^T\int_M \partial_t f(x,t)d\mu(t)dt.$$

To analyze (ii), denote $T_t^{\tau,k}(x) = \exp_x\left(\frac{((k+1)\tau - t)}{\tau}\nabla\phi_{k,k+1}^c(x)\right)$, with this notation $(T_t^{\tau,k})_\# \mu_{k+1}^\tau = \mathbf{Geo}_\tau(\{\mu_k^\tau\})(t)$ and so by definition and using the observation of Lemma 171, if $N_\tau = \lfloor T/\tau\rfloor$,

$$\begin{aligned}
(ii) &= \int_0^T\int_M \langle\nabla_M f(x,t), v_t^\tau(x)\rangle_x d\,\mathbf{Geo}_\tau(\{\mu_k\})(t)dt\\
&= \sum_{k=0}^{N_\tau}\int_{k\tau}^{(k+1)\tau}\int_M \langle\nabla_M f(T_t^{\tau,k}(x),t), \Pi_{t,\gamma_{k,\tau}}\left(-\frac{\nabla\phi_{k,k+1}^c(x)}{\tau}\right)\rangle_{T_t^{\tau,k}(x)} d\mu_{k+1}^\tau(x)dt\\
&\qquad - \int_T^{N_\tau+1}\int_M \langle\nabla_M f(T_t^{\tau,k}(x),t), \Pi_{t,\gamma_{k,\tau}}\left(-\frac{\nabla\phi_{k,k+1}^c(x)}{\tau}\right)\rangle_{T_t^{\tau,k}(x)} d\mu_{k+1}^\tau(x)dt \qquad (2.25)\\
&= -\sum_{k=0}^{N_\tau}\int_{k\tau}^{(k+1)\tau}\int_M \langle\Pi_{t,\gamma_{k,\tau}}^{-1}\left(\nabla_M f(T_t^{\tau,k}(x),t)\right), \frac{\nabla\phi_{k,k+1}^c(x)}{\tau}\rangle_x d\mu_{k+1}^\tau(x)dt\\
&\qquad + \int_T^{N_\tau+1}\int_M \langle\Pi_{t,\gamma_{k,\tau}}^{-1}\left(\nabla_M f(T_t^{\tau,k}(x),t)\right), \frac{\nabla\phi_{k,k+1}^c(x)}{\tau}\rangle_x d\mu_{k+1}^\tau(x)dt,
\end{aligned}$$

where $\Pi_{t,\gamma_{k,\tau}}$ is as in Lemma 171, and denotes parallel transport along the curve $\gamma_{k,\tau} : [\tau k, \tau(k+1)) \to M$ given by $\gamma_{k,\tau}(t) = T_t^{\tau,k}$ which satisfies $\gamma_{k,\tau}((k+1)\tau) = x$. The second equality follows from the fact that parallel transport is an isometry.

Focusing first in the inner most integral of the first term observe that by adding and subtracting the same term we can write

$$\begin{aligned}
&\int_M \langle \Pi^{-1}_{t,\gamma_{k,\tau}} \left(\nabla_M f(T_t^{\tau,k}(x), t) \right), \frac{\nabla \phi^c_{k,k+1}(x)}{\tau} \rangle_x d\mu^\tau_{k+1}(x) \\
&\quad = \underbrace{\int_M \langle \Pi^{-1}_{t,\gamma_{k,\tau}} \left(\nabla_M f(T_t^{\tau,k}(x), t) \right) - \nabla_M f(x,t), \frac{\nabla \phi^c_{k,k+1}(x)}{\tau} \rangle_x d\mu^\tau_{k+1}(x)}_{A} \\
&\quad + \underbrace{\int_M \langle \nabla_M f(x,t), \frac{\nabla \phi^c_{k,k+1}(x)}{\tau} \rangle_x d\mu^\tau_{k+1}(x)}_{B} .
\end{aligned} \tag{2.26}$$

We start by obtaining uniform (on τ) bounds for A, note that by Cauchy-Schwarz, for every $x \in \mathrm{spt}(\mu^\tau_{k+1})$

$$\begin{aligned}
&\langle \Pi^{-1}_{t,\gamma_{k,\tau}} \left(\nabla_M f(T_t^{\tau,k}(x), t) \right) - \nabla_M f(x,t), \frac{\nabla \phi^c_{k,k+1}(x)}{\tau} \rangle_x \\
&\leq |\Pi^{-1}_{t,\gamma_{k,\tau}} \left(\nabla_M f(T_t^{\tau,k}(x), t) \right) - \nabla_M f(x,t)|_x \left| \frac{\nabla \phi^c_{k,k+1}(x)}{\tau} \right|_x .
\end{aligned}$$

But the time derivative of $\Pi^{-1}_{t,\gamma_{k,\tau}}$ along integral curves gives the covariant derivative along $\dot{\gamma_{k,\tau}}$, so by Mean Value Theorem and Proposition 182,

$$\begin{aligned}
&\langle \Pi^{-1}_{t,\gamma_{k,\tau}} \left(\nabla_M f(T_t^{\tau,k}(x), t) \right) - \nabla_M f(x,t), \frac{\nabla \phi^c_{k,k+1}(x)}{\tau} \rangle_x \\
&\leq \sup_{\xi \in M} |\nabla_{\dot{\gamma}_{k,\tau}} (\nabla_M f(\xi, t))|_\xi ((k+1)\tau - t) \left| \frac{\nabla \phi^c_{k,k+1}(x)}{\tau} \right|_x \\
&\leq \sup_{s \in [0,T]} \sup_{\xi \in M} |\mathrm{Hess}\, f_s|_\xi |\dot{\gamma}_{\tau,k}(\xi)|_\xi ((k+1)\tau - t) \left| \frac{\nabla \phi^c_{k,k+1}(x)}{\tau} \right|_x \\
&\leq ((k+1)\tau - t) L^2 \sup_{s \in [0,T]} \sup_{\xi \in M} |\mathrm{Hess}\, f_s|_\xi ,
\end{aligned}$$

where $f_t(\cdot) = f(\cdot, t)$ and in the last bound we used the fact that $\gamma_{k\tau}$ is a geodesic so the norm of it's tangent vector is constant, therefore bounded by Proposition 182.

With this bound in hand, going back to (2.26) we find that

$$\sum_{k=0}^{N_\tau}\int_{k\tau}^{(k+1)\tau}\int_M \langle \Pi^{-1}_{t,\gamma_{k,\tau}}\left(\nabla_M f(T_t^{\tau,k}(x),t)\right)-\nabla_M f(x,t),\frac{\nabla\phi^c_{k,k+1}(x)}{\tau}\rangle_x d\mu^\tau_{k+1}(x)dt$$

$$\leq \sup_{t\in[0,T]}\sup_{\xi\in M}|\mathrm{Hess}\, f_t|_\xi\, L^2\sum_{k=0}^{N_\tau}\int_{k\tau}^{(k+1)\tau}((k+1)\tau-t)dt$$

$$\leq \sup_{t\in[0,T]}\sup_{\xi\in M}|\mathrm{Hess}\, f_t|_\xi\, L^2\left(\sum_{k=0}^{N_\tau}(k+1)\tau^2-k\tau^2-\frac{\tau^2}{2}\right)$$

$$=\frac{1}{2}\sup_{t\in[0,T]}\sup_{\xi\in M}|\mathrm{Hess}\, f_t|_\xi\, L^2\,(N_\tau+1)\tau^2=C\,(N_\tau+1)\tau^2.$$

Because $N_\tau\tau\to T$ as $\tau\to 0$, this upper bound goes to 0 as $\tau\to 0$, meaning that A vanishes in the $\tau\to 0$ limit.
To study B from (2.26) note that using Lemma 181

$$\int_M\langle\nabla_M f(x,t),\frac{\nabla\phi^c_{k,k+1}(x)}{\tau}\rangle_x d\mu^\tau_{k+1}(x)=\int_M\langle\nabla_M f(x,t),\nabla(W*\mu^\tau_{k+1})\rangle_x d\mu^\tau_{k+1}(x).$$

Hence our goal to finish the proof is to show that

$$\sum_{k=0}^{N_\tau}\int_{k\tau}^{(k+1)\tau}\int_M\langle\nabla_M f(x,s),\nabla_M(W*\mu^\tau_{k+1}\rangle_x d\mu^\tau_{k+1}ds$$

$$\xrightarrow{\tau\to 0}\int_0^T\int_M\langle\nabla_M f(x,t),\nabla_M(W*\mu(s))\rangle_x d\mu(s)ds. \tag{2.27}$$

By density in the space of smooth functions we may assume without loss of generality that $f(x,t)=\phi(x)a(t)$ from which we can change the order of integration in the left handside of (2.27) and by mean value theorem for integrals to rewrite

$$\sum_{k=0}^{N_\tau}\int_{k\tau}^{(k+1)\tau}\int_M\langle\nabla_M f(x,s),\nabla_M(W*\mu^\tau_{k+1})\rangle_x d\mu^\tau_{k+1}ds$$

$$=\sum_{k=0}^{N_\tau}\tau a(t_k^*)\int_M\langle\nabla_M\phi,\nabla_M(W*\mu^\tau_{k+1})\rangle_x d\mu^\tau_{k+1}, \tag{2.28}$$

where $t_k^*\in[k\tau),(k+1)\tau]$. Observe that this limit concludes the proof as the extra term in (2.25) vanishes in the $\tau\to 0$ limit.
By Riemann integrability together with (2.28), to establish the limit (2.27) it is enough to show that

$$\sum_{k=0}^{N_\tau}\tau a(t_k^*)\left(\int_M\langle\nabla_M\phi,\nabla_M(W*\mu^\tau_{k+1})\rangle_x d\mu^\tau_{k+1}-\right.$$

$$\left.\int_M\langle\nabla_M\phi,\nabla_M(W*\mu(t_k^*))\rangle_x d\mu(t_k^*)\right)\to 0. \tag{2.29}$$

Observe that by Fubini's theorem, the term in parenthesis can be recast as

$$\begin{aligned}
&\int_M \langle \nabla_M \phi, \nabla_M (W * \mu_{k+1}^\tau) \rangle_x d\mu_{k+1}^\tau - \int_M \langle \nabla_M \phi, \nabla_M (W * \mu(t_k^*)) \rangle_x d\mu(t_k^*) \\
&= \int_M \langle \nabla_M \phi, \nabla_M W(x,y) \rangle_x d(\mu_{k+1}^\tau \otimes \mu_{k+1}^\tau) - (\mu_{t_k^*} \otimes \mu_{t_k^*})(x,y) \\
&\leq d_1(\mu_{k+1}^\tau \otimes \mu_{k+1}^\tau, \mu_{t_k^*} \otimes \mu_{t_k^*}) \leq d_1(\mu_{k+1}^\tau, \mu_{t_k^*})) \leq d_2(\mu_{k+1}^\tau, \mu_{t_k^*}),
\end{aligned}$$

where the first inequality follows from continuity of the integrand and the fact that $M \times M$ is compact together with the Kantorovich-Rubenstein Theorem (see [Cordero-Erausquin-McCann-Schmuckenschläger]), the second inequality by Lemma 184. Coming back to showing (2.27), we get

$$\begin{aligned}
&\sum_{k=0}^{N_\tau} a(t_k^*) \left(\int_M \langle \nabla_M \phi, \nabla_M (W * \mu_{k+1}^\tau) \rangle_x d\mu_{k+1}^\tau - \right. \\
&\qquad\qquad \left. \int_M \langle \nabla_M \phi, \nabla_M (W * \mu(t_k^*)) \rangle_x d\mu(t_k^*) \right) \\
&\leq C_1 ||a||_\infty \tau (N_\tau + 1) \sup_{t \in [0,T]} d_2(\mu_{k+1}^\tau, \mu(t)) \\
&\leq C_2 \tau (N_\tau + 1) \sup_{t \in [0,T]} \{ d_2(\mathbf{Geo}_\tau(\{\mu_k^\tau\})(t), \mu(t))) + d_2\, \mathbf{Geo}_\tau(\{\mu_k^\tau\})(t), \mu_{k+1}^\tau) \},
\end{aligned}$$

where C_1, C_2 are positive constants, which shows that the sum in (2.27) converges in the limit because $\tau N_\tau \to 1$ together with the uniform limit from equation (2.13) and (167). ∎

2.4 Conclusions and extensions

This work proved the existence of small time solutions for (2.1) via the minimizing movement scheme under suitable conditions on the interaction potential. The assumption of dependence on Riemannian distance make it completely intrinsic and suitably general. A first line of investigation could be to derive long-time existence and geometry of solutions of the model for specific interaction potentials like power laws, for example. In this context, is it possible to reproduce the aggregation results from [McCann-Davies-Lim] in curved geometries?

Another interesting extension is the performance of numeric algorithms based on entropic optimal transportation compared to the usual PDE approximation methods.

The idea of using the minimizing movement scheme motivated from the seminal work [Ambrosio-Gigli-Savare] required an analysis of the optimality condition as there was no global λ-convexity of the functional. One way to avoid this problem is to restrict the geometry of the manifold to satisfy a non-negative cross-curvature condition which allows the set of optimal transport maps to be convex yielding λ-convexity. This approach enables the machinery of [Ambrosio-Gigli-Savare] which not only ensures existence and uniqueness but provides error bounds on the discrete approximation.

This work has shown that techniques from non-smooth analysis allow the Euler-Lagrange equation to imply enough regularity to solve the aggregation equation in small times. Further work should concentrate on the several possibilities beyond the specified time T. After time T, it is not clear if one can use the minimizing movement scheme to obtain a relevant flow. The problem relies on the fact that the cut locus does not allow the potential to indicate a unique trajectory. The existence of

multiple geodesics stops us from using the geodesic interpolation, which as seen throughout this work, is the most natural way to interpolate for the minimizing movement scheme. At time T it is not clear whether the measures concentrate and a solution for the flow still exists or if it doesn't move after this time. The characterization of these possibilities is an open question in Riemannian manifolds, resolved for power laws in Euclidean setting in [McCann-Davies-Lim].

Chapter 3

Measure pre-conditioning in Machine-Learning

3.1 Introduction

Recent progress in the use of optimal transportation techniques for machine learning in domain adaptation [Courty-Flamary] and development of Wasserstein Generative adversarial networks [Arjovsky et. al] have helped our understanding of potential learning derived from theoretic properties of the underlying data. The topic of optimal transportation has grown significantly in recent years (see [Villani2003], [Villani2009], [Santambrogio] and references therein).
Machine learning models aim to solve a task (to prescribed accuracy) using only the information of known data (training set). In this context it is preferred to have non-parametric models over parametric statistical families.
In this document we explore an idea that we call **measure pre-conditioning** the training data which consists in modifying the statistical model in order to improve performance of algorithms while preserving the limiting model. One can argue that measure pre-conditioning implicitly imposes unjustified structure to a problem but the idea is that measure pre-conditioning will simplify computations and ensure convergence to the original model. For example measure pre-conditioning one of the measures may allow using optimal transportation techniques to adapt a domain which would otherwise be very costly, this would yield a desired training in a task with little information. We use the terminology "measure pre-conditioning" as the technique reminds us of pre-conditioning matrices from linear algebra and optimization.

3.1.1 Organization of this document

3.1.2 Relation to literature

The authots of [Courty-Flamary] develop the idea of optimal transport domain adaptation on which a linear approximation of the transport map is used to infer labels on target domain and [Courty et al.] developed CO-OT, a technique on which optimal transport is not only done between source and target domains of data but in the space of data and labels.

Recently [Amos-Cohen-Luise-Redko] developed the META-optimal transport technique which by pre-solving an optimization problem improves on the optimal transport efficiency. In this work, the idea is similar: can we modify training sets to ensure properties of learning? The modifications considered in this document, differ from the ones on [Amos-Cohen-Luise-Redko] as we only consider measure pre-conditioning data without establishing a minimization purpose beforehand. These techniques should remind the reader of the concept of preconditioning in optimization, on which one modifies a matrix via a correct scaling to benefit the algorithm computations. In the same fashion, here one modifies the measure associated to a training set to benefit statistical properties of the learning agent.

3.1.3 Necessity of non-parametric measure pre-conditioning techniques

The need for non-parametric measure pre-conditioning techniques arises from the modeller's attempt to not intervene in the learning while improving it's computational performance. Measure pre-conditioning is posed in this document as a general technique and it is the modeller's task to determine which pre-conditioning is useful for their own goal. In section 3.4.1 we give several examples with different goals in mind.

3.2 Measure pre-conditionings

In this section we introduce the main concept and discuss several possible "measure pre-conditionings". In this context a measure pre-conditioning will be a technique to manipulate data in order to obtain a "nicer" measure. For example, we can regularize our problem to obtain a measure that is absolutely continuous with respect to Lebesgue or a measure that has a different type of support.
Measure pre-conditioning is also similar to parameter fitting for curves. In the case of real variable one attempts to infer information from isolated data points by first creating a continuous (typically smooth) curve joining the points. Pre-conditioning between points in $\mathbb{R}$ has drawbacks (overfitting, high-variation, etc) and so will measure pre-conditioning (see section 3.7). Measure pre-conditionining will have the advantage of enabling stronger techniques to infer learning as we will see throughout the paper. We start by defining several possible measure pre-conditioning techniques and analyzing their properties.

Problem 1. *(General measure pre-conditioning problem for independent identically distributed data)*
Let $(X_1, X_2, \dots, X_n)$ be a sample, that is $\{X_i\}_{i=1}^n$ is a set of independent identically distributed data such that $X_1 \sim \mu$. Suppose that the sample will be used to train a machine learning model, the measure pre-conditioning problem is to find a good way to obtain a measure $\tilde{\mu}_n$ from the sample such that $\tilde{\mu}_n$ improves performance of the model or the computational cost of the algorithms while keeping the most relevant features of the problem intact.

As such, this measure pre-conditioning problem is not mathematically well posed, as we haven't defined what "improves performance of the model" or " keeping the most relevant features" mean. Performance improvement can be done in several ways: simplification of algorithms, computational cost, control on domain adaptation or even yielding mathematical properties for the learning agent. All of these type of improvements are valid and impactful in machine learning research. The aim of this paper is to analyze how different measure pre-conditionings impact model performance.

3.3 A mathematical framework admitting pre-conditioning

Let us start with a basic framework from Machine Learning models in order to be able to define measure-preconditioning and show it's relevance. The simplest case is the minimization over all fitting functions f within a class of fitters $\mathcal{C}$ minimizing the expected value of the loss function L measuring the loss of fitting the random variable Y with the variable X via $f(X)$.

3.3.1 Formulation of the problem

Problem 2. *Let $\Omega \subseteq \mathbb{R}^n$ be convex and compact. Assume we have data $X \sim \mu \in \mathcal{P}(\Omega)$ and we aim to do a Machine-learning model towards a dependent variable $Y \in \mathcal{Y}$ where $(\mathcal{Y}, d_Y)$ is a separable complete metric space, we denote by $\pi \in \mathcal{P}(\Omega \times \mathcal{Y})$ the joint distribution of (X, Y). Given $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (called a loss function), let $\mathcal{C} \subseteq \mathcal{Y}^{\Omega}$ and assume d is a distance function on $\mathcal{C}$, the $\mathcal{C}-$ optimal model for L under π is the following non-linear program*

$$\arg\min_{f \in \mathcal{C}} \mathbf{E}_{\pi} \left[L(f(x), y) \right]. \tag{3.1}$$

Now assume we don't know the full model π but we have a training sample, i.e. we have $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \pi$, statistically we know the values on the sample but not the full distribution. Assume we approximate π using the sample via a probability measure π_n, the associated $\mathcal{C}-$model for L under π_n reads

$$\arg\min_{f \in \mathcal{C}} \mathbf{E}_{\pi_n} \left[L(f(x), y) \right]. \tag{3.2}$$

This formulation immediately give rise to the following questions

i If L and $\mathcal{C}$ are fixed, what conditions on π_n ensure that the minimizer in (3.2) approaches (3.1)? In what topology?

ii What properties could (3.2) have that (3.1) may lack?

iii Given a choosing of π_n's, could we find sequences L_n's and $\mathcal{C}_n$ so that the computations on the $\mathcal{C}_n-$ problem with loss function L_n associated to π_n converge to (3.1)? Could these problems improve the algorithmic performance?

Idea 1. ***(Measure pre-conditioning)***
A measure pre-condition is a way to define π_n from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ such that the associated $\mathcal{C}$-problem with loss function L has improved performance in any way while preserving the convergence of minimizers of (3.2) to that of (3.1).

3.3.2 Convergence of the learning problem

Our main focus will be answering: when do minimizers of (3.2) converge to minimizers of 3.1 and in which way?.

We first notice that in many situations it is possible to obtain the same total loss under convergence of the measures (without necessarilly having convergence of minimizers), this situation is rather general and known and is not the main question in the ML community but it gives a good starting point for the techniques used in this document. For many applications it is enough to know convergence of the total loss and so we exemplify conditions that yield such convergence.

Proposition 185. *(Standard convergence results on total loss (not minimizers))*

1. *If* $||L||_\infty < \infty$ *or if* $\text{spt}(\mu)$ *is compact,*
$$|\mathbf{E}_{\pi_n}[L(f(\overline{X}), Y)] - \mathbf{E}_{\pi}[L(f(\overline{X}), Y)]| \leq ||\pi_n - \pi||_{TV}.$$
2. *Given* $f \in \mathcal{C}$ *if* $(x, y) \to L(f(x), y)$ *is Lipschitz, then*
$$|\mathbf{E}_{\pi_n}[L(f(\overline{X}), Y)] - \mathbf{E}_{\pi}[L(f(\overline{X}), Y)]| \leq d_1(\pi_n, \pi).$$
3. *If* L *is* $\mathcal{C}^2$ *and* $||\frac{\partial L}{\partial 1}|| < \infty$ *then*
$$|\mathbf{E}_{\pi_n}[L(f(\overline{X}), Y)] - \mathbf{E}_{\pi}[L(f(\overline{X}), Y)]| \lesssim ||\pi_n - \pi||_{TV} + \sup_{x \in \Omega} d(f^*(x), f_n^*(x))$$
4. *If* $\mathcal{C}$ *is a compact class on* $C(\mathcal{Y})$*, and* $\pi_n \to \pi$ *in* d_1 *then along a subsequence* n_k
$$\mathbf{E}_{\pi_{n_k}}[L(f_{n_k}^*(X), Y)] \to \mathbf{E}_{\pi}[L(f^*(X), Y)]$$
where $f_{n_k}^*$ *is the* $\mathcal{C}-$*optimizing argument for* π_n *and* f^* *is the* $\mathcal{C}-$*optimizing argument for* π*.*

The proof of proposition 185 is direct and hence omitted.

3.3.3 The main question

Measure preconditioning modifies the minimization problem at level n, i.e. it changes the structure of the measure used to evaluate loss with a sample of size n. If the model was unchanged we would expect convergence of the learning agent trained with the sample of size n, i.e. f_n^* to the best fit with respect to the loss for the parametric distribution f^*. If measure pre-conditioning modifies the measure at level n, the true question is when and in which ways does $f_n^* \to f^*$?.
To answer the convergence of minimizers, as it is usual in functional analysis and economics, we introduce Γ-convergence.

Main Theorem

We present an informal version of the main theorem of the work. This informal version corresponds to the rigorous statements answered in Theorem 193, Proposition 195 and section 3.5

Theorem 186. *Full learner recovery system is a concept that allows us to show convergence of learning agents to the ideal parametric agent in cases not covered previously in the literature. This concept allows us to generalize stability arguments for less regular losses and a bigger class of classification/regression problems. Full learner recovery systems are general enough to be applied to several settings in Machine-Learning, including Domain Adaptation transfer learning. These systems explain many phenomena in ML-research where convergence is improved. Full learner recovery systems give a guideline on how and when to modify training data without disturbing the original problem.*

The formulation of Theorem 186 is not mathematically precise, we dedicate this work to make the Theorem rigorous and prove it in the subsequent sections.
We start with the introduction of the main mathematical tool.

3.3.4 A version of the envelope Theorem

Definition 187. *(Γ-convergence on a metric space)*
Let (X,d) be a metric space and let $F_j, F : X \to \mathbb{R} \cup \{\pm\infty\}$,we say F_n Γ-converges to F, denoted $F_n \xrightarrow{\Gamma} F$ if and only if the following two conditions hold

I For all sequences $\{x_j\}$ such that $x_j \xrightarrow{d} x$ we have

$$\liminf_{j\to\infty} F_j(x_j) \geq F(x).$$

II For every $x \in X$ there exists a sequence $x_j \xrightarrow{d} x$ such that

$$F(x) \geq \limsup_{j\to\infty} F_j(x_j).$$

Remark 188. *The most general definition for Γ-convergence is one where X is assumed to be a topological space and not necessarily metric. The definition presented above (Definition 187) is the sequential-definition. We have chosen the sequential definition as it simplifies the theory significantly, knowing that some important examples that we have in mind are only topological spaces on which the Γ-limit is defined via*

$$\Gamma - \lim_{n\to\infty} F_n(x) = \sup_{U\in N(x)} \liminf_{n\to\infty} \inf_{y\in U} f_n(y). \tag{3.3}$$

In some of the examples below the underlying convergence will not correspond to a metric space, on which one must think of (3.3) instead of (I) and (II).

The motivation behind the definition of Γ-convergence is that minimizers converge to minimizers, the content of the following theorem from [Braides]:

Theorem 189. *(Γ-convergence and minima)*
Let (X,d) and F_j, F be as in Definition 187, then

1. *If I from definition 187 is satisfied for all $x \in X$ and K is a compact subset of X then*

$$\inf_K F \leq \liminf_{j\to\infty} \inf_K F_j \tag{3.4}$$

2. *Similarly, if II from definition 187 is satisfied and U is an open subset of X then*

$$\limsup_{j\to\infty} \inf_U F_j \leq \inf_U F \tag{3.5}$$

This Theorem can be found as [Braides, Proposition 1.18]. Finally we recall one more Theorem from [Braides]. We say that a sequence $\{F_j\}$ of functions on a metric space (X,d) is equi-mildly coercive if there exists a non-empty compact set K such that

$$\inf_X F_j = \inf_K F_k \text{ for all } j.$$

Theorem 190. *(Minimizers and Γ-limits)*
In a metric space (X, d) if $\{F_j\}$ is equi-mildly coercive and $F_n \xrightarrow{\Gamma} F$ then

$$\min_X F = \lim_{j\to\infty} \inf_K F_j \tag{3.6}$$

Furthermore, every limit point of a sequence of minimizers of (3.6) is a minimizer of F.

For a proof see [Braides, Theorem 1.21]. With the theory in hand we take a general approach to answer the questions (i) and (ii). Instead of a constructive proof to find the optimal topologies (on $\mathcal{C}$ and $\mathcal{P}(X \times Y)$) we reformulate the convergence problem for it to satisfy the hypothesis of Theorem 187. This way we can relate to classical problems by looking at the given topologies of each framework and verifying the hypothesis.
Going back to the framework of Problems (3.1) and 3.2, we want to be able to recover minimizers from our measure conditioning. We note the interaction of the class of fitters $\mathcal{C}$, the loss function L and the mode of convergence of the conditioners that we choose to evaluate, henceforth it is logical to check conditions for them as a collective, rather than separately. This motivates the following definition.

Definition 191. *(Full learner recovery system)*
In the context of Problem 2, we say that $(\mathcal{C}, d, L, \xrightarrow{m})$ forms a full learner recovery system if it holds that

1. *If $\pi_n \xrightarrow{m} \pi$ for all d-converging sequences sequences $f_n \xrightarrow{d} f$, we have*

$$\liminf_{n\to\infty} \mathbf{E}_{\pi_n}[L(f_n(X), Y)] \geq \mathbf{E}_{\pi}[L(f(X), Y)]. \tag{3.7}$$

2. *If $\pi_j \xrightarrow{m} \pi$ and for every $f \in \mathcal{C}$ there exists a sequence $f_j \in \mathcal{C}$, such that $f_j \xrightarrow{d} f$ and*

$$\mathbf{E}_{\pi}[L(f(X), Y)] \geq \limsup_{j\to\infty} \mathbf{E}_{\pi_j}[L(f_j(X), Y)] \tag{3.8}$$

Remark 192. *In analytical terms, these conditions ensure 2-sided Fatou-Lemmas for integration with respect to L on the first coordinate.*

Γ-convergence can be also used to address the existence of minimizers of the parametric model but that is not the approach of this work, we assume existence of minimizers of the limiting problem and study recovery sequences, from now on we assume the existence of a unique minimizers for (3.2).

Theorem 193. *If $(\mathcal{C}, d, L, \xrightarrow{m})$ forms a full learner recovery system (Definition 191) where (C, d) is a compact metric space, assume the limiting problem from 3.1 has a solution $f \in \mathcal{C}$, then there exists a sub-sequence $\{f_{n_k}\}$ of $\{f_n\} \in \mathcal{C}$ such that*

$$f_{n_k} \in \arg\min_{f\in C} \mathbf{E}_{\pi_{n_k}}[L(f(X), Y)]$$

such that as $k \to \infty$, $f_{n_k} \xrightarrow{d} f$ and

$$\mathbf{E}_{\pi_{n_k}}[L(f_{n_k}(X), Y)] \to \mathbf{E}_{\pi}[L(f(X), Y))]. \tag{3.9}$$

Proof. The definition of $(\mathcal{C}, d, L, \xrightarrow{m})$ forming a full learner recovery system is such that $\mathbf{E}^L_{\pi_n} \xrightarrow{\Gamma} \mathbf{E}^L_{\pi}$, i.e. by taking the functional $F_n(f) : \mathcal{C} \to \mathbb{R}$, defined via

$$F_n(f) := \mathbf{E}_{\pi_n}[L(f(X), Y)],$$

the definition 191 is equivalent to $F_n \xrightarrow{\Gamma} F$. By compactness of $\mathcal{C}$ we get the hypothesis for Theorem 190 so we get the thesis. ∎

In many cases $\mathcal{C}$ is not necessarily compact. The assumption of compactness simplifies the arguments but the argument above can be obtained without compactness of $\mathcal{C}$ if instead one assumes equi-mild-coercivity of $\{F_n\}$, that is there exists a compact set K for which all F_n's satisfy $\inf_C F_n = \inf_K F_n$. See [Braides, Theorem 1.21], we instead assume compactness of $\mathcal{C}$ to avoid this subtlety.

Remark 194. *Evidently the statement of Theorem 193 is useless unless we explore examples and explain the ideas and how to use it. So far, we have just re-written the problem so that we can conclude (subsequential) convergence of learned agents by checking a modified version of Fatou's Lemma. This rewriting allows us to cover different cases at the same time, as we do in the following examples.*

The goal of this list is not to be exhaustive but to show the many different formulations that can be included in Definition 191. Notice that checking Definition 191 involves only studying a two sided version of Fatou's Lemma that can be corroborated in every particular case. Once one establishes that the given ML problem of the form (3.2) and (3.1) are indeed a full recovery system with $\{\pi_n\}, \pi, \mathcal{C}, L$ one has ensured convergence of minimizers (which amounts to perfect approximation of the model).
In the following proposition we show the wide range of options one has for full recovery systems, although the d-convergence in some items of the following proposition are not necessarily with respect to a metric, we have in mind Remark 188.

Proposition 195. *The following are full learner recovery systems*

1. *Let $K \subset \mathbb{R}^p$ be compact, $\mathcal{C}$ a compact subset of $\{f : \mathbb{R}^p \to \mathbb{R}$ s.t. $(x, y) \to L(f(x), y) \in L^1(\pi)\}$ with respect to d, where d denotes point-wise convergence, $L : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ be any positive, bounded, continuous function and let $\xrightarrow{m}$ denote set-wise convergence i.e $\mu_n(A) \to \mu(A)$ for every Borel set A, where $\mu_n, \mu \in \mathcal{P}(K)$.*

2. *$\xrightarrow{m} := \rightharpoonup$ (weak convergence of measures), $\mathcal{C}$ be compact such that $\{L(f(x), y)\}_{f \in \mathcal{C}}$ uniformly integrable with respect to $\{\pi_n\}$ and there exists g such that $L(g(x), y) \in L^1_\pi$ such that $f_n(x) \leq g(x)$ holds π-a.e.*

3. *$\xrightarrow{m} := \xrightarrow{d_1}$, d point-wise convergence and $(x, y) \to L(f(x), y)$ uniformly Lipschitz and uniformly bounded for $f \in \mathcal{C}$ (compact metric space).*

4. *$\xrightarrow{m} := \xrightarrow{TV}$, $L(x, y)$ is d-continuous on the first coordinate and uniformly bounded by some constant $M > 0$ on a compact metric space $(\mathcal{C}, d)$.*

Proof. In all of the cases above we only need to ensure a Fatou-like lemma (Definition 191).

1. Here Γ-convergence must be thought as in Remark 188. This is a direct consequence of Fatou's lemma for varying measures (found in [Royden] or [Feinberg-Kasyanov-Liang, Theorem 1.1]).

2. See [Feinberg-Kasyanov-Liang, Theorem 2.2].

3. The uniform Lipschitz condition gives

$$\begin{aligned}&\left|\int L(f_j(x),y)d\pi_n - d\pi(x) + \left|\int L(f_j(x),y) - L(f(x),y)d\pi(x,y)\right|\right.\\&\leq d_1(\pi_n,\pi) + \left|\int L(f_j(x),y) - L(f(x),y)d\pi(x,y)\right|\end{aligned}$$

   where the first term comes from Kantorovich-Rubinstein [Villani2009, Particular Case 5.16] and the second one vanishes by dominated convergece.

4. In this case we don't only have the inequalities of definition 191 but the limits coincide:

$$\begin{aligned}&\left|\int L(f_n(x),y)d\pi_n(x,y) - \int L(f(x),y)d\pi(x,y)\right|\\&\qquad\leq \left|\int L(f_n(x),y)d(\pi_n-\pi)\right| + \left|\int L(f_n(x),y) - L(f(x),y)d\pi\right|\\&\qquad\leq M||\pi_n-\pi||_{TV} + \left|\int L(f_n(x),y) - L(f(x),y)d\pi\right|\end{aligned}$$

   where the first one goes to zero by the assumption $\pi_n \xrightarrow{TV} \pi$ and the second one by the assumed d-continuity and dominated convergence.

∎

The goal of this list is not to be exhaustive but to show the many different formulations that can be included in Definition 191. Notice that checking Definition 191 involves only studying a two sided version of Fatou's Lemma that can be corroborated in every particular case. Once one establishes that the given ML problem of the form (3.2) and (3.1) are indeed a full recovery system with $\{\pi_n\},\pi,\mathcal{C},L$ one has ensured convergence of minimizers (which amounts to perfect approximation of the model).

Remark 196. *Observe that the conditions imposed for $\mathcal{C}$ and L on Proposition 195 case 4 are less restrictive than the ones on 195 case 2. This is intuitively obvious as the total variation convergence is stronger than weak convergence. This means that ensuring a stronger convergence in measure is a degree of improvement for the ML-problem associated to fixed $\mathcal{C}$ and L. It is also evident that regularity conditions usually assumed in ML-theory (like Lipschitz properties of L) yield strong approximations in most types of convergence $\xrightarrow{m}$, making this framework not only inclusive but rather general.*

One of the main advantages of measure pre-conditioning is the ability to change the training sample. It is common to use the empirical measure in non-parametric statistics, nevertheless the next section shows that the empirical measure is in general, not the best formulation for (3.2) as it may happen that the conditions for convergence hold for a different sequence of measures and not

the sequence of empirical measures. We will see this is the case of Proposition 195 case 4, where the sequence of empirical measures would not ensure subsequential convergence but a different sequence does, justifying completely the use of measure pre-conditioning as it improves the likelihood that the algorithm gives a reasonable final learnt agent.

Remark 197. *(Compactness)*
Stronger conditions like compactness of the underlying sets yield a more elegant theory. Many of the modes of convergence are equivalent under the assumption on compactness (see [Billingsley] or [Villani2003, Chapter 7]). The assumption of compactness simplifyies most theorems as it will automatically bound sequences and so Definition 191 is much easier to satisfy and verify which automatically yields:

Proposition 198. *If $\mathcal{C} \subseteq C(Y)$, $(x, y) \to L(x, y)$ is continuous and*

$$\sup_{f \in C} \sup_{(x,y)} |L(f(x), y)| < \infty$$

then (3.2) $\to$ (3.1) in the $\mathcal{C}$ uniform topology, i.e.

$$\arg\min_{f \in \mathcal{C}} \mathbf{E}_\pi \left[L(f(x), y) \right] \xrightarrow{\mathcal{C}} \arg\min_{f \in \mathcal{C}} \mathbf{E}_\pi \left[L(f(x), y) \right].$$

No Empirical Probability Measure can Converge in the Total Variation Sense for all Distributions

Towards studying when to measure pre-condition we realize that it is important to know what types of empirical measures converge and in which cases. In the seminal work [Devroye-Giorfi], the authors proved the following theorem:

Theorem 199. *(No Empirical Probability Measure can Converge in the Total Variation Sense for all Distributions)*
Let $\{\pi_n\}$ be a sequence of empirical distributions and $\delta > 0$, then there exists a proability measure π such that

$$\inf_n \sup_A |\pi_n(A) - \pi(A)| > \frac{1}{2} - \delta \ \ a.s.$$

For a proof see [Devroye-Giorfi].
Theorem 199 tells us that the class of measures approximated in total variation norm by the empirical measure is not all measures. For different measures, other probability measures formed from data can converge in total variation but the empirical measure does not converge to all measures.

Remark 200. *In [Devroye-Giorfi] it is shown that the standard empirical measure does not converge in total variation sense to absolutely continuous limits. Hence, Theorem 193 does not apply with Proposition 195 case 4 if we use the standard empirical measure. Nevertheless, as shown in [Devroye], the kernel-empirical measure given by*

$$\pi_n = \frac{1}{hn} \sum K(f/h)$$

does converge in total variation (see Definition 203 below). Hence, Theorem 193 via Proposition 195 case 4 applies to the sequence $\{\pi_n\}$ but not the sequence of standard empirical measures. This shows that the model solution for ML-program 3.2 will converge to the best parametric $\mathcal{C}$-model. This argumentation ***explains*** *why* ***standard techniques in Machine-Learning****, such as shifting and adding noise give* ***better results in practice****, as convergence is ensured by this system.*

Example: Linear regression

Let us consider $\pi \in \mathcal{P}_{ac}(\mathbb{R}^2)$, we consider the linear regression problem with square-loss function with respect to target measure π:

$$\min_{(a,b)\in\mathbb{R}^2} \mathbf{E}_{\pi}[(Y - aX + b)^2]. \tag{TargetLR}$$

By differentiating with respect to a, b from first order conditions we know that the solutions to (TargetLR) are

$$a = \frac{\int y \cdot x d\pi(x,y) - \int y d\pi(x,y) \int x d\pi(x,y)}{\int x^2 d\pi(x,y) - \left(\int x d\pi(x,y)\right)^2} \tag{3.10}$$

$$b = \int y d\pi(x,y) - \left(\frac{\int y \cdot x d\pi(x,y) - \int y d\pi(x,y) \int x d\pi(x,y)}{\int x^2 d\pi(x,y) - \left(\int x d\pi(x,y)\right)^2} \right) \int x d\pi(x,y) \tag{3.11}$$

If we consider a sequence of measures π_n, obtained using the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ then the linear regression problem with square-loss function with respect to approximating measure π_n is

$$\min_{(a,b)\in\mathbb{R}^2} \mathbf{E}_{\pi_n}[(Y - aX + b)^2]. \tag{AppxLR}$$

The solution (a_{π_n}, b_{π_n}) to (AppxLR) is given by

$$a_{\pi_n} = \frac{\int y \cdot x d\pi_n(x,y) - \int y d\pi_n(x,y) \int x d\pi_n(x,y)}{\int x^2 d\pi_n(x,y) - \left(\int x d\pi_n(x,y)\right)^2} \tag{3.12}$$

$$b_{\pi_n} = \int y d\pi_n(x,y) - \left(\frac{\int y \cdot x d\pi_n(x,y) - \int y d\pi_n(x,y) \int x d\pi_n(x,y)}{\int x^2 d\pi_n(x,y) - \left(\int x d\pi_n(x,y)\right)^2} \right) \int x d\pi_n(x,y) \tag{3.13}$$

If π_n corresponds to the empirical measure, then rate of convergence of a_{π_n} and b_{π_n} have been widely studied. See [Bishop, Chapter 3] for example. We also know by Theorem 199 that $\pi_n \not\xrightarrow{TV} \pi$. By [Devroye, Section 2] we can find a sequence of measures (Parzen windows) $\{\tilde{\pi}_n\}$ such that $\tilde{\pi}_n \xrightarrow{TV} \pi$. For simplicity, assume that

$$\int x d\pi_n(x,y) = 0, \int x^2 d\pi_n(x,y) = 1, \int x d\pi(x,y) = 0 \text{ and } \int x^2 d\pi(x,y) = 1.$$

With this assumption we immediately obtain the following bound:

$$|a_{\pi_n} - a_{\pi}| \leq \left(\sup_{(x,y)\in \text{spt}(\pi_n)\cup\text{spt}(\pi)} |x \cdot y| \right) ||\pi_n - \pi||_{TV}. \tag{3.14}$$

Which in the case where $\{\pi_n\}, \pi$ are uniformly compactly supported yields

$$|a_{\pi_n} - b_{\pi_n}| \lesssim ||\pi_n - \pi||_{TV}. \tag{3.15}$$

Equation (3.15) is a bound on the order of convergence on the coefficient of linear regression of (AppxLR) to that of (TargetLR) which is not available in the case of the empirical measure, as indicated by Theorem 199. The bound (3.15) different to the usual order of convergence bounds for linear regression exemplifies the impact of measure pre-conditioning. Equation (3.15) shows (uniform) stability of learning agents corresponding to the measure pre-conditioned problem, allowing us to use more tools than the standard ones.

3.3.5 Measure pre-conditioning approaches

Measure pre-conditioning approaches impose certain structures to the original data. The idea is to analyze how does this structure impacts final outcomes of the modelling. In some way, this process resembles plain statistical inference.

3.3.6 Background and Notation

Let $\Omega \subset \mathbb{R}^n$ be fixed. We denote by $\mathcal{P}^p(\Omega)$ to be the set of probability measures with p-th finite moment. That is $\mathcal{P}^p(\Omega) = \{\mu \in \mathcal{P}(\Omega) : \int_\Omega |x - x_0|^p d\mu < \infty, \text{ for some } x_0 \in \Omega\}$. We define the Wasserstein p distance between $\mu, \nu \in \mathcal{P}^p(\Omega)$

$$d_p(\mu, \nu) = \left(\inf_{\pi \in \Gamma(\mu,\nu)} \int_{\Omega\times\Omega} |x - y|^p d\pi(x, y) \right)^{1/p}$$

where $\Gamma(\mu, \nu)$ denotes the set of probability measures on $\Omega \times \Omega$ having first marginal μ and second marginal ν. We say a map $T : \Omega_1 \to \Omega_2$ is a Monge map with respect to the cost function $c : \Omega_1 \times \Omega_2 \to \mathbb{R}$, between Borel measures μ and ν whenever

$$T \in \underset{T\#\mu=\nu}{\arg\min} \left\{ \int_{\Omega_1} c(x, T(x)) d\mu(x) \right\} \tag{3.16}$$

where $T\#\mu$ means that for every Borel set A, $\nu(A) = \mu(T^{-1}(A))$.

3.4 Empirical measures and non-parametric estimation

In this section we discuss common non-parametric estimates and their relations to the structure of the ML-problems (3.2) and (3.1). We aim to explain how each measure can be used to pre-condition and the pros and cons coming with their use.

3.4.1 Non-exhausting list of non-parametric estimation techniques

Definition 201. *(Empirical measure)*
Given $X_1, \ldots, X_n$ we define the standard empirical measure as the number of successes on the n occurrences:

$$\mu_n(A) = \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k}(A).$$

Definition 202. *(Histogram)*
Given $X_1, \dots, X_n$ we define the histogram measure associated to the sets $B_1, \dots, B_m$

$$\mu_n(A) = \frac{1}{n} \sum_{k=1}^{n} \sum_{l=1}^{m} \frac{1}{\rho(B_l)} \delta_{X_k}(A \cap B_l).$$

where ρ is a probability measure (usually taken to be normalized Lebesgue).

Definition 203. *(Kernel estimation via Parzen windows)*
Given $X_1, \dots, X_n$, we define the n-th density estimation with kernel K via

$$f_{\pi_n}(x) = \frac{1}{nH_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{H_n}\right)$$

where K is fixed and $\{H_n\}$ is any sequence of random variables, that (may) depend on the sample $X_1, \dots, X_n$ that satisfy that $H_n \to 0$ almost surely and $nH_n \to \infty$ almost surely.

The idea of this formulation of the kernel estimation comes from [Parszen] and [Rosenblatt] and it is fully justified by Theorem 223.

Wasserstein 2-Barycenter

Definition 204. *(Wasserstein Barycenter)*
Given a sample $X_1, X_2, \dots, X_n$ random variables in $\mathbb{R}^p$ we define the 2-Wasserstein Barycenter of the sample (also called Frechet mean) as any probability measure satisfying

$$\mu^* \in \underset{\rho \in \mathcal{P}^2(\mathbb{R}^p)}{\arg\min} \left\{ \sum_{k=1}^{n} d_2(\rho, \delta_{X_k})^2 \right\} \tag{3.17}$$

where δ_{X_k} denotes the unit mass at X_k.

Remark 205. *Note that $\rho \to d_2(\cdot, \nu)^2$ is lower-semicontinuous for all ν and so Wasserstein Barycenters exist. In general, Wassertein barycenters with respect to random Dirac measures are not unique. If instead, one of the deltas is replaced by an absolutely continuous measure, uniqueness can be shown. We don't do this replacement in this document, instead we study the entropic regularization of the minimization problem in Definition 210.*

The theory of Wasserstein Barycenters has recently received attention from several fields of applied mathematics, see for example [Panaretos-Zemel] for a more complete theory.

Remark 206. *The barycenter can be defined given any distance function $d : \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p) \to \mathbb{R}$ and a sample $(X_1, \dots, X_n)$ the d-barycenter is any probability measure μ satisfying*

$$\mu^* \in \underset{\rho}{\arg\min} \left\{ \frac{1}{n} \sum_{k=1}^{n} d(\rho, \delta_{X_k}) \right\} \tag{3.18}$$

where the infimum is taken over all probability measures on $\mathbb{R}^p$.. We have only chosen the Wasserstein 2-distance as we aim to focus on Domain Adaptation.

Remark 207. *It is important to notice that efficient algorithms to compute Wasserstein Barycenters have recently been developed (see [Cuturi-Doucet]) in the case of empirical measures. This efficient computability is essential for the applications we have in mind.*

Uniform convex hull

Definition 208. *(Convex Hull)*
The convex hull of a set $B \subseteq \mathbb{R}^p$ is defined to be the smallest convex set on which B is contained, equivalently

$$\mathrm{Conv}(B) = \bigcap_{\substack{C \text{ convex} \\ B \subseteq C}} C.$$

We define the uniform convex hull of the sample $(X_1, X_2, \dots X_n)$ to be the uniform measure on the convex hull of $\{X_1, X_2, \dots, X_n\}$, i.e.

$$\mu_{conv} = \frac{\mathcal{L}^p \mid_c}{\mathcal{L}^p(\mathrm{Conv}(\{X_1, X_2, \dots, X_n\}))} \tag{3.19}$$

where $\mathcal{L}^p$ denotes the Lebesgue measure in $\mathbb{R}^p$.

Remark 209. *Note that μ_{conv} is the restriction of the Lebesgue measure to the convex hull of the sample so it's support is automatically convex. This particular property could be significant for future applications as the theory of convex optimization unlocks several numerical techniques. Evidently, it's support also includes all points of the sample. Note that Definition 208 always gives a well defined measure.*

Entropically regularized barycenter

Definition 210. *Given a sample $X_1, X_2, \dots, X_n$ random variables in $\mathbb{R}^p$ and a reference probability measure ν we define the ν-entropically regularized 2-Wasserstein Barycenter of the sample as any probability measure satisfying*

$$\mu^* \in \underset{\rho \in \mathcal{P}^2(\mathbb{R}^p)}{\arg\min} \left\{ \frac{1}{n} \sum_{k=1}^{n} d_2(\rho, \delta_{X_k})^2 + \mathrm{Ent}(\rho \mid \nu) \right\} \tag{3.20}$$

where δ_{X_k} denotes the unit mass at X_k and $\mathrm{Ent}(\mu \mid \nu)$ denotes the relative entropy of ρ with respect to ν given by

$$\mathrm{Ent}(\rho \mid \nu) = \int \log\left(\frac{d\rho}{d\nu}\right) d\nu \tag{3.21}$$

whenever $\rho \ll \nu$ and $\mathrm{Ent}(\rho \mid \nu) = \infty$ otherwise.

Remark 211. *If $\nu \ll \mathcal{L}^p$, the functional to minimize is lower semi-continuous and with the addition of entropy a unique absolutely continuous minimizer of (3.20).*

Class-regularized barycenter

Motivated from the work of [Courty-Flamary] we can also think of measure pre-conditioning in terms of pre-established class based groups. The idea behind the next definition is that elements in the same class may be very similar while elements from different classes could be very different from each other.

Definition 212. *(Class barycenter)*
Given a sample $X_1, X_2, \ldots, X_n$ random variables in $\mathbb{R}^p$ suppose that each X_i belongs to one and only one of a finite collection of classes $\{C_l\}_{l=1}^m$, then we can define the class-based barycenter to be any measure μ satisfying

$$\mu^* \in \underset{\mu \in \mathcal{P}^2(\mathbb{R}^p)}{\arg\min} \left\{ \frac{1}{m} \sum_{k=1}^{m} d_2(\rho, \nu_k)^2 + \mathrm{Ent}(\rho \mid \nu) \right\} \tag{3.22}$$

where ν_k is a measure determined only from class C_k. For example, one would obtain a barycenter of barycenters if one were to choose ν_k to be the 2-Wasserstein barycenter of $\{X_i : X_i \in C_k\}$.

MMD-regularized Conditional measures

Definition 213. *Given a characteristic kernel function k (see [Sriperumbudur] for details), define the maximum mean discrepancy between μ, ν with respect to k via*

$$\mathrm{mmd}_k(\mu, \nu) = \mathbf{E}_{\mu \times \mu}[K(X, \tilde{X})] + \mathbf{E}_{\nu \times \nu}[k(Y, Y)] - 2\mathbf{E}_{\mu \times \nu}[k(X, Y)]$$

The empirical optimal transference plan between conditional distributions for a given lower-semicontinuous cost function c, denoted $\pi_n^{*,c}$ is defined in [Manupriya-Keerti-Biswas-Chandhok-Jagarlapudi] via the minimization over $\Gamma(\mu.\nu)$ of the following functional:

$$\int c(x, y) d\pi + \lambda_1 \frac{1}{n} \sum_{i=1}^{n} \mathrm{mmd}_k^2(Proj^1 \# \pi, \delta_{Y_i}) + \sum_{i=1}^{n} \mathrm{mmd}_k^2(Proj^1 \# \pi', \delta_{Y_i'}). \tag{3.23}$$

Existence and uniqueness depends on the cost function and usual conditions (smoothness and twist) are required, see [Sriperumbudur] for details.

3.4.2 Some properties of the measure pre-conditioners

Proposition 214. *When they exist, the measures from definitions 210 and 212 are absolutely continuous with respect to ν.*

Proof. By definition, $Ent(\rho|\nu) = \infty$ if $\rho \not\ll v$, because ν is always feasible, the functional is not infinity and hence the minimizer is a.c. with respect to ν. ∎

Corollary 215. *If $\nu = \mathcal{L}^p$ in Definitions 210 or 212, the minimizer has a density (w.r.t. Lebesgue).*

Although the proof is simple, the importance of Proposition 214 and Corollary 215 is fundamental for practice. If we can estimate the density, we can use it to improve the convergence of algorithms by numerical methods. See for example [Peyre-Cuturi] where the entropic regularization allows a closed (and very simple) form of the density which then yields a dual-descent algorithm. Knowing explicitly the density allows us to find minimizers of Problem 3.2 via formulae and so we can focus our attention on estimating numerically these minimizers without carrying a second numerical error.

3.4.3 Optimality (Euler-Lagrange)

Most of the measure pre-conditioners defined on section 3.3.5 require the minimization of a functional. Let $\Omega \subseteq \mathbb{R}^p$, in this section we study the first order conditions for minimization in $(\mathcal{P}_2(\Omega), d_2)$ which can be found in [Santambrogio, Theorem 7.20].

Definition 216. *(First variation of a functional in $\mathcal{P}(\Omega)$)*
Let F be a functional $F : \mathcal{P}_2(\Omega) \to \mathbb{R}$, let $\rho \in \mathcal{P}_2(\Omega)$ be fixed and $\epsilon > 0$, for any $\tilde{\rho} \in \mathcal{P}^2_{ac} \cap L^\infty(\Omega)$, define $\nu = \tilde{\rho} - \rho$, we say that $\frac{\delta F}{\delta \rho}(\rho)$ is the first variation of F evaluated at ρ if

$$\frac{d}{d\epsilon}\bigg|_{\epsilon=0} F(\rho + \epsilon\nu) = \int \frac{\delta F}{\delta \rho}(\rho) d\nu.$$

Theorem 217. *(Optimality criteria)*
For a functional $F : \mathcal{P}_2(\Omega) \to \mathbb{R}$ suppose that $\mu \in \underset{\nu \in \mathcal{P}_2(\Omega)}{\arg\min} F(\nu)$. Assume that for every $\epsilon > 0$ and for every ρ absolutely continuous with $L^\infty(M)$ density

$$F((1-\epsilon)\mu + \epsilon\rho) < \infty$$

let $c := essinf \left\{ \frac{\delta F}{\delta \rho}(\mu) \right\}$. If $\frac{\delta F}{\delta \rho}(\mu)$ is continuous,

$$\frac{\delta F}{\delta \rho}(\mu)(x) \geq c \ \forall x \in M, \tag{3.24}$$

$$\frac{\delta F}{\delta \rho}(\mu)(x) = c \ \forall x \in supp(\mu). \tag{3.25}$$

The proof can be found as Theorem 7.20 in [Santambrogio].
Just as in the remark after Corollary 215, the main use of this tool is to focus the algorithmic implementation towards the computation of the first variation of the functional it minimizes.

3.4.4 Convergence

The objective of the reformulation of the general ML-problem in terms of Problem 3.2 and 3.1 is that we can adapt every stage of the learning process by using a measure estimation that fits the problem better. In order for us to know that we can recover the ML-problem in this process we need to know the types of convergence on which the sequences of measures formulated with the data converge to the underlying distribution. Many theorems and specific cases on density estimation have been studied, we recollect some of them here in terms of the definitions of section 3.4.1.

Convergence of density estimations

Observe that Theorem 191 and Proposition 195 allow different systems of convergence, i.e. depending on the 'strength' of the type of convergence $\xrightarrow{m}$ of the probability measures, different requirements on $\mathcal{C}, d, L$ are needed. In this section we give a non-exhaustive list of modes of convergence for density estimation and the sequences in Section 3.4.1 that can be used as measure pre-conditioners. In this section one should notice that every type of convergence should be coupled with hypothesis that ensure the system is a full learner recovery system (Definition 191).

Theorem 218. *(Glivenko Cantelli in $\mathbb{R}$)*
Let μ be any probability measure on $\mathbb{R}$ and μ_n be the standard empirical measure (Definition 201), if $F(t) = \mu((-\infty, t])$ and $F_n(t) = \mu_n((-\infty, t])$ then $F_n \to F$ uniformly on $\mathbb{R}$ as $n \to \infty$

This theorem is well-known see for example [Durrett, Theorem 7.4] or [Dudley2002, Theormem 11.4.2.]. By account's of Donsker's theorem one can get the following improvement:

Proposition 219. *(Rate of convergence for continuous F)*
If μ is a law on $\mathbb{R}$ for which F is continuous, the order of convergence of Theorem 218 satisfies

$$n^{1/2} \sup_t |F_n(t) - F(t)| \rightharpoonup \max_{0 \le s \le 1} |B_s - sB_1| \tag{3.26}$$

where $\{B_s\}$ is a Brownian motion, i.e. the rate of convergence approaches the law of the absolute value of a Brownian bridge on $[0, 1]$ and so it's law can be computed explicitly:

$$P_0\left(\sup_{0 \le s \le 1} |B_s - sB_1| < b\right) = \sum_{m=-\infty}^{\infty} (-1)^m e^{-2m^2 b^2} \tag{3.27}$$

See [Durrett, Theorem 8.10] and the following proposition for the explicit formula of it's law.

Remark 220. *The theorem presented here as Theorem 218 is just a specific version. In general, one refers to any theorem of this type as "a Glivenko-Cantelli type theorem" see for example [Dudley2002].*

Theorem 221. *(Varadarajan)*
If π is any probability measure on $X \times Y$ and $X \times Y$ is a separable metric space then the standard empirical measures (Definition 201) for (X, Y) converge weakly in probability to π.

For a proof see [Dudley2002, 11.4.1]. It is important to notice that the convergence is almost surely. In some cases, like the case of real numbers, the convergence can be upgraded.

Remark 222. *Notice that from Theorem 218 one can infer the convergence of the Histogram (Definition 202) weakly in $\mathbb{R}^p$.*

Theorem 223. *(Devroye)*
If $H_n^2 n \to \infty$ and $\mu \ll Leb$, the empirical density estimate of Definition 203 converges uniformly in measure to μ, i.e. for every $\epsilon > 0$,

$$P\left(\left\{\omega : \sup_{x \in \mathbb{R}} |f_n(x, \omega) - f(x)| < \epsilon\right\}\right) \xrightarrow{n \to \infty} 1. \tag{3.28}$$

For a proof see [Rosenblatt].
The following theorem is a specific case of the much more general convergence of Barycenters proved in [Ahidar-Coutrix-Le Gouic], in the paper the authors prove the d_p-convergence in metric measure spaces satisfying a positive curvature condition.

Proposition 224. *(Barycenters d_p converge)*
If μ has compact support and μ_n is the a p-Wasserstein Barycenter of Definition 204, then $\mu_n \xrightarrow{d_p} \mu$ as $n \to \infty$.

For a proof see [Ahidar-Coutrix-Le Gouic] and apply it to the simple case where $(R^p, |\cdot|, \mu)$ is given as the initial measure space. In [Manupriya-Keerti-Biswas-Chandhok-Jagarlapudi] the following proposition was shown:

Proposition 225. *(Total variation)*
The mmd_k *minimizer of Definition 213,* $\pi_n^{\mathrm{mmd}_k}$ *converges in total variation norm to the solution* π^* *of unrestricted transport with respect to* c *(Definition 3.16), i.e.*

$$\pi_n^{\mathrm{mmd}_k} \xrightarrow{||\cdot||_{TV}} \pi^*.$$

See [Manupriya-Keerti-Biswas-Chandhok-Jagarlapudi, Theorem 1].

Convergence and full learner recovery systems

In the previous section 3.4.4 we have listed several convergence results for different types of empirical measures. Empirical measures encompass our understanding of the sample. Theorems 218, 221 and 223, Propositions 219, 224 and 225 need to be coupled with regularity properties of L and the underlying class of functions $\mathcal{C}$ as in Proposition 195. This list shows that given an underlying model, it's intrinsic features will determine the type of measure pre-conditioners needed to ensure convergence on the specific convergence mode that the limiting measure admits.
For example, Proposition 225 involves convergence in Total Variation norm from which one can infer that the measure pre-conditioning of Definition of 213 applies for a d-continuous (in the first coordinate) loss function L as in Proposition 195. 4. In contrast, Theorem 199 shows that the empirical (uniform) measure is not well-suited for every limiting distribution and so in the case of a continuous density, preconditioning by 213 is proved to have better results (theoretically) than the empirical measure.

Estimating the marginal instead

In the discussion of density estimation (Section 3.4.1) we haven't done any specific distinction on the particular form the data for Problems 3.2 and 3.1. Definitions 201-213 work for all kinds of data. In the particular case of the ML Problems 3.1 and 3.2, our objective is to model in the class $\mathcal{C}$ the dependence of Y on X penalized by the loss function L. We aim to study how good (with respect to L) a $\mathcal{C}$-model $f(X)$ approximates Y. In this context the distribution π refers to that of (X, Y). Measure pre-conditioning amounts to approximating π using the sample in a way that benefits computations. We note that this gives rise to two different approaches:

(a) We can estimate π directly via π_n according to definitions 201 -213.

(b) We can make assumptions on the conditional distribution of $Y|X$ and then use definitions 201-213 for approximations on the X-marginal of π.

Most of the study of this document has focused on approach (a). Let us give an example of the approach (b) to show it's interaction with measure pre-conditioning.

Theorem 226. *Assume that* $Y|X = x \sim \nu_x$ *and that we have estimated* ν_x *via* ν_n^x *such that* $\nu_n^x \xrightarrow{d_p} \nu_x$ *uniformly on* x*, i.e. given* ϵ *there exists* $N > 0$ *such that for every* $n \geq N$

$$d_p(\nu_x, \nu_n^x) < \epsilon \text{ for every } x$$

assume also that $\mu_n \xrightarrow{d_p} \mu$, *and* $L : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ *is continuous. Let* $f \in \mathcal{C}$ *and assume that there exists* $g \in L^1(\mu)$ *such that*

$$\left| \int L(f(x), y) d\nu_n^x(y) \right| \leq g(y).$$

and that $y \to \int L(f(x), y) d\nu_n^x(y)$ *is continuous and bounded, then*

$$\int \int L(f(x), y) d\nu_n^x(y) d\mu_n(x) \xrightarrow{n\to\infty} \mathbf{E}_\pi[L(f(X), Y)].$$

Proof. The proof is a direct consequence of dominated convergence applied twice, observe that

$$\begin{aligned}
&\int \int L(f(x), y) d\nu_n^x d\mu_n(x) - \int L(f(x), y) d\pi = \\
&\int \int L(f(x), y) d\nu_n^x d\mu_n(x) - \int \int L(f(x), y) d\nu_n^x d\mu(x) + \\
&\qquad \int \int L(f(x), y) d\nu_n^x d\mu(x) - \int L(f(x), y) d\pi.
\end{aligned}$$

the first term goes to zero if we ensure $y \to \int L(f(x), y) d\nu_n^x(y)$ is continuous and bounded, the second term goes to zero by dominated convergence (using μ as reference measure). ∎

Remark 227. *(On the general approach and the restrictiveness of the hypothesis on Theorem 226). Theorem 226 is only one example of the multiple approaches one can use to estimate* π *from* μ_n *and assumptions on* $Y|X$*, even though the hypothesis of Theorem 226 are very difficult to meet in practice, it is presented here to illustrate the general idea.*
Measure pre-conditioning on the marginals ν^x *allows the modeller to include the specific features of each data class. It is clear the many lines of investigations one can explore to get similar results (with less restrictive hypothesis), we choose not to develop any further and leave it for future research.*

3.4.5 The recipe: How to choose a measure and how to implement the algorithm

The general approach for this document is to put in a single, standard, theoretical background many ideas that have come to light in ML-research. Namely, ML-reaserchers have realized that their algorithms improve in performance or convergence properties after a small "tweak" to either data or the loss function occurs. Stability of ML-algorithms has been widely known and is one of the main focus of ML-research. The idea of measure pre-conditioning is that the standard empirical distribution, though it may contain all the possible information in terms of inference (except for order) may not be well adapted to the specific problem one aims to minimize. It is well-known for example that if the functional to be minimized is convex, algorithms used for minimization can take advantage of convexity. This encourages the solver to find an empirical estimation from definitions 201-213 that makes their functional convex. Finding such a measure is what we call **pre-conditioning**, if the preconditioning satisfies any of the assumptions of Proposition 195 then one is ensured to have a full learner recovery system and hence have not lost anything on the process while achieving improved performance. One could instead use a pre-conditioner based on many

reasons (such as having a specific algorithm to compute already at hand for example), this work explains how as soon as a condition like Proposition 195 is satisfied, one will end up with the same classifier/regressor.

3.5 The problem of Domain Adaptation and the impact of measure pre-conditioning

Domain Adaptation (DA) is a sub-problem of transfer learning on which one aims to infer the parameters for a new learning agent in terms of an agent that learn in similar data.
Many of the DA adaptation formulations are well-suited for Optimal Transport (OT), our framework of Problems 3.2 and 3.1 was motivated at first by the recent research in optimal transportation in Machine Learning (see [Amos-Cohen-Luise-Redko], [Redko-Vayer-Flamary-Courty], [Courty et al.], [Courty-Flamary]) and so in this section we explore the implications of measure-preconditioning in the specific case of domain adaptation problems related to optimal transportation and the recent research in the area (see [Courty et al.], [Courty-Flamary] and references therein for a more complete exposition of the use of optimal transportation in machine learning).

Problem 3. *(General domain adaptation problem)*
Suppose that we have a sample $(X_1^s, X_2^s, \dots, X_n^s)$ *of features together with the a sample of the dependent variable* $(Y_1^s, Y_2^s, \dots, Y_n^s)$ *and we use the learning agent to minimize a loss function* $L : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ *among a class of functions* $\mathcal{C}$*. The learning problem is to obtain the best possible parametric function* f*, among the class* $\mathcal{C}$ *explaining the data, i.e.*

$$\min_{f \in \mathcal{C}} \left\{ \sum_{k=1}^{n} \mathbf{E}[L(f(X_i^s), Y_i^s)] \right\} \tag{3.29}$$

If f_s^* *realizes the minimum in* (3.29)*, we say that it is the learnt agent or that* f_s^* *correspond to the learnt parameters.*
Now suppose we have another sample $(X_1^T, X_2^T, \dots, X_{n_2}^T)$ *which we believe is similar in some features to the original sample. The domain adaptation problem is: How much can one learn from the previous learning? That is, how can we transfer the learning from the source domain to target domain?.*

The research field which attempts to answer Problem 3 is known as Domain adaptation for transfer learning. For a general introduction and approach see [Courty et al.], [Bishop] and references therein.
The problem of domain adaptation 3 is different to Problems 3.1, 3.2 as it aims to transfer the statistical knowledge obtained by a minimization on source-domain to a minimization on the target-domain. The formulation of Problem on (3.29) has the implicit assumption of the empirical distribution being imposed at level n.
In this section we aim to explain how measure pre-conditioners as defined in section 3.4.1 can be used in the field of DA for transfer learning.

Problem 4. *(Domain Adaptation and transfer learning with varying losses and classes)*
Suppose that we have a sample $(X_1^s, X_2^s, \dots, X_n^s)$ *of features together with the a sample of the dependent variable* $(Y_1^s, Y_2^s, \dots, Y_n^s)$ *and we use the learning agent to minimize a loss function*

$L_s : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ *among a class of functions* $\mathcal{C}_f$. *The learning problem is to obtain the best possible parametric function* f, *among the class* $\mathcal{C}$ *explaining the data, i.e.*

$$\min_{f \in \mathcal{C}_f} \{\mathbf{E}_{\pi_n}[L_s(f(X^s), Y^s)]\} \tag{3.30}$$

and compare it with the perfect learner on target domain with class $\mathcal{C}_t$ *and loss function* $L_t : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$*:*

$$\min_{f \in \mathcal{C}_t} \left\{\mathbf{E}_{\pi^t}[L_t(f(X^t), Y^t)]\right\} \tag{3.31}$$

If f^* *denotes the minimizing argument for* (3.30)*, the Domain Adaptation problem is: How can we use* f^* *to obtain good estimates for* (3.31)*?*
What is the structure of such agent?
How does it compare to the actual minimizer of (3.31)*?*

Suppose that every $X_i^s \sim \mu_s$ and $X_i^t \sim \mu_t$, under "similarity assumptions" on μ_s and μ_t, one expects to be able to transfer learning to some accuracy. Of course "similarity assumptions" depends on the context of the ML-task in hand.
For example, two measures might be considered similar in a classification problem that may not be considered similar in a generative model. In the same fashion, suppose that μ_s and μ_t satisfy that there exists a solution, T, for Problem (3.16) with a given cost function $c : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. A good candidate for a new learnt agent can be immediately obtained via $f^* \circ T^{-1}$. As seen in [Courty-Flamary], the error made by this agent relative to the total error obtained from training an agent from scratch can be controlled as soon as μ_s and μ_t are d_2-close and $\mathcal{C}$ is rich enough. In the field of Domain Adpatation (DA) usually at least one of the following assumptions is made:

Assumption 4. *(Conditional structure of learning task)*
In the context of Problem 4, if (X^s, Y^s) *is the source variable and* (X^t, Y^t) *the target variable, it is common to ask that*

$$(Y_i^s \mid X_i^s) \sim (Y_i^t \mid X_i^t), \tag{3.32}$$

where $Y \mid X$ *denotes the random variable whose law is the regular conditional probability of* Y *given* X.

This assumption means that the probabilistic structure of the dependence of Y on X is the same in both domains. We understand this assumption as a strong hypothesis of similarity in the modellings.

Assumption 5. *(Identical dependence)*
In the context of Problem 4, if (X^s, Y^s) *is the source variable and* (X^t, Y^t) *the target variable, it is common to ask that*

$$(X_s, Y_s) \sim (X_t, Y_t)$$

The identical dependence assumption has been used extensively but is in general not a good idea to pre-impose. The identical assumption implies that any sample of the source domain can be considered a sample of the target domain so if $L_s = L_t$ and $\mathcal{C}_s = \mathcal{C}_t$ then the learning transfer is perfect as we can identify the source data as target data in the empirical destimation of $\pi_s = \pi_t$
The following assumption can be found in recent papers in DA-ML, see [Courty et al.] for example.

Assumption 6. *(c-optimal map)*
There exists an optimal transport map (with respect to a cost function $c : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$) T_c as in (3.16) that satisfies

$$(X_i^s, Y_i^s) \sim (T_c(X_i^s), Y_i^t).$$

Remark 228. *Though it is straightforward to use Assumption 6 (postulated in [Courty et al.]) in the context of optimal transportation, it is of significant importance to understand the necessary conditions that yield this assumption.*

Remark 229. *Note that these assumptions and the framework of DA is closely related to the line of investigation proposed in Remark 227 below.*

3.5.1 General Idea in the non-linear case

Domain Adaptation should be used when the target and source measures are believed to be similar. If the source measure satisfies the assumptions of Brenier's Theorem (see [Villani2003, Theorem 2.32]) and the loss function is quadratic (or strictly convex function of quadratic distance) the optimal transport map T transporting μ_s onto μ_t can be used as an learning agent on the target domain. We do this by first mapping onto the source domain using the optimal transport map and only then evaluating the agent that has learnt paramters, i.e. define f_{ad} a candidate for the minimization of loss for learning agents in the target domain by

$$f_{ad} = f^*(T^{-1}).$$

The work in [Courty et al.] shows a convergence for this agent under Assumption 4.

3.5.2 Main question: What cost should we impose?

Note that Assumption 6 is an existence condition. If there exists a cost function for Assumption 6 one would need to check that it satisfies the conditions for existence and uniqueness of optimal transport maps like regularity and the twist condition (see [McCann-Guillen], [Villani2003],[Santambrogio]). In the general approach for DA using transfer learning via optimal transport in the framework of Problem 4, two problems seem to arise more often in practice:

P.i) When the conditions of the trainings are fixed and not to be chosen: study a learnt agent when $L_1, L_2, \mathcal{C}_1, \mathcal{C}_2$ are given and fixed.

P.ii) When we are able to choose $L_1, \mathcal{C}_1$ with the goal of maximizing (in any way) the transfer learning for a given loss function L_2 and class $\mathcal{C}_2$.

3.5.3 A measure of transferrability

In Problems 3.30 and 3.2, we start under a similarity assumption on the source measures. This follows an intuitive statement: in order to be able to transfer any learning, the original measures should share some features. We can't expect to transfer any learning if the problems have nothing in common.
We may expect to transfer the learning (classifier) differentiating between dogs and cats to a new agent aiming to differentiate wolves and lions. In this case the distribution of dogs and cats is

believed to be similar to that of wolves and lions.
How much could we transfer? Could we guess beforehand how much learning we can transfer?
As a thought experiment, let us study a way to measure the transfer of learning. There are many ways to measure transferability, see [Courty-Flamary], [Courty et al.] or references therein. We propose another one, assume that π^s and π^t are as in Problem 3, let $h : \mathbb{R} \to \mathbb{R}$ be any *strictly* convex function with $h(0) = 0$, set

$$d_h(\pi^s, \pi^t) = \inf_{\Pi \in \Pi(\pi^s, \pi^t)} \int h(L_1(f_1(x_1), y_1) - L_2(f_2(x_2), y_2)) d\Pi((x_1, y_1), (x_2, y_2)). \tag{3.33}$$

where f_1 is the solution for the $\mathcal{C}_1, L_1$-source problem and f_2 the corresponding solution for the $\mathcal{C}_2, L_2$-target problem. Evidently, a-priori the value of $d_h(\pi^s, \pi^t)$ can not be computed as f_1, f_2 are unknown and the value of (3.33) depends on the choice of models $(\mathcal{C}_1, L_1)$ and $(\mathcal{C}_2, L_2)$. We claim (3.33) is a reasonable way to measure transfer depending on $\mathcal{C}_1, L_1, \mathcal{C}_2, L_2$, in the sense that the closest d_h is to 0 the more likely it is that a learnt agent for the L_1 problem with source data (X^s, Y^s) would perform decently in the L_2 problem with data (X^t, Y^t). This is to be expected as it may be reasonable to transfer the learnt agent for certain loss functions but not with all of them. Even though f_1 and f_2 are unknown, in some cases some estimates can be obtained. To the knowledge of the author no measure of transferrability of the form (3.33) has been studied which points to a promising line of investigation.

3.5.4 Problem 1

Let us first address problem **P**.i) where all the conditions $(\mathcal{C}_1, \mathcal{C}_2, L_1, L_2)$ are fixed and we aim to measure the efficiency of a solution to (3.1) and (3.2).

Measure pre-conditioning in the conditional average guess

Let us consider here a different approach to the general Problem 4, suppose that we have solved the source problem i.e.

$$f^* \in \arg\min_{f \in \mathcal{C}_1} \mathbf{E}_{\pi^s}[L_1(f(X), Y)]. \tag{3.34}$$

Similar to the ideas in [Courty et al.] one can make assumptions like Assumption 6 in order to benefit from the source sample by using conditional distributions. Given $y \in \mathrm{spt}(\mathrm{Proj}_2 \#\pi^s)$ and $f \in \mathcal{C}_1$, assume we can find $T^{f,y}$ optimal transport map for the cost function $c_y(x, \tilde{x}) = |L_1(f^*(x), y) - L_2(f^*(\tilde{x}), y)|$ between the conditional distributions $\pi^s(x|Y = y)$ and $\pi^t(x|Y = y)$. The question is now how to generate an element in $\mathcal{C}_2$ from the learnt information on the conditional distributions. The first immediate guess is to average with respect to the target distribution, that is if

$$d\pi^t(x, y) = d\pi^t(x|Y = y) d\nu^t(y)$$

a guess for a learnt agent would be

$$f_{ad} = f^* \circ (T^{f^*})^{-1}, \text{ where } T^{f^*}(x) = \int_Y T^{f,y}(x) d\nu^t(y). \tag{3.35}$$

In the general case, no estimates on the control of learning for agent (3.35) are known.
It is expected that if the measures satisfy that d_h from (3.33) is small then the agent obtained using

(3.35) is good although so far no precise statements have been shown. Formula (3.35) is a reasonable guess because it takes into account the best agent at each y before averaging over all $y \in Y$.

Open Question 1. *In the context of Problem 4, is it true that if $d_h(\pi^s, \pi^t)$ is small, then f^*_{ad} performs well in (3.31) when constructed using (3.35) and pre-conditioning? Is this performance quantifiable? Is it true that as $n \to \infty$,*

$$\mathbf{E}_{\pi_n}[L_2(f^*_n(X), Y)] - \min_{f \in \mathcal{C}_2}[L_2(f(X), Y)] \to 0?.$$

Can such performance be studied by d_h of (3.33) when $h(r) = |r|$?.
How does (3.35) compare to

$$f^* \circ T_2, \text{ where } T_2 = \int_Y (T^{f,y})^{-1}(x) d\nu^t(y)? \tag{3.36}$$

This questions are relevant both in the field of transfer learning and to measure pre-conditioning. The computation of $T^{f,y}$ may be difficult in practice and we expect measure-preconditioning for every y to benefit the performance of the intermediate algorithms without disruption on convergence. Numerical simulations are being performed to corrobate this idea and study the performance of (3.35) and will appear in subsequent works.

Data-driven conditional OT

On [Trigila-Tabak] the authors studied the following problem given a cost function c in the product space $X \times Z$ and a probability measure on $X \times Z$:

$$\min_{\substack{T(\cdot,z) \\ \forall z T(\cdot,z)\#\rho(\cdot|z)=\mu(\cdot|z)}} \int c(x, T(x,z)) d\rho(x,z) \tag{3.37}$$

which they denoted the data-driven optimal transport problem. In the same work, the authors showed that the minimization of (3.37) is equivalent to

$$\min_{T(\cdot,z)} \max_{\lambda \geq 0} \int c(x, T(x,z)) d\rho(x,z) + \lambda \operatorname{Ent}\left(\mu(\cdot|z) \middle| \frac{1}{2}(T\#\rho(,z) + \mu(\cdot,z))\right) \tag{3.38}$$

The dual formulation of (3.37) via (3.38) already hints a connection with our work. As the algorithm implemented in [Trigila-Tabak] is a sequential algorithm using gradient descent, it can be interpreted in the sense of measure pre-conditioners that entropically regularize at every discrete step n, just as Definition 210 in the framework of Wasserstein distance and problem 3.2. This means that an algorithm to compute data-driven conditional optimal transport can benefit directly from measure-preconditioning.

3.5.5 Control on optimal transport domain adapted learning

In this section we present different hypothesis and assumptions that yield stability results on transferred learning. The results are not as strong as those conjectured in section 3.5.4 but directly related to measure pre-conditioning.
It is evident that there are many options on $\mathcal{C}_1, \mathcal{C}_2, L_2, L_2$ that will ensure the transfer learning is efficient. In this section we reduce to present the most straight-forward formulations.

Proposition 230. *Let T be any map with $T\#\mu = \nu$ and $d\pi_1(x,y) = d\pi_2(T(x),y)$ if $\mathcal{C}_1 \circ T = \mathcal{C}_2$ then*

$$\arg\min_{f\in\mathcal{C}_1} \mathbf{E}_{\pi_s}[L(f(X^s),Y^s)] = \arg\min_{f\in\mathcal{C}_2} \mathbf{E}_{\pi_t}[L(f(X^t),Y^t)]$$

The proof is a direct consequence of the composition of classes $\mathcal{C}_1 \circ T = \mathcal{C}_2$.

Proposition 231. *If $\mathcal{C}_1 = \mathcal{C}_2 = \mathcal{C}$ and $L_1 = L_2$ and if $\mathcal{C}$ is so that $(x,y) \to L_1(f(x),y)$ is Lipschitz and bounded then for every f*

$$|\mathbf{E}_{\pi_1}[L_1(f(x),y)] - \mathbf{E}_{\pi_2}[L_1(f(x),y)]| \leq d_1(\pi_1,\pi_2)$$

and so the total loss of transfer learning when the learned agent is adapted is controlled by the d_1-distance between joint measures.

Proof. The proposition follows directly from the Kantorovich-Rubinstein representation of the d_1 norm as d_1 is the suprema over Lipschitz functions. ∎

In [Courty-Flamary] the authors proved the following theorem:

Theorem 232. *(Courty-Flamary)*
If $L(x,y) = |x-y|^2$ and $\mu^s = \frac{1}{n}\sum_{k=1}^{n} \delta_{x_i^s}$ where $x_1, x_2, \ldots, x_n \in \mathbb{R}^n$ and there exist A positive definite matrix and a vector b such that $\mu^t = \frac{1}{n}\sum_{k=1}^{n} \delta_{Ax_i^s+b}$, set $T(x) = Ax + b$ then $f_* \circ T^{-1}$ is a perfect learning agent in the sense that it minimizes (3.31).

See in [Courty-Flamary, Theorem 3.1]. We now generalize this idea before we continue.

Theorem 233. *Let π_s, π_t be the joint measures for the the source (X^s,Y^s) and (X^t,Y^t) target domains respectively. Denote μ_s and μ_t the projections into the X-coordinates of π_s, π_t and by μ_x^s and μ_x^t the conditional distributions of $Y^s|X^s$ and $Y^t|X^t$. Assume there exists a map $T : \mathbb{R}^p \to \mathbb{R}^p$ such that*

1. *$T\#\mu_s = \mu_t$*
2. *$\mu^t_{T(x)} = \mu^s_x$*
3. *$\mathcal{C}_2 = T \circ \mathcal{C}_1$*

if f is the solution for (3.30) then $f \circ T^{-1}$ is a perfect learner in the sense that it minimizers (3.31).

Proof. The proof relies only on the disintegration of measures, as

$$\mathbf{E}_{\pi^t}[L(f(X^t),Y^t)] = \int \left(\int L(f(x),y) d\mu_x^t(y) \right) d\mu_t(x) = \int \left(\int L(f(T(x),y) d\mu_x^s \right) d\mu_s(x)$$

where we have used the condition $d\mu^t_{T(x)} = d\mu^s(x)$ in the last equality. Minimization over $\mathcal{C}_2$ and the condition $\mathcal{C}_2 = T \circ \mathcal{C}_1$ yields the result. ∎

Open Question 2. *(Can learning error be totally controlled?)*
Assume f_* minimizes the target problem, under what conditions on $\mu, \nu, L, \mathcal{C}_1, \mathcal{C}_2$ does there exist $C > 0$ such that

$$\left| \frac{1}{n} \sum_{i=1}^{n_2} \mathbf{E}[L(f_*(X_i^t), Y_i^t) - L(f_{ad}(X_i^t), Y_i^t)] \right| \leq C d_2(\mu_s, \mu_t)?$$

The previous theorems and the ideas of [Courty et al.] respond this question in very restricted situations. Having a general context to answer this question similar to the one of 191 would be essential for the theory of domain adaptation.

3.6 Outside of the framework

In this section we explain how the framework developed in this article can be extended to encompass more general situations (whose formulation is not exactly represented by (3.1) and (3.2)) but benefit from the same ideas.

3.6.1 Using pre-conditioners on WGANs

The Wasserstein Generative Adversarial Networks (WGAN) introduced in [Arjovsky et. al] is a generalization of the generative adversarial networks (GAN) introduced in the seminal work [Pouget-Abadie et al.]. The reason to consider the Wasserstein framework is due to the convergence properties of the Wasserstein metric together with the representation of Kantorovich-Rubinstein. The WGAN problem consists in computing

$$\arg\min_{\theta} \arg\max_{w \in \mathcal{W}} \mathbf{E}[f_w(X)] - \mathbf{E}[f_w(g_\theta(Z))] \tag{3.39}$$

where $X \sim \mathbb{P}_1$ is prescribed, $Z \sim \mathbb{P}_2$ and $\{g_\theta\}_{\theta \in \Theta}$ is a parametric function space. Further work would study the same principles applied in this document to the more general version of the problem admitting (3.39) using maybe 2 parametric families $\mathcal{C}, \tilde{\mathcal{C}}$. The only difference between our problem and (3.39) is the presence of an extra outer minimization problem. It is clear that algorithms like TTC presented in [Milne] that take a dual approach can benefit from sequential measure-pre-conditionining. In the original formulation, as in [Arjovsky et. al]

$$\sup_{f \in \mathcal{C}} \int f d\mu - \int f d\nu + -\lambda \int (|\nabla f| - 1)^2 d\sigma$$

where $Z \sim \sigma$ iff $Z = tX + (1-t)Y$ where $t \sim U[0,1]$, note that we can replace μ and ν at level n via the empirical measures or measure pre-conditioners. This means that measure-preconditioning can be applied in more general circumstances than Problem 3.1 as the estimation of σ can be done via $tX_n + (1-t)Y_n$ where $t \sim U[0,1]$ and the triangle inequality yields convergence.

3.6.2 Covariate shift domain adaptation problem

In general, the label-shift domain adaptation problem is usually written as

$$\min_{h,g} \frac{1}{n} \sum_{i=1}^{n} L(h(g(x_i^s)), y_i^s) + \lambda \operatorname{Ent}(\mu_s^g | \mu_t^g) + \Omega(h, g) \tag{3.40}$$

where h is the hypothesis, g is a representation mapping and Ω is a regularization term. The first term corresponds to losses in approximation while the second and the third correspond to regularizations. Compared to the framework used in Problems 3.30 and 3.31, (3.40) is a more general version. Nevertheless, the idea of measure pre-conditiniong can substitute the entropy term by using a sequence of entropic regularizations and $\Omega + L$ can be used as a modified loss function. The difference in algorithmic performance of both approaches is an interesting project.

3.6.3 COOT and measure pre-conditioning

In [Redko-Vayer-Flamary-Courty], the following problem was introduced to handle at the same time the disparity between correlated distributions and the data marginals. In the case where $X_i \in \mathbb{R}^p$, authors in [Redko-Vayer-Flamary-Courty] consider the matrix $X = (X_1, \ldots, X_n)$ not only as a sample where the randomness comes form a single distribution but as a doubly-random matrix in the sense that each row is considered a sample and the columns are consider features, in this context let μ_S denote the probability measure associated to samples and μ_F the associated feature distribution one should perform optimal transport simultaneously in sampling and feature spaces. We expect the techniques of the two previous sections to also work in this context mutandis mutatis.

3.7 Researcher's criteria on measure pre-conditioning

In section 3.4.5 we explained what a ML-developer should consider as recipe for applying measure pre-conditing. It explained that each modification of the n-level measure had different implications which should be pointed towards some (algorithmic) benefit. In general, it may be difficult to know a-priori exactly what to use and so this (and subsequent) work should be considered as a guideline.

3.7.1 Trade-offs

In low-dimensional regimes, absolutely continuous (w.r.t. Lebesgue) tend to behave better, while in higher dimensions highly concentrated measures tend to have better properties, see [Panaretos-Zemel, Chapter 4]. This is already a hint on what to do, if the problem involved has few features, absolutely continuous measures may improve the performance of the algorithm.

3.8 Conclusions and further work

Recent work [Redko-Vayer-Flamary-Courty] has introduced new techniques for domain adaptation, the idea is to optimally match features and samples, it is still open lines of investigation how different measure pre-conditioning techniques would impact the co-optimal transport problem. The features and samples are in general of very different nature for which combining more than one of the techniques of section 3.3.5 could improve the performance of the algorithms. For example, it may be the case that features share a structure that can be exploited by a specific technique while the relation between samples may algorithmically benefit from another.

3.8.1 Order of convergence

Establishing that the ML problem gives a full learner recovery system is good in order to know convergence is ensured, in algorithmic practice we need more. We need to study the order of

convergence and the imrpovements on this order by Measure pre-conditioners, this work is left for future work and other researchers.

Data-driven model changes and convergence

In the start of section 3.3.1 we asked question (iii): Given a choosing of π_n's, could we find sequences L_n's and $\mathcal{C}_n$ so that the computations on the $\mathcal{C}_n-$ problem with loss function L_n associated to π_n converge to 3.1? Could these problems improve the algorithmic performance?
In section 3.3 we studied conditions on $\mathcal{C}$ and L to ensure Definition 191 and consequently Theorem 193. The question of how and when to change $\mathcal{C}_n$ and L_n at every step is still open and interesting. A good answer would yield heuristics to change the model given the data in terms of the parametric space, this means to not only change the way we measure the information from the data but also how we learn from it. This line of investigation is left for future work.

3.8.2 k-nearest neighbohrs and relation to meta-transport

k-nearest neighbors and point-process notation

The list of empirical estimating probabilities (section 3.4) is obviously non-exhausting. Algorithmic treatment of data such as k-nearest neighbors represent a potentially significant pre-conditioning method. The theory of this algorithms is usually developed through point-processes. The extension of this work to point-processes together with section 3.8.1 is a promising area for mathematical theory of learning.

Meta-transport

Another recent development in Optimal Transport based machine learning is the development of meta-optimal transport in [Amos-Cohen-Luise-Redko]. The basic idea, similar to the basic idea of this document is to present a way to improve the performance of ML-algorithms through pre-working on them. The seminal work [Amos-Cohen-Luise-Redko] develops completely algorithmic-focused techniques, as explained in section 3.7.1. This work is focus on the underlying structured of pre-condiitoning the samples, the statistics in Wasserstein space and how they impact the outputs of the algorithms. In some way, [Amos-Cohen-Luise-Redko] tackles the pre-conditioning/pre-measure pre-conditioning in a different manner, with a clever approach based on numerical algorithms. We expect that a theory similar to the one developed in section 3.4 can also encapsulate the algorithmic pre-conditioning. This can be modelled via point-processes (as it's done for k-nearest neighbors).

3.8.3 General disintegration estimates

One can study different conditions on $L, \mathcal{C}, \mu, \pi, Y|X$ such that a convergence similar to Theorem 226 occurs. This area is particularly technical as disintegration is not a continuous operation with respect to some metrics on spaces of probability measures. Generally, one does not necessarily need to estimate the disintegration but can explore different methods of convergence. An approach to full learner recovery systems (191) in the special case of assumptions on $Y|X$ would be interesting and related with sections 3.5.4, 3.5.4 and literature as to [Trigila-Tabak] and references therein.

3.8.4 Problem 2 of section 3.5.2

If L can be chosen thinking ahead of the Target problem, choose the cost function by chosing an L_1 depending on L_2 or viceversa. The idea of the problem is to ensure learning can be transferred by picking the problems with the goal of transferring. A full theory with the approach of training with the goal of transferring would be interesting on it's own.

Choosing the first Loss function to improve the second

With the same approach as in Section 3.8.4, if we know that we aim to solve the target problem for $L_2, \pi^t, \mathcal{C}_2$, what loss function L_1 should we chose given π^s and $\mathcal{C}_1$? Similarly, allow $\mathcal{C}_1$ to be chosen. We should chose L_1 in a way that data under π^s behave similar to L_2 under π^t. How one takes the target problem into consideration is an open question.

3.8.5 Choosing the target loss model according to the source

Assume we have solved Problem 3.1 with set of features $L_1, \mathcal{C}_1, \pi^s$ and we know there is a distribution (unknown to us) on which we aim to transfer the knowledge, what loss function L_2 would ensure good properties of the learnt agent on target space? One can think of an L_2 loss function that penalizes the error of the learnt agent and *simultaneously* penalizes the difference between probabilities. This function would take into account that a mistake in the model is not relevant when one knows the error on difference of distributions is big. The L_2 loss function could be used to simultaneously control model error with (probability) transfer error.

Chapter 4

A generalization of an economic model of Roy for labor distribution under occupational choice

In collaboration with Jeffrey Liang and Aloysius Siow.

4.1 Introduction

In the seminal work of A.D. Roy [Roy], a model for the distribution of occupations in a community is presented. The objective of the model is to explain using specific conditions on the needs of the community, which workers will dedicate to each possible occupation assuming we can give a numerical value to the skills of performing each job. Recent work ([Siow-Mak2016],[Siow-Mak]) has focused on a generalized version of the model of [Roy] where a separation function, dividing workers from both roles, has had a successful impact on the analysis of the theory. In this article we study analytical details of these separation functions, the necessity of them and consequences of this formulation.

We study a problem based on [McCann-Trokhimtchouk], [Siow-Mak2016] and [Roy] on which workers aim to be hired in companies for any of two roles. The two roles can be thought of as manager (primary) and assistant (secondary) and the companies will hire pairs of workers so that each worker will do one job. Each matched couple of workers will be assigned to a company on which one of the agents will be the manager and the other one will be the assistant. The underlying matching is modelled by a production function that evaluates the numerical output of their work, corresponding to a benefit function in the context of optimal transportation, in the context of [McCann-Trokhimtchouk].

The work on both occupations is remunerated by wages, the structure of this salaries is what we refer as earnings schedule. Given an earnings schedule, i.e. determination of salaries for both options, the occupational choice model requires each worker to choose what job to do. We assume that each worker's decision is motivated only by the earnings schedule and hence will choose the job that will

pay the most according to their own set of skills. The problem studied here involves also finding the optimal earnings schedule for the economy and therefore differs from the standard occupational choice in the sense that the earning schedules are not known a priori but are part of the solution. An important step in the development of this model is Section 4.2.6 on which we show that the optimal solution of the model corresponds to a 2-step problem on which the inner problem is the occupational choice for workers given earning schedules.
The problem presented here generalizes the Roy model from [Roy] where the separation between workers is assumed to be linear. In this generalization, presented to us by Dr. Siow, we allow the separation to be non-linear. One could expect that in certain specific production functions depending on interaction, one would always get linear separation, we show this is not the case by means of an example in section 4.3.4.
We also study the relationship between this model and the general version of the social planner's problem presented in [McCann-Trokhimtchouk].

4.1.1 Plan of the Paper

In section 5 we present the generalized Roy model and introduce the formulation of Dr. Siow. We explain the role of each variable and some technicalities. We describe the problem, the assumptions made and provide intuition and interpretations of the model.
In section 4.2.6 we rewrite the original model as a 2-step problem which allows us to introduce the framework of optimal transportation of mass, which is a theory proven to be very useful in many physical and economical problems, for introduction to the theory see [McCann-Guillen], [Ball], [Villani2003] and for interesting applications to economics and physics see [Galichon], [Santambrogio] to name a few. The idea is that any equilibrium of distribution of workers into roles should involve optimal matching between managers and assistants, allowing optimal transportation theory to handle an inner problem. The rest of the section is dedicated to obtain further understanding of the problem using the optimal transport framework.
In section 4.5 we review the model of McCann and Trokhimtchouk, from their seminal work [McCann-Trokhimtchouk]. We introduce their model and state some important results before making the connection with the generalized Roy model presented here. At this point we also include simple examples to fix intuition and also resolve a conjecture by Dr. Siow on whether or not the separation function would result to be linear and hence agreeing always with the original model of Roy from [Roy] and [Heckman-Honore].
Finally in section 4.4 we study how the different variables affect the model and yield a continuity result for separation. This result intuitively says that similar economies and similar work-forces will yield similar separations of the labor force. Here we take advantage of the separation function to study the wage inequality, the difference in salaries between two people working in the same firm and conclude an interesting bound that provides economical insight in 4.4.2. We finish by exploiting the first order conditions and duality to understand approximations of optimal total production from similar economies in terms of production or in terms of original distribution of people.
In section 4.7, we explain how this model can be implemented and possibly changed as well as explaining what we believe the next lines of investigation of this model should look like.

4.1.2 Preview and conclusions from the economical stand-point

In this section we give an economical preview of the results, that is we explain their economical significance before developing the mathematical tools required.

- In section 4.2.1 we formulate the model, the objective is to obtain a model for the distribution of roles that optimally includes occupational choice and the best matching for revenue. We study the implication of having firms competing to hire the agents, which establishes a structure in the earning schedules that we model through the use of the F-transform (Definition 235). We know that agents will decide their occupation in an strictly economical way, meaning that they will always take the better-paying job. This choice is modelled as a constraint, as we separate the agents doing the different roles by a function that we call the separation function (Definition 238). The separating function itself, has to be defined by the wages, and the occupational choice constraint forces all possible separation functions to have the same structure (Definition of $\mathcal{C}$ in 4.13). This idea has to be coupled with the fact that everyone will be employed, both of this conditions determine the set on which the optimization is done.

- In section 4.2.6 we propose a slightly different model (4.2.6). In the first model, as all the optimizations are done at the same time, agents don't really know the wages and so they happen to be guessing what the best paying job given their skills is. This guess is eliminated by the 2-step program. This second problem gives the agents a possible wage structure on which they base their decision for a role and then optimizes over all feasible wage structures. We show that these problems are equivalent in 4.2.6.

- For section 251, we establish existence of equilibrium for the 2-step model. In order to do this, several continuity properties need to be shown. These properties also explain how the model interacts to small changes. For example, in Lemmatta 254 and 257 we see that if a wage structure is changed a little, one does not expect agents to radically change their decisions, on the contrary one expects the new distribution of roles to be similar. This is the content of this section.

- As the draft by [Siow-Mak2016], in the section 4.2.9 we derive first order conditions on the separation function (4.22). We learn that the rate of change of the line that separates people into jobs depends on the amount of change of production with respect to each skill and the rate of change of salaries. We also obtain a expected result: the change in wages depends on the partial derivative of revenue with respect to the corresponding skill, evaluated at the matching.

- We explore an optimality conjecture: Conjecture (4.26). One can expect the optimal wage schedule to be stable. That is, once agents see the optimal salaries in the market, they will adjust to use that schedule. We formulate this conjecture rigorously and show some progress towards it (Observation 265).

- In section 4.5 we look at the Social Planner's problem of [McCann-Trokhimtchouk] and compare it to our model. The model in [McCann-Trokhimtchouk] allows a general framework in which revenue depends on both skills of both agents, in this context the twist condition (4.39) says that a match on which a manager and an assistant are better in both skills is better for production. Nevertheless, our model has the singularity that once both agents are hired, the skill for the job they are not performing is irrelevant. This is a reasonable assumption,

economies on which both jobs are differentiated need to find good couplings but the roles for the hired pairs are truly distinguished. This formulation differentiates from said previous work (as seen in Proposition 271).

- In section 4.3 we study what happens with pure interaction and we note that the formulation is an actual generalization of the previous models as in general the separation function need not to be linear.
- After, we show mathematically that if two economies are very similar (in appropriate sense) the resulting distribution of roles will also be similar (Theorem 266). We do this by showing that the change in separation function has to be small and we quantify this change in terms of the difference of this two markets.
- Finally, we provide a bound on the wage inequality (4.37). We show that the difference of salaries between a manager and an assistant is at most a factor of the maximal change in revenue by one skill multiplied by how different a person match is to a person who you would be indifferent to swapping jobs with.

4.2 Generalized Roy Model

The seminal model of Roy [Roy] is a simple community where workers will decide on their own whether to fish or hunt. The fundamental idea is to allow ourselves to think about the amount of rabbits hunted or the amount of fish in terms of a numerical pair that represents both skills of each person. In [Heckman-Honore] we see the first mathematical modelling of the ideas from [Roy]. The model presented here is called the generalized model of Roy for matching, as we start from the fundamental idea in [Roy] that the skills of the workers are quantified by numerical values, that the workers will decide their occupation by themselves but we add the constraint similar to [McCann-Trokhimtchouk] that two workers are needed in every firm and each of the workers will do a single job. Our work differs from [Heckman-Honore] in the sense that we do not assume any linearity in how the distribution of workers will be split and we incorporate the occupational choice as a contraint.

4.2.1 The model

We assume there exists a group of people which represents the labor force and want to be employed. Every person in this labor force will be required to get one of two jobs: Manager or Assistant. Manager and assistant represent a key and a secondary role, respectively. We assume that every worker is capable of doing any of the jobs and will decide on their own which job to take. Each worker is represented by a skill set, an ordered pair of numerical values (k, s) on which the first coordinate represents the level of skill of the worker for performing the key occupation (manager) and the second coordinate represents the skill of the person for the secondary job (assistant). Notice that this assumption can be relaxed to a multi-dimensional skill set $(\tilde{k}, \tilde{s}) \subset \mathbb{R}^{n\times m}$ without much loss of generality as explained in [Siow-Mak2016]. We do not handle such generalization in this document.
We model a labor force by a distribution $\mathcal{R} \in \mathcal{P}_{ac,c}(\mathbb{R}^2)$ where $\mathcal{P}_{ac,c}(\mathbb{R}^2)$ is the set of Borel probability measures on $\mathbb{R}^2$ that are absolutely continuous with respect to the Lebesgue measure and have compact support. We denote by R the density function of $\mathcal{R}$.

We assume the existence of a production function $F : \mathbb{R}^2 \to \mathbb{R}$ where the value $F(k, s)$ represents the amount of money generated by a couple of workers where the worker performing the key role has skill k to do the job and it's pair has skill s to do the secondary occupation.

Definition 234. *(Strict supermodularity)*
Let $F : \mathbb{R}^2 \to \mathbb{R}$, we say F is strictly supermodular if

$$F(c, d) + F(a, b) - F(a, d) - F(c, b) > 0 \text{ whenever } a < c, b < d. \tag{4.1}$$

If we assume that F is twice differentiable, then supermodularity implies that the cross partials are positive and so by addition of a constant we can assume without loss of generality that

$$\partial_1 F(k, s) > 0, \partial_2 F(k, s) > 0 \ \ \mathcal{R} - a.e. \text{ on } (k, s). \tag{4.2}$$

From now on, we always assume our production F is strictly supermodular and satisfies (4.2). In economic terms, strict supermodularity of the production function represents an economy on which strictly better skills yield strictly better monetary outputs for the firms.
Under a strictly supermodular production function F, the Generalized Roy model aims to study wages and distribution of occupations among the labor force $\mathcal{R}$ on which:

- Every worker will be employed.
- Firms will hire by pairs. For each firm, one worker will be a manager and the other one will be an assistant.
- Each worker will attempt to maximize salary.
- Each firm attempts to maximize the difference between production and wages paid.
- There is competitive equilibrium, differentiation and no entry barrier among firms so that every level of skills will be employed by a firm.

4.2.2 Occupational choice

Suppose that the salary for performing the managing role with a skill level of k is given by $\pi(k)$ and the salary for performing the assistant role with a skill level of s is $w(s)$. Given an earnings schedule, an agent of skill (k, s) will evualate

$$\max\{\pi(k), w(s)\}$$

to decide what role to work in.
Competitive occupational choice is the economic model on which the person with skill set (k, s) will try to dedicate to the highest paying job, that is, manager if $\pi(k) > w(s)$ and assistant if $w(s) > \pi(k)$. The person will be indifferent between jobs if $\pi(k) = w(s)$. Observe that if π, w are given, then each worker can just evaluate it's own skill set and determine what job to do. Nevertheless, the earning schedules observed in the market will depend on the labor force R, on the production function F, on the distribution of both occupations and the possible matchings. This implies that the wages structures, π and w are not known a priori and have to be determined during the optimization process.

4.2.3 Competitive equilibrium for firms

Each firm will attempt to maximize $F(k,s) - \pi(k) - w(s)$ among the pair of skills it is able to employ. If there is competitiveness among firms, for each available skill set on the final distribution of occupations, there will exist a firm willing to employ that skill set pair, yielding the problem

$$\max_{(k,s)\in \mathrm{spt}(\mathcal{R})} F(k,s) - \pi(k) - w(s).$$

The objective of the model presented in this document is to couple occupational choice as in Section 4.2.2 with firms in a competitive equilibrium as in Section 4.2.3. In order to do this, we study the possible earning schedules in the optimals for both problems, for which it is convenient to recall the F-transform, a tool from optimal mass transport theory,

Definition 235. *(F-transform)*
Given $\pi : \Omega_1 \to \mathbb{R}$, we define π^F, the F-transform of π via

$$\pi^F(s) := \sup_{k\in\Omega_1} \{F(k,s) - \pi(k)\}. \tag{4.3}$$

Definition 236. *($\tilde{F}$-transform) Given $w : \Omega_2 \to \mathbb{R}$, we define π^F, the $\tilde{F}$-transform of π via*

$$w^{\tilde{F}}(s) := \sup_{s\in\Omega_2} \{F(k,s) - w(s)\}. \tag{4.4}$$

The difference between F-transform and $\tilde{F}$-transform is the set on which we maximize and the coordinate used, when there is no confusion on the domain of a function we use $\tilde{F}$ and F to denote the same transform, that is, we write w^F for an $\tilde{F}$-transform as long as there is no confusion that the domain of w is a subset of Ω_2.
From the definition of F-transform, we have $\pi(k) + \pi^F(s) \geq F(k,s)$ for any pair (k,s) on which π and π^F are defined. Note also that the definition of F-transform depends on the domain of the original function Ω_1.

Remark 237. *(On the definition of Ω_1 and Ω_2)*
The definitions of the F and $\tilde{F}$ transforms involve two arbitrary sets Ω_1 and Ω_2. We state it in this way for technical reasons but after the right Lemmas have been proved we will use the projections of $\mathrm{spt}(\mathcal{R})$ onto coordinates instead of Ω_1 and Ω_2, respectively. That is, after some mathematical reasults are obtained we will set $\Omega_1 = \mathrm{spt}(P_1\#\mathcal{R}), \Omega_2 = \mathrm{spt}(P_2\#\mathcal{R})$ where $P_1(x,y) = x, P_2(x,y) = y$.

Definition 238. *(Separation of wages)*
Given $\Omega_1, \Omega_2 \subseteq \mathbb{R}$, and functions $\pi : \Omega_1 \to \mathbb{R}, w : \Omega_2 \to \mathbb{R}$ where w is invertible in $\pi(\Omega_1)$ we say $\phi : \{k \in \Omega_1 : \pi(k) \in w(\Omega_2)\} \to \mathbb{R}$ separates π and w if for every $k \in \mathrm{Dom}(\phi)$,

$$\phi(k) = w^{-1}(\pi(k)). \tag{4.5}$$

The function that separates is called the separation function for (π, w). In the definition Ω_1, Ω_2 are just two subsets of $\mathbb{R}$ and the domain of the separation function is well-defined. If $w = \pi^F$, we denote $\phi = w^{-1} \circ \pi$ simply by ϕ_π when the domains are specified.

Lemma 239. *(Reformulation of wages and separation)*
Given $\pi : \Omega_1 \to \mathbb{R}, w : \Omega_2 \to \mathbb{R}$ strictly increasing functions, a function ϕ is the separation function for (π, w) from Definition 238 at $k \in \{\tilde{k} \in \Omega_1 : \pi(\tilde{k}) \in w(\Omega_2)\}$ if and only if $\max\{\pi(k), w(\phi(k))\} = \pi(k) = w(\phi(k))$.

Proof. For the first direction if $\phi = w^{-1}(\pi)$ (well-defined as w is strictly increasing), then $\max\{\pi(k), w(\phi(k))\} = \max\{\pi(k), w(w^{-1}(\pi(k)))\} = \max\{\pi(k), \pi(k)\} = \pi(k) = w(w^{-1}(\pi(k))) = w(\phi(k))$.
The reverse implication follows directly from the condition $\pi(k) = w(\phi(k))$ and the fact that w is invertible. ∎

Definition 240. *(Separation 1/2-cut)*
Let $\mathcal{R} \in \mathcal{P}_{ac,c}(\mathbb{R}^2)$ with density R, let (π, w) be a pair of strictly increasing functions as in Lemma 239, if ϕ is the separation function for (π, w), we say that ϕ 1/2-cuts $\mathcal{R}$ if

$$\int_{\mathbb{R}} \int_{-\infty}^{\phi(u)} R(u, v) dv du = 1/2. \tag{4.6}$$

Observe that this definition only makes sense when ϕ is defined in the correct domain (this technical point is dealt with in Assumption 242). In the general case, to avoid the use of ϕ and it's domain, given an earnings schedule (π, w) we say the earning's schedule 1/2-cuts $\mathcal{R}$ if

$$\int_{\mathbb{R}} \int \mathbf{1}_{\{\pi(u) \geq w(v)\}}(v) R(u, v) dv du = 1/2. \tag{4.7}$$

where $\mathbf{1}_{\{\pi(u) \geq w(v)\}}(v)$ denotes the indicator of the set of points v such that $\pi(u) \geq w(v)$ at level u.
The two definitions are equivalent via Lemma 239.
The 1/2-cut separation refers to the idea of splitting the mass exactly in half. In this case, ϕ separates the wages (by definition) and it's image splits the mass in halves. A separation function is interpreted as a divisory line between the two groups of the population. Before we set up the model, we need one more definition. The push-forward measure familiar in the context of optimal transportation of mass is a useful tool to conceptualize mass-balance.

Definition 241. *(Push-forward)*
Given a Borel measure ν and a borel function $f : \mathbb{R} \to \mathbb{R}$ the push-forward measure of ν by f denoted $f\#\nu$ is defined for every borel set A via

$$f\#\nu(A) := \nu(f^{-1}(A)).$$

To fix notation we write $P_1(x, y) = x, P_2(x, y) = y$ the coordinate projections and $||F||_\infty$ to denote the supremum over the set $\text{spt}(P_1\#R) \times \text{spt}(P_2\#R)$, that is

$$||F||_\infty := \sup_{(k,s) \in \text{spt}(P_1\#R) \times \text{spt}(P_2\#R)} |F(k, s)|$$

and $||\cdot||_\infty$ to denote the supremum of a function over it's domain.
As mentioned in Remark 237, from now on we always assume $\Omega_1 = \text{spt}(P_1\#R), \Omega_2 = \text{spt}(P_2\#R)$ for the definitions of F-transforms and separations (Definitions 235 and 238).

Remark 242. *(ϕ and it's domain)*
According to definition 238, in order to define ϕ we need to be able to evaluate $(\pi^F)^{-1}$ on $\pi(k)$. As stated, it is not a general condition nor (to our knowledge) something that can be derived from initial data $\mathcal{R}, F$. Because of the use of separation functions in literature we state the possibility of evaluation as an assumption, we explain in section 4.2.5 the consequences of this assumption.

Definition 243. *(Market-feasible separation)*
We say that a separation function ϕ (as in Definition 238) is market-feasible (with respect to $\mathcal{R}$) if $r_1, r_2 \in Dom(\phi)$ i.e. for the wages π, w that ϕ separates, there exist $s_1, s_2 \in \text{spt}(P_2\#\mathcal{R})$ satisfying

$$\pi(r_1) = w(s_1), \pi(r_2) = w(s_2). \tag{4.8}$$

Assumption 7. *(Connectedness of the projected measure and feasibility)*
We assume that there exist $r_1, r_2 \in \mathbb{R}$ such that $\text{spt}(P_1\#\mathcal{R}) = [r_1, r_2]$ and every earnings schedule π is market-feasible according to Definition 243.

Definition 244. *(Occupational distributions induced by separation)*
Let $\mathcal{R} \in \mathcal{P}_{ac,c}(\mathbb{R}^2)$ with density R, given a continuous, $\phi : \text{spt}(P_1\#R) \to \mathbb{R}$ the occupational distributions induced by ϕ are the measures whose distribution functions are given by the following formulas:

$$H^\phi(k) = \int_{-\infty}^{k} \int_{-\infty}^{\phi(\tilde{k})} R(s, \tilde{k}) ds d\tilde{k} \tag{4.9}$$

$$G^\phi(s) = \int_{-\infty}^{s} \int_{-\infty}^{\phi^{-1}(\tilde{s})} R(k, \tilde{s}) dk d\tilde{s}. \tag{4.10}$$

The measure associated to H^ϕ via $H^\phi(k) =: \mu_{H^\phi}((-\infty, k])$ is called the distribution of the labor force for the key occupation. The one associated to G^ϕ corresponds to the secondary job. To simplify notation we will not distinguish between H^ϕ and μ_{H^ϕ} and will write dH^ϕ instead of $d\mu_{H^\phi}$. Exactly, as before, according to Remark 242 it is not clear wether or not a separation function is defined everywhere on the domain of π, in the case where it is not we instead write

$$H^{(\pi,w)}(k) = \int_{-\infty}^{k} \int \mathbf{1}_{\{\pi(k) \geq w(s)\}} R(s, \tilde{k}) ds d\tilde{k} \tag{4.11}$$

$$G^{(\pi,w)}(s) = \int_{-\infty}^{s} \int \mathbf{1}_{\{\pi(k) \leq w(s)\}} R(k, \tilde{s}) dk d\tilde{s}. \tag{4.12}$$

To further simplify notation, in the case where $w = \pi^F$ we write $H^{(\pi,\pi^F)} = H^\pi$ and $G^{(\pi,w)} = G^\pi$.

Economic interpretation of distributions

Observe that if ϕ induces a 1/2-cut, this rewrites as $H^\phi(\mathbb{R}) = 1/2$ and in that case by Fubini's theorem one obtains $G^\phi(\mathbb{R}) = 1/2$ as well.

The idea is that $H^\phi(k)$ should be interpreted as the amount of population that dedicates to the key role having a skill lesser or equal than the value k. It is amount of workers willing to perform the key role under the salaries (π, w) which have skill at most k.

The objective of the Generalized Roy model is to incorporate the most economical matching of managers and assistants under competitive occupational choice and firm competitiveness.

4.2.4 Formulation of the model

Problem 5. *(Generalized Roy Model)*
Given $\mathcal{R} \in \mathcal{P}_{ac,c}(\mathbb{R}^2)$ with density $R : \text{spt}(\mathcal{R}) \to \mathbb{R}$, a strictly supermodular function $F : \mathbb{R}^2 \to \mathbb{R}$ the Generalized Roy Model is the following non-linear optimization program:

$$\sup_{(\phi,\pi,w,\mu)\in\mathcal{C}} \left\{ \int F(k, \mu(k)) dH^{\phi} \right\}. \tag{4.13}$$

where $\mathcal{C}$ is the set of quadruples (ϕ, π, w, μ) satisfying: $\pi, \phi \in C(\text{spt}(P_1\#R), \mathbb{R}), w \in C(\text{spt}(P_2\#R), \mathbb{R}), \mu : \mathbb{R} \to \mathbb{R}$ and

i. $\pi = w^F$,

ii. ϕ *separates* (π, w),

iii. ϕ *1/2-cuts* $\mathcal{R}$,

iv. $\mu\#H^{\phi} = G^{\phi}$ *(as measures according to Definition 241).*

v. $||\pi||_{\infty} \leq ||F||_{\infty}$

A quadruple $(\tilde{\phi}, \tilde{\pi}, \tilde{w}, \tilde{\mu}) \in \mathcal{C}$ is called an equilibrium for the Generalized Roy model if it achieves the supremum in (4.13).

The Generalized Roy model (5) differs from a usual Monge-Kantorovich optimal transportation problem in the sense that the optimization involves the generation of earning schedules directly, as π, w determine ϕ which in turn defines H^{ϕ} and G^{ϕ} which appear in the objective. This apparent circularity stops us from applying the theory of optimal transportation ([McCann-Guillen], [Villani2003]) directly. In section 4.2.6 we will show that the use of ϕ, π, w is somehow immaterial, as we can reduce it only to dependence on π, given that w and ϕ can be determined by only knowing $\pi, \mathcal{R}$ and F.
The idea to incorporate the pair (π, w) in the constraint set is motivated from the discussion in the beginning of section 4.2.2. The earning schedules are not exogenous. The original Roy model assumes ϕ to be linear, as one can see from the definition of occupational distributions in [Heckman-Honore]. This model makes no such assumption, motivating the name "Generalized Roy Model".

Detailed review of the model

The supremum in (5) attempts to maximize total production for a distribution of skills H^{ϕ}, at this point the production in a firm corresponds to matching a worker with skill k to be manager with a worker whose skill for the assistant role is $\mu(k)$. The matching function μ represents which worker is matched with whom to work together. The integral is computed with respect to H^{ϕ} as we have to consider the total produced from all workers that will dedicate themselves to the manager role. The total produced is the sum over all managers of the amount produced by the manager and the assistant matched with them.
The separation function ϕ will be shown to be avoidable in section 4.2.6 but has a significant economical interpretation: a person of skill k for the manager role, $\phi(k)$ is the skill for the assistant role needed for this person to be indifferent between being a manager or an assistant. That is, every worker whose skill set is of the form $(k, \phi(k))$ is indifferent between being a manager or an assistant, therefore one expects ϕ to be a certain kind of boundary separating both occupations.

4.2.5 The set of constraints

In this section we analyze and explain every constraint in the definition of the set $\mathcal{C}$ from the model (5). The objective is to give intuition on why each of this constraints is imposed and the implications they have on the assumptions made.

Wages are revenue-conjugates

Given a possible earning schedule (π, w) the condition that firms will have profit but this profit would potentially be zero in competitive equilibrium yields $F(k,s) \geq \pi(k) + w(s)$ from which we know that $F(k,s) - \pi(k) \geq w(s)$ which in turn yields $\pi^F(s) \geq w(s)$, as we will see in section 4.2.6, the objective function increases as w increases, so π^F is a feasible wage for the secondary role that increases the total output. A similar argument to the one in [Villani2003, Theorem 1.14], for duality of the Kantorovich problem allows us to reduce our search space to only F-conjugate pairs (functions that are F-transforms of each other).
In this definition, the use of w^{-1} involves an apparent hidden assumption that w is strictly increasing, nevertheless this presents no difficulties due to the following lemma:

Lemma 245. *(Strict supermodularity yields strictly increasing F-conjugate wages)*
If F is strictly supermodular as in Definition 4.1, then π^F is non-decreasing for every function π. Further if F is twice differentiable and π is continuous then π^F is strictly increasing.

Proof. Given $s > s'$ we have $\pi(k) + \pi^F(s) \geq F(k,s)$, consequently

$$\pi^F(s) \geq F(k,s) - \pi(k) > F(k,s') - \pi(k)$$

Taking the supremum yields $\pi^F(s) \geq \pi^F(s')$.
In the twice differentiable case, by envelope theorem one has

$$(\pi^F)'(s) = \partial_2 F(k^*(s), s) > 0$$

by the assumption of (4.2) where $k^*(s)$ attains the maximum from Definition 235 and continuity of π. ∎

Separation of wages

The definition of $\phi = w^{-1} \circ \pi$ helps to have a better interpretation. In section 4.2.6, the model is shown to be equivalent to a formulation without ϕ even though the term $w^{-1} \circ \pi$ is essential as it enforces occupational choice as shown in Lemma 239. The imposition of $w^{-1} \circ \pi$ to determine distributions of occupations enforces occupational choice. In this way, looking for optimal (π, w) will yield distributions that satisfy occupational choice in the sense of section 4.2.2.

Separation of wages and the technical assumptions

During the development of this work we realized that the definition of the separation of wages is somewhat unjustified for the modelling. It turns out that one can define a generalized roy model by looking only at salaries schedules (π, w) without ever defining ϕ. The use of the separation function is common in literature (see [Siow-Mak],[Siow-Mak2016],[Roy]). In order to obtain existence and

continuity results an assumption must be made. Either we impose a technical condition that ensures separation functions are well defined (Assumption 8 or Assumption 7 combined with Definition 243) or we impose a condition on the rate of growth of earning schedules (Assumption 9).
Assumptions 8 and 7 allow us to have a better economical interpretation and link our results with the ideas already present in literature. The assumption 9 allows us to show existence and uniqueness in other cases. We evaluate both possibilities throughout this work.

Separation cuts the labor force in half

The fact that ϕ (in defect $w^{-1} \circ \pi$) achieves a 1/2-cut of R represents the fact that exactly half of the workers will be managers and half will be assistants. If this constraint were not placed, one would not obtain a one-to-one map for matching. Models of many-to-one are discussed in section 4.7.

Mass Balance

The condition $\mu \# H^{\phi} = G^{\phi}$ ensures the matching is one-to-one, this allows each manager to have exactly one assistant associated and every firm to get a manager-assistant pair to hire.

4.2.6 The 2-step model

In this section we rewrite the Generalized Roy model (Problem 5) as a 2-step problem. We show that the model es equivalent to considering the wage structure as given, maximizing production via optimal matching and then maximizing over feasible earning schedules. This means that in competitive equilibrium, the determination of optimal wages happens in a way that is equivalent to the occupational choice given matching.

The condition of market-feasibility imposed via Definition 243 is the economic idea that the worst and best k-workers will be matched to someone instead of left to work by themselves. Mathematically, the concept allows separation functions to have the same domain and therefore be compared. An interesting line of research can be the relaxation or removal of this condition, where one would encounter the difficulty of multiple domains of separations. Under this setting, one would need to not compare separation functions in a pointwise matter (as we do later) but maybe the use of a different distance (like $L^p(P_1 \# R)$) would suffice. More on this is explained in Section 4.7.3.

Lemma 246. *(Market-feasibility and domains)*
Let $\phi : C(\text{spt}(P_1 \# \mathcal{R}) \to \mathbb{R}$, if ϕ is market-feasible (as in Definition 243) separation function for (π, w) a pair of continuous functions, then $Dom(\phi) = [r_1, r_2]$

Proof. Notice that $\phi = w^{-1} \circ \pi$, which means ϕ is continuous as a composition of continuous functions. Because $[r_1, r_2]$ is connected so is $\pi([r_1, r_2])$ and by continuity so is $w(\pi([r_1, r_2])$. Hence $Dom(\phi)$ is connected and $r_1, r_2 \in Dom(\phi)$ giving the result. ∎

Problem 6. *(2 step problem with explicit separation)*
Given a strictly supermodular function F and a labor force $\mathcal{R}$ as in (5), we define the 2-step program with explicit separation to be the non-linear problem:

$$\sup_{(\phi,\pi,w)\in\mathcal{C}_2} \left\{ \sup_{\mu \# H^{\phi} = G^{\phi}} \left\{ \int F(k, \mu(k)) dH^{\phi} \right\} \right\} \tag{4.14}$$

where $\mathcal{C}_2$ is the set of triples (ϕ, π, w) of continuous functions, $\pi : \text{spt}(P_1 \# R) \to \mathbb{R}, w : \text{spt}(P_2 \# R) \to \mathbb{R}, \phi : \mathbb{R} \to \mathbb{R}$ and

i. $\pi = w^F$,

ii. $\phi = w^{-1} \circ \pi(k)$,

iii. ϕ *1/2-cuts* R,

iv. $||\pi||_\infty \leq ||F||_\infty$,

Problem 7. *(2 step problem for earnings schedule)*
Given a strictly supermodular function F and a labor force $\mathcal{R}$ as in (5), we define the 2-step program to be the non-linear problem:

$$\sup_{\pi \in \mathcal{C}_3} \left\{ \sup_{\mu \# H_\pi = G_\pi} \left\{ \int F(k, \mu(k)) dH_\pi \right\} \right\} \tag{4.15}$$

where $\mathcal{C}_3$ is the set of continuous functions $\pi : \text{spt}(P_1 \# R) \to \mathbb{R}$ such that

i. There exists $w : \text{spt}(P_2 \# \mathcal{R}) \to \mathbb{R}$ satisfying $\pi = w^F$,

ii. $(\pi^F)^{-1} \circ \pi$ *1/2-cuts* R,

iii. $||\pi||_\infty \leq ||F||_\infty$

here H_π and G_π are the induced measures from Definition 244 using $(\pi^F)^{-1} \circ \pi$ as separation function.

Observe that the only difference between Problem 6 and Problem 7 is that ϕ is explicit in the former but not in the latter. This technicality is essential to note that the set of constraints is not on really triples as one may expect from looking at Problem 6 but only in the wage structure as it is evident in Problem 7. The fact that Problem 6 and 7 are equivalent is evident by the definition of ϕ in both cases.
We turn our attention to the relation between these two problems and the generalized Roy model (Problem 5).

Equivalence

Theorem 247. *(Equivalence of the problems)*
Given a strictly supermodular production function F and a labor force $\mathcal{R} \in \mathcal{P}_{ac,c}(\mathbb{R}^2)$, Problem 5, Problem 6 and Problem 7 are equivalent, i.e.

$$\begin{aligned} \sup_{(\phi,\pi,w,\mu) \in \mathcal{C}} \left\{ \int F(k, \mu(k)) dH^\phi \right\} &= \sup_{(\phi,\pi,w) \in \mathcal{C}_2} \left\{ \sup_{\mu \# H^\phi = G^\phi} \left\{ \int F(k, \mu(k)) dH^\phi \right\} \right\} \\ &= \sup_{\pi \in \mathcal{C}_3} \left\{ \sup_{\mu \# H_\pi = G_\pi} \left\{ \int F(k, \mu(k)) dH^\phi \right\} \right\} \end{aligned}$$

Proof. It is clear that Problem 6 and Problem 7 are equivalent so it is enough to show that Problem 5 and 7 are equivalent. Given $(\phi, \pi, w, \mu) \in \mathcal{C}$, clearly $\pi \in \mathcal{C}_3$ as w satisfies the constrain $w^F = \pi$, as it is imposed in $\mathcal{C}$. Again, feasibility means $\mu \# H_\pi = G_\pi$ as $H_\pi = H^\phi$ by definition. Hence,

$$\begin{aligned}\int F(k, \mu(k)) dH^\phi &\leq \sup_{\mu \# H_\pi = G_\pi} \left\{ \int F(k, \mu(k)) dH^\phi \right\} \\ &\leq \sup_{\pi \in \mathcal{C}_3} \left\{ \sup_{\mu \# H_\pi = G_\pi} \left\{ \int F(k, \mu(k)) dH_\pi \right\} \right\}\end{aligned}$$

As this happens for every $(\phi, \pi, w, \mu) \in \mathcal{C}$ taking the supremum on $\mathcal{C}$ yields that the supremum in Problem 5 is bounded above by the suprema in Problem 7.
For the reverse inequality, take $\pi \in \mathcal{C}_3$, then there exists w with $w^F = \pi$ and set $\phi = w^{-1} \circ \pi$, well defined as noted in Lemma 245, then $(\phi, \pi, w, \mu) \in \mathcal{C}$ and hence

$$\int F(k, \mu(k)) dH_\pi \leq \sup_{(\phi, \pi, w, \mu) \in \mathcal{C}_3} \left\{ \int F(k, \mu(k)) dH^\phi \right\}$$

Taking the suprema in the order of Problem 7 yields the result. ∎

Remark 248. *Although the proof is relatively simple, the value of Theorem 247 is 2-fold: firstly, it simplifies the problem of the Generalized Roy Model into a two-step program on which we can identify a Monge-Kantorovich optimal transport problem in usual form in the inner problem and secondly it provides the sanity check that it is indeed the same to think about the matching given the earning schedules as our intuition predicts.*

Interpretation

Theorem 247 allows us to conclude that the generalized Roy model equilibrium yields the same equilibrium of looking at the occupational choice problem and then optimizing over possible earning schedules. The original formulation on the quadruple does not allow to apply the results developed over the last decades from the theory of transportation, while the 2-step reformulation does. In the community where workers are deciding between being managers or assistants, they can plan by finding their optimal match first for every earning schedule and then finding the earning schedule that maximizes the total output. In this way, each individual could potentially plan for their own occupational choice, knowing that maximizing over earnings will yield the "simultaneous" equilibrium from problem 5.
The main motivation for establishing Theorem 247 is that we can now make use of the framework developed in recent years in the study of the Monge-Kantorovich problem.

Duality

In order to take advantage of the technicality provided by Theorem 247, we start by rewriting the duality theorem for the Monge problem, as in [Villani2009] Theorem 5.3

Theorem 249. *(Kantorovich Duality)*
Given F strictly supermodular and $\mathcal{R} \in \mathcal{P}_{ac,c}(\mathbb{R}^2)$, let $\pi : \text{spt}(P_1 \# R) \to \mathbb{R}$ be given and suppose

$w^F = \pi$ *a.e. for some* $w : \text{spt}(P_2 \# R) \to \mathbb{R}$*, if we set* $\phi = w^{-1} \circ \pi$ *we have*

$$\sup_{\mu \# H^\phi = G^\phi} \int F(k, \mu(k)) dH^\phi = \sup_{\gamma \in \Gamma(H^\phi, G^\phi)} \left\{ \int F(k,s) d\gamma \right\} \tag{4.16}$$

$$= \inf_{\varphi \in C(\text{spt}(P_1 \# R))} \left\{ \int \varphi dH^\phi + \int \varphi^F dG^\phi \right\} \tag{4.17}$$

where $\Gamma(H^\phi, G^\phi)$ *is the set of measures in the product space with marginals* H^ϕ *and* G^ϕ *respectively.*

For a proof see [Villani2009, Theorem 5.3].

Existence and Uniqueness

Theorem 250. *(Optimality on Monge-Kantorovich)*
The F*-optimal transport map* μ *for the Monge-Kantorovich problem between* ν_1 *and* ν_2 *satisfies*

$$\pi(k) + \pi^F(\mu(k)) = F(k, \mu(k)) \ \nu_1 - a.e. \tag{4.18}$$

For a proof see [Villani2009, Theorem 5.10].
In this section we explore whether the problem 5 has a unique solution or not. We do this by exploiting the knowledge of existence and uniqueness on the inner problem of Problem 6 and then appealing to Theorem 247.

Theorem 251. *(Existence)*
Under either the assumption 8 or assumption 9, if F *is twice differentiable and super-modular and* $\mathcal{R} \in \mathcal{P}_{ac,c}(\mathbb{R}^2)$ *then Problem 5 has a solution.*

Before we write the proof of the Theorem we need some Lemmata.

Lemma 252. *(*F*-transforms are equi-Lipschitz)*
Let π, w *be such that* $\pi = w^F$*, then* π *is Lipschitz with Lipschitz constant at most* $\sup_y |D_x F(x,y)|$.

For a proof see [McCann-Guillen, Lemma 3.1].

Lemma 253. *(Pointwise uniform bound)*
The set $\mathcal{C}_3$ *is pointwise uniformly bounded.*

Proof. Take $\pi \in \mathcal{C}_3$ by definition of the set $\mathcal{C}_3$, $\pi(x) \leq ||F||_\infty$ everywhere on the domain of π. This bound is uniform as it does not depend on π nor w. ∎

Lemma 254. *(*$\pi \to w$ *continuity)*
Let $\pi, \tilde{\pi} \in \mathcal{C}_3$ *with* $w^F = \pi$ *and* $\tilde{w}^F = \tilde{\pi}$*, for every* $\epsilon > 0$ *there exists* $\delta > 0$ *such that* $||\pi - \tilde{\pi}||_\infty < \delta$ *then* $||w - \tilde{w}||_\infty < \epsilon$

Proof. Given any $y \in \text{spt}(P_2 \# R)$, note that

$$\sup_x \{F(x,y) - \pi(x)\} - \sup_x \{F(x,y) - \tilde{\pi}(x)\} \leq \sup_x \{\tilde{\pi}(x) - \pi(x)\} \leq ||\pi - \tilde{\pi}||_\infty$$

So setting $\delta = \epsilon$ finishes the proof. ∎

Observe that the proof of Lemma 254 indicates that $\text{Im}(\tilde{\pi}^F) \subseteq \{w_1 \in \mathbb{R} : |w - w| < \delta, w \in \text{Im}(\tilde{\pi})\} =: \text{Im}(\pi^F)^\delta$. According to Remark 242, separation functions may not share a full domain. The fact that two different separation functions can not be compared in supremum norm is a technical liability.Observe that this assumption is implicit in the formulation of the model by [Roy] and [Siow-Mak]. (See for example the definition of separation function on [Siow-Mak]).

Assumption 8. *(Uniformity of Domains)*
Assume that for every $\pi \in \mathcal{C}_3$, if ϕ is the separation function induced by (π, w) then $Dom(\phi) = \text{spt}(P_1 \# R)$, i.e. for every $k \in \text{spt}(P_1 \# R)$ there exists (a unique) $s \in Dom(w)$ such that $\pi(k) = w(s)$.

The assumption 8 will allow us to show continuity of the Roy Model in the sense that small changes in the earning's schedule will correspond to small changes in the solution. This continuity is done via the $||\cdot||_\infty$ norm for which we need to be able to compare separation functions everywhere. One can argue that the assumption can be avoided by the introduction of a different norm to evaluate the differences of separation functions (for example an $L^p_{P_1 \# R}$ norm, we leave this for future work or other reasearchers and explain further details in section 4.7.

Remark 255. *(Assumption 7 and Definition 243 $\Rightarrow$ 8)*
Observe that Lemma 246 shows that under Assumption 7 if ϕ is market-feasible (by definition 243), the domain of all separation functions is $[r_1, r_2]$ which in particular yields Assumption 8.

4.2.7 The restricted version of the problem

Whenever we add the Assumption 8 to the generalized Roy Model we call it the restricted Generalized model. We note also that this may not be the only way to avoid such problem. We observe that this assumption on the earning schedules allows us to study general economies and general populations, nevertheless making assumptions on the population or the production function can lead to similar conclusions via different techniques. We explore the idea of restricting the production functions further to not assume the uniform domains in section 9.

Lemma 256. *(Continuity of the inverse of w)*
Under Assumption 8, let $\pi \in \mathcal{C}_3$ and w such that $w^F = \pi$, for every $\epsilon > 0$ there exists $\delta > 0$ such that every $\tilde{w} \in B_\delta^{||\cdot||_\infty}(w)$ such that

$$||\tilde{w}^{-1} - w^{-1}||_\infty < \epsilon. \tag{4.19}$$

Proof. Given $\epsilon > 0$ let δ be the one from the uniform continuity of w, if y is fixed in $\text{Range}(w)$, note that by $\tilde{w} \in B_\delta(w)$ we get

$$w^{-1}(y - \delta) \leq \tilde{w}^{-1}(y) \leq w^{-1}(y + \delta)$$

Substracting $w^{-1}(y)$ in all terms we get

$$w^{-1}(y - \delta) - w^{-1}(y) \leq \tilde{w}^{-1}(y) - w^{-1}(y) \leq w^{-1}(y + \delta) - w^{-1}(y)$$

By definition of δ the suprema on both bounds approaches 0 as $\epsilon \to 0$ which yields the result as the definition of D in terms of w and $\tilde{w}$ ensures existence of the inverses. ∎

Lemma 257. *($\pi \to \phi$ continuity)*
Under Assumption 7, given $\epsilon > 0$ and $\pi \in \mathcal{C}_3$, there exists $\delta > 0$ such that $||\pi - \tilde{\pi}||_\infty < \delta$ implies $||\phi_\pi - \phi_{\tilde{\pi}}||_\infty < \epsilon$ where ϕ_π and $\phi_{\tilde{\pi}}$ are from Definition 238.

Proof. Given $\pi, \tilde{\pi} \in \mathcal{C}$ and $w^F = \pi, \tilde{w}^F = \tilde{\pi}$ by triangle inequality,

$$|w^{-1}(\pi(x)) - \tilde{w}^{-1}(\tilde{\pi}(x))| \leq |w^{-1}(\pi(x)) - w^{-1}(\tilde{\pi}(x))| + |w^{-1}(\tilde{\pi}(x)) - \tilde{w}^{-1}(\tilde{\pi}(x))|.$$

Given $\epsilon > 0$ we define $\delta < \min\{\delta_1, \delta_2\}$ where δ_1 is from uniform continuity of w^{-1} and δ_2 from Lemma 256. If $\pi(x)$ belongs to the image of $\tilde{\pi}$, using assumption 7. ∎

Lemma 258. *(d_2-continuity of the split measures)*
Take $(\phi, \pi, w), (\tilde{\phi}, \tilde{\pi}, \tilde{w})$ satisfying the constraints of (4.2.6) if we assume 8 then for every $\epsilon > 0$ there exists $\delta > 0$ such that if

$$||\phi - \tilde{\phi}||_\infty < \delta$$

then

$$d_2(H^\phi, H^{\tilde{\phi}}) < \epsilon.$$

Proof. Note that we can simply compute the difference of integrals

$$\left| \int_{-\infty}^{k} \int_{-\infty}^{\phi(\hat{k})} R(\hat{k}, s) ds d\hat{k} - \int_{-\infty}^{k} \int_{-\infty}^{\tilde{\phi}(\hat{k})} R(\hat{k}, s) ds d\hat{k} \right|$$
$$\leq ||R||_{\infty, R} \left(||\phi - \tilde{\phi}||_{\infty, R_1} diam(R_1) \right)$$

where the ∞, R norms are the suprema on $\text{spt}(\mathcal{R})$ and over the projection to the first coordinate which are compact sets so continuity of ϕ and R yield an uniform estimate on accumulation functions of H^ϕ and $H^{\tilde{\phi}}$. Because $\text{spt}(\mathcal{R})$ is compact H^ϕ and $H^{\tilde{\phi}}$ also have compact support so the L_∞ bound gives an L^1 bound which translates to a d_1-bound for the measures and by compactness of $\text{spt}(\mathcal{R})$ again we obtain the desired d_2 bound. See [Villani2003] Chapters 2 and 8 for the $d_1 - d_2$ relation under compactness. ∎

4.2.8 What this restriction does

As explained in the previous section, the assumption 8, of uniformity of domains is not inherent to the original model of Roy. It is a technical hypothesis to reconstruct continuity using the supremum norm. We formulate here another assumption and explore conditions that result on such an assumption that would yield a $\pi \to H^\pi$ continuity as Lemmata 257 and 258. Here we avoid the use of ϕ and it's domain by using the second equivalent definition of H^π given in definition 244.

4.2.9 Restrictions on production functions and populations instead

We have discussed Assumptions 7 and 8 as they played a significant role in previous versions of the Roy model ([Siow-Mak], [Roy]). This assumption has been helpful to obtain continuity of the problem. In this section we explore a different assumption that can be inferred more directly from initial data and is also sufficient for the continuity conditions as in Lemma 258. In this section we focus on the formulation of the measures (244) that don't depend on a separation function (4.11). The idea is that a uniform lower bound on derivatives of earning schedules for the secondary role yields the same continuity estimates but we are able to obtain such bounds in more general situations where ϕ may not be everywhere defined.

Assumption 9. *(Uniform lower bound on F-transforms via F)*
We assume that there exists a positive constant $C > 0$ such that for every $\pi \in \mathcal{C}_3$

$$(\pi^F)'(s) \geq C. \tag{4.20}$$

for every s for which the derivative is defined.

Economically, assumption 9 says that the rate of change of salaries of the secondary role is bounded for all possible salaries.

Lemma 259. *($\pi \to H^\pi$-continuity)*
For every $\epsilon > 0$, there exists $\delta > 0$ such that if $||\pi - \tilde{\pi}||_\infty < \delta$ then $d_2(H^\pi, H^{\tilde{pi}}) < \epsilon$

Proof. By Lemma 252, we know π^F is a.e. differentiable therefore We claim for any $\epsilon > 0$, there exists a $\delta > 0$ s.t. if $||\pi - \tilde{\pi}||_\infty < \delta$ then $||f_{H_\pi} - f_{H_{\tilde{\pi}}}|| < \epsilon$ where f with a subscript refers to the density. Let $\epsilon > 0$ be given, first note that $H_\pi - H_{\hat{k}}$ can be written as the integral over $\hat{k}$ of the skill density R times the difference in the two indicator functions of he sets where $\pi(\hat{k}) > \pi^F(s)$ and $\tilde{\pi}(\hat{k}) > \tilde{\pi}^F(s)$ just as in equation (4.11).This difference is nonzero only when exactly one condition is true at a given point. Since for a given $\hat{k}$, the left hand sides in the indicator functions in (4.11) are fixed and the right hand sides are increasing, the integrand is nonzero on an interval (s_0, s_1). Without loss of generality suppose that at s_0 we have $\pi(\hat{k}) = \pi^F(s_0)$ but $\tilde{\pi}(\hat{k}) > \tilde{\pi}^F(s_0)$. In this case,

$$\begin{aligned}\tilde{\pi}(\hat{k}) - \tilde{\pi}^F(s_0) &\leq ||\tilde{\pi} - \pi||_\infty + |\pi(\hat{k}) - \tilde{\pi}^F(s_0)| \\ &= ||\tilde{\pi} - \pi||_\infty + |\pi^F(s_0) - \tilde{\pi}^F(s_0)| \leq ||\tilde{\pi} - \pi||_\infty + ||\pi^F - \tilde{\pi}^F||_\infty \leq 2||\tilde{\pi} - \pi||_\infty\end{aligned}$$

Therefore if $\tilde{\pi}^F$ has derivative bounded below by $C > 0$, we will have

$$\tilde{\pi}(\hat{k}) - \pi^F(s_0) \leq 2||\tilde{\pi} - \pi||_\infty \leq \tilde{\pi}^F(s_0 + 2||\tilde{\pi} - \pi||_\infty/C) - \tilde{\pi}^F(s_0),$$

which means

$$\tilde{\pi}(\hat{k}) \leq \tilde{\pi}^F(s_0 + 2||\tilde{\pi} - \pi||_\infty/C)$$

i.e. $s_1 \leq s_0 + 2||\tilde{\pi} - \pi||_\infty/C$. Therefore,

$$||H_\pi - H_{\tilde{\pi}}||_\infty \leq \int_{s_0}^{s_1} R(\hat{k}, s)ds \leq 2C||R||_\infty \cdot ||\tilde{\pi} - \pi||_\infty$$

a bound independent of $\hat{k}$. Therefore we can choose δ to be $\epsilon C/2||R||_\infty$. ∎

The previous proof seems rather unintuitive because we are only using the low-regularity estimate of π and $\tilde{\pi}$ being differentiable almost everywhere, observe that by the mean value theorem if every element of $\mathcal{C}_3$ were continuously differentiable we could obtain the same conclusion by mean value theorem:

$$|s_1 - s_2| = \left|\frac{s_1 - s_2}{\pi^F(s_1) - \pi^F(s_2)}\right|\pi(s_1) - \pi(s_2)| < \frac{||\pi - \tilde{\pi}||_\infty}{C} < \delta/C$$

which yields the $\pi \to H^\pi$ continuity by the same argument as above. We specifically don't assume that the elements in $\mathcal{C}_3$ belong to $C^1(\text{spt}(P_1\#\mathcal{R})$ as the regularity we use is inherent from Lemma 252.

Theorem 260. *(Stability of optimal Transport)*
Let $F_n \to F$ uniformly, where each F_n is as in Definition 4.1 and satisfying the assumption (4.2), and $\{\rho_n\}, \{\nu_n\}$ two sequences of probability measures converging weakly to μ and ν respectively, suppose that there exists an F_n-optimal transport map T_n between ρ_n and ν_n and assume there exists an optimal transport map T between ρ and ν then as $n \to \infty$,

$$\int F(k, T_n(k)) d\rho_n \to \int F(k, T(k)) d\rho$$

For a proof see [Villani2009, Theorem 5.20].

Corollary 261. *(Stability of optimal transport maps)*
Assume that $\{F_k\}_{k\in\mathbb{N}}$ is a sequence of production functions which are supermodular and twice continuously differentiable and so is F, such that $F_k \xrightarrow{||\cdot||_\infty} F$, let $\nu_n \xrightarrow{d_2} \nu$ and $\rho \in \mathcal{P}(\mathbb{R})$ be fixed, then the F_k-optimal transport map μ_k between ρ and ν_k converges in ρ-probability to μ, the unique F-optimal transport map between ν and ρ, that is for every $\epsilon > 0$ we have

$$\rho\left(\{k : |\mu(k) - \mu_n(k)| > \epsilon\}\right) \xrightarrow{n\to\infty} 0. \tag{4.21}$$

For a proof see [Villani2009, Corollary 5.23]. With all the Lemmata in place, now we can write a proof for Theorem 251.

Proof of Theorem 251. Inner problem has a unique solution via Brenier's Theorem, the map

$$\pi \to \sup_{\mu \# H_\pi = G_\pi} \left\{ \int F(k, \mu(k)) dH_\pi \right\}$$

is continuous in the uniform topology under the assumptions 8 or 9 via Lemmatta 258 or 259 respectively. Let us show that $\mathcal{C}_3$ is nonempty. Let $D_c := \{w \in C(\text{spt}(P_2 \# R)) : ||w||_\infty \leq c\}$ and define A to be $\mathcal{C}_3$ without the condition that the function 1/2-cuts R , that is $A = \{\pi : \text{spt}(P_1 \# R) \to \mathbb{R} : \exists w, \pi = w^F, ||\pi||_\infty \leq ||F||_\infty\}$. D_c is convex and therefore connected, furthermore it maps continuously to a subset P_c of A via the F transform as long as c is small enough (according to Lemma 254). Thus P_c is connected. Also note that if $w = 0 \in D_c$ is chosen, then $\pi(k) = \sup_{s\in\text{spt}(P_2\#R)} F(k, s)$, and so the resulting $H^\phi(\mathbb{R}) = 1$ and if $w = \sup_{k\in\text{spt}(P_1\#R)} F(k, s) \in D_c$ then $\pi(k) = \sup_{s\in spt(P_2\#R)} F(k, s) - F(k^*, s)$ where k^* is the maximum of k in $\text{spt}(P_1 \# R)$. Then $H^\phi = 0$ on $\mathbb{R}$. But H^ϕ is a continuous function on A (according to Lemmata 257 and 258), hence we have exhibited two functions in a connected subset P_c which map to 0 and 1. Therefore there exists a function in P_c which maps to 1/2 and is therefore an element of $\mathcal{C}_3$. Now by Lemma 252 the set $\mathcal{C}_3$ is equi-Lipschitz and by Lemma 253 point-wise bounded and so by Arsela-Ascoli it is relatively compact in the uniform topology and hence achieves the supremum by extreme value theorem. ∎

Observation 262. *A simpler version of the intermediate value theorem can be used when more restrictions on the revenue function are imposed. If F satisfies the (A3S) condition from [McCann-Guillen], then by [McCann-Guillen, Theorem 5.1] the set of $F-$convex functions is convex itself and intermediate value theorem van be applied in a much simpler way. In general the (A3S) condition is a differential equation on the fourth partials of F which is a-priori not assumed here but yields regularity of optimal transport maps, see [McCann-Guillen] or [Villani2009] for more details.*

First order conditions for optimality

In this section we explore how the optimal quadruples on Problem 5 depend on the problem data. In this section we assume all variables are continuously differentiable even though form some variables as π, w only almost everywhere differentiability is ensured by Lemma 252. We assume continuously differentiable throughout but most of the results can be generalized by the concept of approximate differentiability.

Proposition 263. *Let F and R satisfy assumptions of Problem 5, then for the optimal quadruple it holds*

$$\phi'(k) = \frac{F_1(k,\mu(k)) + F_2(k,\mu(k))\mu'(k) - w'(\mu(k))\mu'(k))}{w'(w^{-1}(F(k,\mu(k)) - w(\mu(k))))}. \tag{4.22}$$

where F_1 and F_2 denote the partials with respect to the first and second coordinate respectively.

The proposition is a direct computation. But the economical interpretation of this derivative is particularly interesting: the optimal salary paid in the labor market for workers with skill k depends not only on the production achieved by their optimal matching but also on how difficult it is to change the pairs (μ') at a certain level. Note that one can find this computation in [Siow-Mak].

Proposition 264. *Let F and R satisfy assumptions for Problem 5, then for the optimal quadruple it holds that*

$$\pi'(k) = F_1(k,\mu(k)) \tag{4.23}$$

$$w'(s) = F_2(\mu^{-1}(s), s) \tag{4.24}$$

for $k \in \operatorname{spt}(H)$ and $s \in \operatorname{spt}(G)$. If there exists an interval $[a,b] \supseteq \operatorname{spt}(H)$ on which μ is defined Lebesgue a.e., we may integrate to obtain the following expression for the wages:

$$\pi(k) = c_k + \int_a^k F_1(\hat{k}, \mu(\hat{k}))d\hat{k} \tag{4.25}$$

where c_k is an integration constant. The analogous result holds for $w(s)$. For example, this condition holds for k if the projection of $\operatorname{spt}(\mathcal{R})$ to the k-axis is an interval and the "lower envelope", defined as $\min\{s : (k,s) \in \operatorname{spt}(\mathcal{R})\}$ is non-increasing in k.

Proof. We prove this for $\pi(k)$ as the result for w follows by symmetry. Recall we have $\pi(k) = \sup_{s \in \operatorname{spt}(G)} \{F(k,s) - w(s)\}$ for $k \in \operatorname{spt}(H)$ by duality of wages. By (4.18) this supremum is attained by $\mu(k)$. Since $\pi(k)$ is differentiable and $F(k,s) - w(s)$ is differentiable with respect to k, the envelope theorem implies that $\pi'(k) = F_1(k,\mu(k))$. The result follows by integrating. The last condition implies that $\operatorname{spt}(H)$ is an interval. Since ϕ is strictly increasing. ∎

4.2.10 An optimality conjecture

The following conjecture arises from discussions with Dr. Siow and the fact that the 2-step problem rewrites the model in [Siow-Mak2016]. The 2-step problem is expected to be wage-optimal in the sense that if agents decide after they see wages and wages are maximal, those should correspond to the actual earnings. In our context, we formulate this economical idea as follows.

Conjecture 1. *(Super optimality of wages)*
The optimal wages in the generalized Roy Model and the 2-step problem are the Kantorovich potentials of the inner problem of Problem 7, i.e. if $\pi^* \in \mathcal{C}_3$ realizes the supremum of Problem 7 then

$$\max_{\pi \in \mathcal{C}_3} \left\{ \max_{\mu \# H^\phi = G^\phi} \int F(k, \mu(k)) dH^\phi \right\} = \int \pi^* dH_{\pi^*} + \int (\pi^*)^F dG_{\pi^*} \tag{4.26}$$

Observation 265. *Note that, by duality (4.17), we automatically obtain*

$$\max_{\pi \in \mathcal{C}_3} \left\{ \max_{\mu \# H^\phi = G^\phi} \int F(k, \mu(k)) dH^\phi \right\} \leq \int \pi^* dH_{\pi^*} + \int (\pi^*)^F dG_{\pi^*},$$

as π is a viable candidate for the inner minimization problem. The trickier part (if the conjecture is true) is the reverse inequality. One can attempt to define a map $\Theta : \mathcal{C}_3 \to C_3$ corresponding to the Kantorovich potential of a given element (with the appropriate constant to make a $1/2$-cut). If the map Θ happens to be a contraction (in the appropriate Banach space), then one can use a fixed-point argument, as one could use the consecutive iterations of Θ starting from π^* to obtain a limit which increases the objective function and hence should coincide with π^*. We haven't been able to show the map is a contraction, not even assuming convexity of $\mathcal{C}_3$ which one can derive from general assumptions of F (such as the cross-curvature condition) as [McCann-Guillen, Theorem 5.1]. By expanding both terms on (4.26), it is equivalent to show that if $\pi^* \in \mathcal{C}_3$ is optimal and ϕ is it's corresponding optimal earning schedule for role k, then

$$\int \int \max\{\pi(k), \pi^F(s)\} dR(k,s) \leq \int \int \left(\mathbf{1}_{\{\pi^F(s) \leq \pi(k)\}} \phi(k) + \mathbf{1}_{\{\pi(k) < \pi^F(s)\}} \phi^F(s) \right) dR(k,s),$$

where $\mathbf{1}_C$ is the indicator function of the set C.
Now clearly, it would be sufficient to show that R-a.e. with respect to (k, s),

$$\max\{\pi(k), \pi^F(s)\} \leq \left(\mathbf{1}_{\{\pi^F(s) \leq \pi(k)\}} \phi(k) + \mathbf{1}_{\{\pi(k) < \pi^F(s)\}} \phi^F(s) \right),$$

which we haven't been able to show.

4.3 Examples of the Generalized Roy Model

4.3.1 Non-homogeneous degree 1 Cobb-Douglas production

The generalized Roy model (Problem 5) allows many different types of markets and behaviours. In this section we simplify the problem by looking at a specific production function that depends non-linearly in both skills k and s but the interaction term is homogeneous of degree 1.

Assumption 10. *We assume that there exist strictly increasing, continuously differentiable functions $a, b : \mathbb{R} \to \mathbb{R}$ and a constant $c \geq 0$ such that*

$$F(k, s) = a(k) + b(s) + cks. \tag{4.27}$$

By Proposition 264, we have

$$\begin{aligned} \pi'(k) &= a'(k) + c\mu(k) \\ w'(\mu(k)) &= b'(\mu(k)) + ck. \end{aligned} \tag{4.28}$$

on spt(H). Changing variables to $s = \mu(k)$ and integrating we obtain

$$\begin{aligned} \pi(k) &= a(k) + c\int_0^k \mu(\tilde{k})d\tilde{k} + e \\ w(s) &= b(s) + c\int_0^s \mu^{-1}(\tilde{s})d\tilde{s} + d \end{aligned} \tag{4.29}$$

for constants e, d. To figure out e and d we plug in the lowest matching $(0, \mu(0))$ whenever $0 \in \text{spt}(e_1 \# \mathcal{R})$ we obtain

$$e + d + \int_0^{\mu(0)} \mu^{-1}(\tilde{s})d\tilde{s} = 0.$$

Using the separation of wages, note that by structure of Brenier's map μ is positive assortative from which the integral term vanishes and then $e + d = 0$, from the optimality condition on (ϕ, μ, π, w) we know that $\phi = w^{-1} \circ \pi$ to find e we use that ϕ 1/2-cuts $\mathcal{R}$, meaning that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\phi(k)} R(k,s)dsdk = \frac{1}{2} = \int_{-\infty}^{\infty}\int_{-\infty}^{\phi^{-1}(s)} R(k,s)dkds.$$

Our goal is to explore the different regimes of the production function as c varies, for which it is important to start by understanding the model for the case with no interaction.

4.3.2 No interaction

In the case where there is no interaction ($c = 0$) above, $F(k,s) = a(k) + b(s)$ let us analize the separation function ϕ. The separation function determines the wages being paid to the workers. For a worker of skill k, the value $\phi(k)$ represents the salary at which the worker would be indifferent between working in the primary or the secondary job.
Given a function $\pi : \mathbb{R} \to \mathbb{R}$, define H^ϕ and G^ϕ as in Definition (244). Observe that for any function μ with $\mu \# H^\phi = G^\phi$ we have

$$\int F(k, \mu(k))dH^\phi = \int a(k)dH^\phi + \int b(s)dG^\phi$$

which means that every function μ that satisfies $\mu \# H^\phi = G^\phi$ yields the same production output. *So the matching does not affect production*, we can choose any volume-preserving map when ϕ is fixed. In this case (4.29) reduce to

$$\begin{aligned} \pi(k) &= a(k) + e \\ w(s) &= b(s) + d \end{aligned} \tag{4.30}$$

Notice then that $w^{-1}(y) = b^{-1}(y-d)$ and so $\phi(k) = b^{-1}(\pi(k)-d) = b^{-1}(a(k)+e-d) = b^{-1}(a(k)+2e)$. This implies that the slope of ϕ in the $a(k) - b(s)$-axis is 1. The separating function on the skill value-skill value plane is a straight line, in other words, in equilibrium workers are paid linearly with respect to their skill sets and how much their skill set produces, independent of other workers abilities and matches. So in the context where production is modelled to not be improved by the relationship between manager and assistant, it doesn't matter who they match, they'll be paid as much as they can generate for production. If you are good enough worker, the person you match should not affect your wage. *In this model, even if your manager is incompetent you can still make it.*

4.3.3 Pure interaction

Now we look into the complete opposite case on which production depends only on the interaction between workers. This models labor markets where the relationship between primary and secondary job is complementary, *in this context you can not do your job well (in terms of producing more) if you can't work efficiently with your coworker. The better you and your coworker get along, the more you produce together.*

Assumption 11. *In this case production is totally complementary, i.e.* $a(k) = 0, b(s) = 0$ *and there exists* $c > 0$ *such that* $F(k, s) = cks$.

In this case given ϕ, the inner problem of (4.2.6) corresponds to the optimal transport problem with cost $c(k, s) = |k - s|^2$ because

$$\int ck\mu(k)dH^\phi = \frac{c}{2}\left(\int |k|^2 dH^\phi + \int |s|^2 dG^\phi - \int |k - \mu(k)|^2 dH^\phi\right)$$

and once ϕ is fixed, the first two terms are constant. For this part we can think of ϕ as being fixed, obtain conditions on μ and the afterwards optimize on ϕ thanks to Problem 6. Again by Brenier's theorem we know that

$$\mu = (G^\phi)^{-1} \circ H^\phi \tag{4.31}$$

where in this context we denote by G^ϕ and H^ϕ the accumulation functions, simplifying Problem 6 to

$$\max_{(\phi,\pi,w)} \left\{\int (k - (G^\phi)^{-1} \circ H^\phi(k))^2 dH^\phi\right\} \tag{4.32}$$

where (ϕ, π, w) satisfy the constraints of Problem 6.

4.3.4 Counterexample to linearity of the separating function

Let $R(k, s) = \chi_{[0,1]^2}$, so that skills are distributed uniformly on the unit square. Let $a(k) = ak, b(s) = bs, c = 0$ so that $f(k, s) = ak + bs$. As we showed in the last section, $b(\phi(k)) = a(k) + 2\alpha$ so by rearranging we obtain $\phi(k) = \dfrac{ak + 2\alpha}{b}$. Furthermore, the separating function should divide the density in half so we should have $\displaystyle\int_0^1 \Phi(\phi(\hat{k}))d\hat{k} = 1/2$, where $\Phi(x) = max(min(x, 1), 0)$ clamps the value between 0 and 1. If $\alpha > 0$ and $b - 2\alpha > a$, it is easy to check that $0 \leq \phi(k) \leq 1$ for all k in the unit interval so we can omit Φ. In this case a simple computation shows that $\alpha = \dfrac{b - a}{4}$.

Let $a = 1, b = 3$. We see that $\alpha = 1/2$ and the conditions for this computation to be valid are satisfied. So $\phi(k) = \dfrac{k + 1}{3}$. Now by the definition of ϕ we have $w(\phi(k)) = \pi(k)$.

We claim that this no longer holds for some k when $c > 0$. Therefore the separating function is no longer the same. Suppose ϕ is the same for some $c > 0$. Then we note that μ is also the same. The corresponding equality would be

$$3\phi(k) + c\int_0^{\phi(k)} \mu^{-1}(s)ds - \alpha = k + c\int_0^k \mu(\hat{k})d\hat{k} + \alpha \tag{4.33}$$

Substitute the old expression for ϕ and we obtain:

$$k+1+c\int_0^{\frac{k+1}{3}}\mu^{-1}(s)ds = k+c\int_0^k \mu(\hat{k})d\hat{k}+2\alpha \tag{4.34}$$

$$c\left(\int_0^{\frac{k+1}{3}}\mu^{-1}(s)ds - \int_0^k \mu(\hat{k})d\hat{k}\right) = 2\alpha - 1 \tag{4.35}$$

which has to hold for all k in the unit interval. So differentiate with respect to k, we obtain that $\mu^{-1}(\frac{k+1}{3}) * 1/3 = \mu(k)$. But $\mu(0) = 1/3$ which implies that $\mu^{-1}(1/3) = 1$ contradicting the fact that $\mu^{-1}(1/3) = 0$ as we saw.

4.4 Dependence of the model on relevant quantities

In this section we study how the model is affected when different inputs change, of particular interest is the economical question: If the production function varies slightly, with the same labor force, is it true that the separation of occupations will vary slightly too? .
This question can be reformulated in terms of the model, if the difference between two production functions is small (in an appropriate normed space) is it true that the difference of the resulting optimal separation functions is small (in appropriate normed space)?. We will answer this question positively in the next section.

4.4.1 On continuity of separation

In this section we provide a positive answer to the question posed in the introduction.

Theorem 266. *(Continuity of separation)*
Suppose that $(\phi, \pi, w, \mu) \in \mathcal{C}$ realize the maximum in Problem 5 for a super-modular function F and $\mathcal{R} \in \mathcal{P}_{ac,c}(\mathbb{R}^2)$, for every $\epsilon > 0$ there exists $\delta > 0$ such that if $(\tilde{\phi}, \tilde{\pi}, \tilde{w}, \tilde{\mu}) \in \mathcal{C}$ is an optimal quadruple for a super-modular function $\tilde{F}$ and the same labor force $\mathcal{R}$ with

$$||F - \tilde{F}||_{\mathcal{C}^1} < \delta$$

then

$$||\phi - \tilde{\phi}||_\infty < \epsilon$$

Lemma 267. *($F \to \pi$ continuity)*
Let $\{F_n\}$ be a sequence of twice differentiable functions satisfying supermodularity and converging uniformly to a supermodular, twice continuously differentiable function F. If π_n and π are their Kantorovich potentials then $\pi_n \to \pi$ in uniform norm.

Proof. By Lemma 252 every potential is Lipschitz with constant depending on it's production function F_n. If $F_n \xrightarrow{C^1} F$, then for n big enough $\{\pi_n, \pi\}$ are equi-Lipschitz with Lipschitz constant at most $|\sup_{(x,y)} D_x F(x,y)|$ and so by Arzela-Ascolil using Lemma 253 have a convergent subsequence with respect to uniform topology. By relabelling, assume $\pi_n \to \pi$ in uniform topology, by Lemma 257 we obtain the desired result.

∎

Proof of Theorem 266: The theorem results by consequently applying Lemmata 267 and 257. ∎

Example 268. *(From small interaction to none)*
For $c > 0$ consider the function $F_c(k,s) = a(k) + b(s) + cks$ and $F(k,s) = a(k) + b(s)$, then

$$\nabla F_c(k,s) = \begin{bmatrix} a'(k) + cs \\ b'(s) + ck \end{bmatrix} \tag{4.36}$$

Note that $F_c \to F$ in uniform norm but also $\nabla F_c \to \nabla F$ in uniform norm over $\mathrm{spt}(\mathcal{R})$, so $F_c \xrightarrow{C^1(\mathrm{spt}(\mathcal{R}))} F$ and Theorem 266 applies. Therefore in the limiting case of interaction, separation remains close. This can be understood as follows: If the output of the work done by two people depends very slightly on how they interact, the distribution of workers for both occupational roles will be similar to the ones observed in no-interaction at all. This is a mathematical justification of a expected economical behaviour.

4.4.2 Maximum wage inequality and matching someone with very different skill

Observe that if (ϕ, π, w, μ) is the optimal quadruple for a supermodular function F and a labor force $\mathcal{R}$, by Lemma 252 w is a Lipschitz function, denote it's Lipschitz constant by $L_W := \sup|D_2F(k,s)|$ as in Lemma 252, hence

$$|w(\phi(k)) - w(\mu(k))| \leq L_W |\phi(k) - \mu(k)|. \tag{4.37}$$

This means that the in-firm wage inequality can't surpass a factor (depending only on the change in production as one of the skills is changed (see Lemma 252)) times the difference in skill for the secondary job of the person matched with our worker of k skill level for the primary job and the secondary skill of the person who our worker would be indifferent in swapping jobs with.
This quanititative result not only tells us that there is no wage inequality when $\phi = \mu$ but also that the wage is proportional (at worst) to the difference in secondary skill of these workers associated to the person of key skill k.

4.4.3 Numerics

We compute approximations to the true key wage, π by iteratively applying a function $\lambda(\pi)$ which represents the evolution of wages under market forces. Under the conjecture that this is a contraction mapping, this procedure indeed converges to the solution. We begin with an arbitrary π. As long as the distance between the last two wages π_k, π_{k-1} is above some specified threshold ϵ, we compute $\lambda(\pi_k) =: \pi_{k+1}$ by computing the C that equalizes the mass working in each occupation when C is added to π (and thus subtracted from π^F) using a root finding algorithm such as bisection. With the induced skill distributions, we compute the Optimal Transport solution using a specialized library and extract the dual variable associated with the key role. This is our desired $\lambda(\pi_k)$. Note that π may not be defined in skill levels where there are no key workers. In this case, π may be defined arbitrarily except that it has to be strictly increasing on its domain.

To obtain (π, w, μ, ϕ) as in the optimal for Problem 6.

$\pi_0 \leftarrow 0$
Define H_0 and G_0 from π_0 as in Definition 244.
while $\|\pi_k - \pi_{k-1}\|_\infty > \epsilon$ **do**
find C such that $\pi_k + C$ induces $G^\phi(\mathbb{R}) = H^\phi(\mathbb{R})$.
set $w_k = \pi_k^F, \phi_k = w^{-1} \circ \phi, H_k = H^{\phi_k}$
compute optimal transport map $\mu \# H^\phi = G^\phi$.
set π_{k+1} as dual minimizer from OT solution. (can be altered on $\mathcal{R}$ null set to promote convergence)
end while
Return $(\pi_k, w_k, \phi_k, \mu_k)$

The following graphs show various quantities of the numerical solution when the distribution is uniform on the unit square and the revenue function is of the form $F(k, s) = 3k^2 + s^2 + cks$ where c varies.

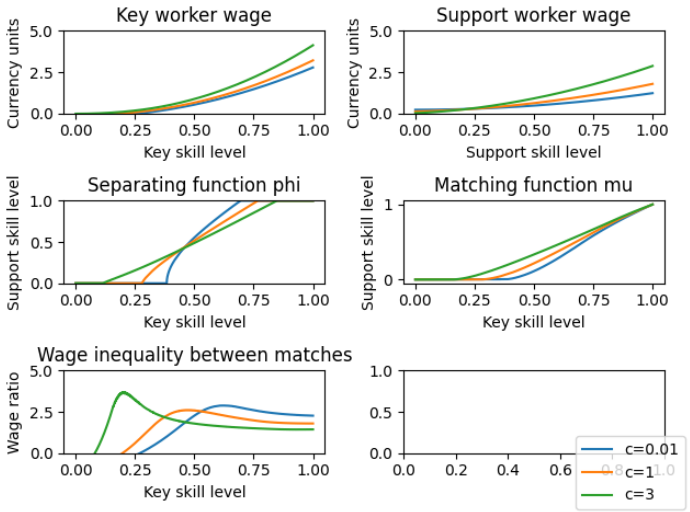

Figure 4.1: Simulation for smaller values of c in the interaction of F

4.5 The Social Planner's problem of McCann-Trokhimtchouk

In this section we present the formulation of the social planner's problem from [McCann-Trokhimtchouk] and it's connections with the present work. We start by presenting the problem, relevant definitions, the duality result and provide economic interpretation. In section 4.5.2 we relate the formulation of this social planner's problem to our generalized Roy model.
The work of McCann and Trokhimtchouk [McCann-Trokhimtchouk] deals in much more generality, where the skill-set is not assumed to be an ordered paired of real numbers. In this section we write the relevant definitions of [McCann-Trokhimtchouk] in the specific case of the skill set $\mathbf{X} = \mathbb{R}^2$ for

consistency. The goal of this section is to relate these definitions to our generalized Roy model, so for consistency we rewrite the relevant definitions in the case of $\mathbb{R}^2$.

4.5.1 Relevant definitions

Definition 269. *(Pure pairing)*
A probability measure $\nu \in \mathcal{P}(\mathbb{R}^2)$ and a Borel function $f : \mathbb{R}^2 \to \mathbb{R}^2$ are called a pure pairing for $\mathcal{R} \in \mathcal{P}(\mathbb{R}^2)$ whenever

$$\nu + f\#\nu = 2\mathcal{R}$$

Problem 8. *(Social Planner's problem)*
Given a production function $p : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, the social planner's problem is the maximization of pure-pairing productions, that is,

$$\sup_{\nu + f\#\nu = 2\mathcal{R}} \left\{ \int p(x, f(x)) d\nu \right\}. \tag{4.38}$$

The fact that the domain of this production function is a subset of $\mathbb{R}^4$ means that the joint production achieved by a couple of workers may depend on both skills of the workers, i.e. even if a worker performs the manager role, his assistant skill set influences production. This model is different to Problem 5 from section 4.2.4. Nevertheless, one can specify the model in $\mathbb{R}^2 \times \mathbb{R}^2$ to our context by setting $p((k, s), (k', s')) = F(k, s')$.

Finally we review the existence result which mimics the existence theorem from optimal transportation, the idea is that an optimal pure pairing must be supported in a p-cyclically monotone set which in turn correspond to subdifferentials of potentials.

Theorem 270. *(Existence and uniqueness of optimal pure pairings)*
Assume that p is non-negative, continuously differentiable and satisfies,

$$y \to \nabla_1 p(x, y) \text{ is inyective } \forall y. \tag{4.39}$$

If $\mathcal{R} \in \mathcal{P}_{ac,c}(\mathbb{R}^2)$ then there exists a p-contact map f such that all optimal mixed pairings are of the form $\gamma = (Id, f)\#(P_{(1,2)}\#\gamma)$. And this function is unique $P_{(1,2)}\#\gamma$ a.e., where $P_{(1,2)}$ is the projection onto the first two coordinates: $P_1((x, y, x', y')) = (x, y)$.

4.5.2 Relation to the Generalized Roy Model

The objective of this section is to relate Problem 5 with the problem 8. One may think that the Generalized Roy model can be put in the general framework of McCann-Trokhimtchouk but the assumptions on Theorem 270 are not satisfied, hence one can not ensure from this framework the general existence and uniqueness result. Note that Theorem 270 presents sufficient conditions for existence of pure pairings, while the existence in the Generalized Roy Model was discussed in section 251.

Proposition 271. *(Dissimilarity of the models)*
In the framework of Problem 8, the production function associated to Problem 5 does not satisfy the twist condition (4.39).

Proof. Let $p : \mathbb{R}^2 \times \mathbb{R}^2$ be given by $p((x,y),(x',y')) = F(x,y')$ where F is the production function on Problem 5, then

$$\nabla_1 p((x,y),(x',y')) = \begin{bmatrix} \partial_1 F(x,y') \\ 0 \end{bmatrix} \tag{4.40}$$

which is clearly not injective as a function of (x',y'). ∎

This means that even though the problems are related (as both are self-partition of labor force) the conditions of the independence of the production function F on the skills not used don't allow us to conclude the existence and uniqueness from the general framework of [McCann-Trokhimtchouk]. The main reason is that the labor force partition is being done in different ways. The generalized Roy Model (Problem 5) has the peculiar property that people don't care about their partner's abilities to perform the role they won't end up performing. Note that the imposition of the outer supremum on Problem 7 is motivated by the imposition of occupational choice, which is not imposed a priori in the social Planner's problem 8. One can think of this difference as the fact that a social planner will determine the distribution of occupations without considering the specific preferences of each individual.

4.6 On the identification problems

Suppose now that we do not know the initial distribution of skills $\mathcal{R}$ but we see the optimal matchings made by firms and we see the conditional (on salaries) distributions of skills. The identification problem asks on what conditions can we recover the distribution of skills $\mathcal{R}$?
Mathematically, the identification problem corresponds to the uniqueness in the inverse problem. The more interesting question is wether or not we can recover together the distribution of skills and the production function, that is, given the matchings, earnings and distribution of skills can we recover the unconditional distribution of skills $\mathcal{R}$ together with the production function F ?

4.6.1 Identification on social planner's problem

In the context of problem (4.5), given ν and f, the condition

$$\nu + f\#\nu = 2\mathcal{R} \tag{4.41}$$

completely determines $\mathcal{R}$. This is observed by definition, as for every A Borel subset of $\mathbb{R}^2$ we have

$$\mathcal{R}(A) = \frac{\nu(A) + \nu(f^{-1}(A))}{2} \tag{4.42}$$

Whether or not the production function p is known, we can always use (4.42) to find $\mathcal{R}$.
In the context of 4.5 if p is twice differentiable, non-negative, satisfies the twist condition and $\mathcal{R}$ is containted in a compact set and the diagonal is $p-$cyclically monotone, then by [McCann-Trokhimtchouk, Corollary 1] the mixed pairing is unique indicating how to obtain the production function via

$$p(k,s) = \pi(k) + w(s) \qquad 2\mathcal{R} - \text{a.e.},$$

which is exhaustive as we know π, w and $\mathcal{R}$. This implies we can only recover the production function in the support of the unconditional distribution, meaning that we can only know the

production function for the workers that we see and it's behaviour outside the support of $2\mathcal{R}$ can not be determined. Of course, under regularity assumptions on $\mathcal{R}$ and p like being Lipschitz-continuous one can extend p uniquely.

4.6.2 Discussion on the identification on general non-linear Roy model

In a similar way to the previous section, we ask ourselves whether we can recover $\mathcal{R}$ and F from knowledge of (π, w, ϕ, μ).
If π is fixed and assumed to be continuous and F is known and strictly supermodular, Assume there exists two different unconditional distributions $\mathcal{R}_1, \mathcal{R}_2$ with continuous densities R_1, R_2 then for every $k \in \text{spt}(H^{(\pi,w)})$

$$0 = H^{(\pi,w)}(k) - H^{(\pi,w)}(k) = \int_{k_1}^{k} \int \mathbf{1}_{\{\pi(\tilde{k}) \geq w(s)\}} (R_1(\tilde{k}, s) - R_2(\tilde{k}, s)) ds d\tilde{k}. \tag{4.43}$$

Let $s^* \in \text{spt}(P_2 \# \mathcal{R})$, if $k^* \in \text{spt}(P_1 \# \mathcal{R})$ is such that $\pi(k^*) > w(s^*)$ then by continuity of π and w there exist δ_1, δ_2 such that if $(k, s) \in (k^* - \delta_1, k^* + \delta_1) =: B_{\delta_1,\delta_2}(k^*, s^*) \times (s^* - \delta_1, s^* + \delta_2)$ then

$$\pi(k) > w(s). \tag{4.44}$$

Using (4.43) we obtain that

$$\begin{aligned} 0 = &\int_{k^*-\delta_1}^{k^*+\delta_1} \int \mathbf{1}_{B_{\delta_1,\delta_2}(k^*,s^*)}(\tilde{s}) \cdot (R_1(\tilde{k}, \tilde{s}) - R_2(\tilde{k}, \tilde{s})) d\tilde{s} d\tilde{k} \\ &+ \int_{k^*-\delta_1}^{k^*+\delta_1} \int \mathbf{1}_{(B_{\delta_1,\delta_2}(k^*,s^*))^c}(\tilde{s}) \cdot (R_1(\tilde{k}, \tilde{s}) - R_2(\tilde{k}, \tilde{s})) d\tilde{s} d\tilde{k} \end{aligned}$$

Using that $\tilde{k} \in (k^* - \delta_1, k^* + \delta_1)$ yields

$$0 = \int_{k^*-\delta_1}^{k^*+\delta_1} \int \mathbf{1}_{B_{\delta_1,\delta_2}(k^*,s^*)}(\tilde{s}) \cdot (R_1(\tilde{k}, \tilde{s}) - R_2(\tilde{k}, \tilde{s})) d\tilde{s} d\tilde{k} \tag{4.45}$$

Because $R_1 - R_2$ is integrable, by continuity of the Lebesgue integral:

$$\lim_{\delta_1,\delta_2 \to 0} \int_{k^*-\delta_1}^{k^*+\delta_1} \int \mathbf{1}_{B_{\delta_1,\delta_2}(k^*,s^*)}(\tilde{s}) \cdot (R_1(\tilde{k}, \tilde{s}) - R_2(\tilde{k}, \tilde{s})) d\tilde{s} d\tilde{k} = 0. \tag{4.46}$$

Equation (4.46) impedes us from concluding $R_1(k^*, s^*) = R_2(k^*, s^*) = 0$ with the usual techniques. This impediment reinforces the idea of [Heckman-Honore] in the linear case where the distribution can not be identified. Nevertheless, notice that under regularity assumptions on $F, \phi, \mathcal{R}$, differentiating twice (244) we obtain

$$R(k, \phi(k)) = \frac{(H^{\phi})''(k)}{\phi'(k)}. \tag{4.47}$$

This formula serves as partial analogue of (4.42). We can identify the distribution at the points $(k, \phi(k))$ explicitly but it seems like nowhere else. .

4.6.3 Identification of production

The identification of production is a more subtle question. Note that uniqueness of Brenier maps in 1-dimensional transport indicates that if the distributions are one dimensional, the optimal matching will be the same for all super-modular functions. The question whether or not one can recover the pair $(F, \mathcal{R})$ from only the information of (π, w, μ) remains open to the authors. Observe that F was known in section 4.6.2 when we looked for $\mathcal{R}$, the identification of the pair $(\mathcal{R}, F)$ is expected to not be solvable i.e. we expect many different pairs to yield the same earnings (π, w) and matching μ, although we still have no rigorous proof.

4.7 Further development and some open questions

In this section we describe some generalizations, problems and ideas that we believe would make interesting lines of future investigation.

4.7.1 Infinite dimensional linear program

Motivated by the success of the Kantorovich formulation in the Monge-Kantorovich problem, in this section we formulate the infinite-dimensional relaxation of Problem 5.

Definition 272. *(Kantorovich formulation of Roy's model)*
Given a supermodular production function $F : \mathbb{R}^2 \to \mathbb{R}$ and $\mathcal{R} \in \mathcal{P}_c(\mathbb{R}^2)$ we define the relaxation of the generalized Roy model (Problem 5) as the linear program defined via

$$\sup_{\bigcup_{\pi \in \mathcal{C}_3} \Gamma(H^\pi, G^\pi)} \left\{ \int F(k,s) d\gamma(k,s) \right\} \tag{4.48}$$

where $\Gamma(\mu, \nu) = \{\gamma \in \mathcal{P}(\text{spt}(\mathcal{R})) : P_1 \# \gamma = \mu, P_2 \# \gamma = \nu\}$ and $\mathcal{C}_3$ is the set defined in Problem 5.

The constraint set in Definition 272 may be an interesting object of study. It is not clear to the author whether or not this set is compact or even convex. Observe that a formulation like that of (4.48) resembles the work in [McCann-Trokhimtchouk], indicating possible lines of investigation.

Reformulation of the definitions of the measures

Given a function $f : X \to \mathbb{R}$ is $S \subseteq X$, the S-hypograph of f is defined via

$$Hyp_S(f) = \{(x, r) \in S \times \mathbb{R} : r \leq f(x)\}.$$

Similarly, the S-strict epi-graph of f is defined via

$$SEpi_S(f) = \{(x, r) \in S \times \mathbb{R} : f(x) < r\}.$$

With this notation, Definition 244 rewrites

$$H^\pi(A) = \mathcal{R}(Hyp_A(\pi))$$
$$G^\pi(B) = \mathcal{R}(SEpi_B(\pi)).$$

We expect this notation to be useful to simplify some of the proofs and enlighten other properties as the hypergraph and the epigraph have notable properties for convexity/concavity.

4.7.2 Superoptimality Conjecture

The first line of investigation seems to be whether or not Conjecture 4.26 holds true. This conjecture is interesting both in mathematical and economical sides. We refer to section 4.2.10 for the details. This conjecture is also related to classical economical theory, one could attempt to use Theorem 9.19 in [Roth-Marilda-Sotomayor] but the generalizations and connections should be established rigorously,

4.7.3 Generalizations and extensions

The generalized Roy Model (Problem 5) presented here applies for an absolutely continuous, compactly supported labor force with skill sets in $\mathbb{R}^2$ and a supermodular function F that does not depend on the skills of your partner in the job not performed. The general version of [McCann-Trokhimtchouk] deals with much more generality but one could attempt to introduce occupational choice. The separation function of Problem 5 was shown to be removable via the equivalence with Problem 7, nevertheless it provides interesting economic interpretations, so one must ask, is there an equivalent of separation in many dimensions? If so, how can one interpret such a function?.
The intuitive modelling using the separation function ϕ, could potentially be used for a multi-role model on which one would obtain multiple separation functions and the matching would realize n-tuples of people to work on a firm. This could be an interesting model for the hiring of teams to perform a job but the optimal matching in this context would require different tools.
Another interesting point of investigation is the relaxation of the continuity of separation functions made in Definition 244, we expect the model to be unstable and very different if such hypothesis is removed.
Along a similar line, the condition of Definition 243 could potentially be removed by looking at separation functions with different domains, this apparent technicallity showed to be necessary for the strategy used during the proof of Lemma 258 and the use of Arzela-Ascoli requires a uniform domain. Studying existence and stability in this framework is still open and interesting.

4.7.4 On the second fundamental Theorem of Welfare

A very interesting line of investigation comes from pure economical reasoning. If the social planner's problem from [McCann-Trokhimtchouk] (Problem 8) is indeed a Pareto equilibrium, can one find the initial conditions (say F supermodular and R) such that the solution for Problem 5 achieves the solution of Problem 8 ? If not, as they may be disassociated, what is the competitive version (with no occupational choice as constraints) that allows the social planner's problem of McCann-Trokhimtchouk to be attained for initial conditions? Can one find such initial conditions?

Furthermore, the solution to Conjecture (4.26) could come from the second fundamental theorem of welfare but appropriate definitions and connections are yet to be established.

4.7.5 First variations, the envelope theorem and approximating total production in similar economies

In this section we propose two intuitive ways to approximate the resulting outputs of different economies. Let us introduce notation:

$$TP_{GRM}(F, \mathcal{R}) = \sup_{(\phi,\pi,w,\mu)\in\mathcal{C}} \left\{ \int F(k, \mu(k)) dH^{\phi} \right\},$$
$$TP_{MT}(p, \mathcal{R}) = \sup_{\nu + f\#\nu = 2\mathcal{R}} \left\{ \int p(x, f(x)) d\nu \right\}.$$

One could attempt to approximate to first order the values for similar economies via

$$TP_{MT}(p, \mathcal{R}) - TP_{MT}(p, \tilde{\mathcal{R}}) \approx 2 \int v(x) d\mathcal{R} \cdot d_2(\mathcal{R}, \tilde{\mathcal{R}})$$
$$TP_{GRM}(F, \mathcal{R}) - TP_{GRM}(\tilde{F}, \mathcal{R}) \approx \left(\int \pi - \tilde{\pi} dH^{\phi} + \int (w - \tilde{w}) dG^{\phi} \right) ||F - F||_{\infty}$$

This ideas are motivated from the concept of first variations (see [Santambrogio, Chapter 7]) and could be used to approximate the values of economies where one can either only see the earnings or one element of the quadruple but expects labor forces to be similar (in the second case). Whether this approximations are good or not to first order is not known to the authors but would yield an interesting approach to studying similar populations whose production functions are the same or viceversa. The second guess is a little more naive, we attempt to use only salaries observed in both economies, difference between production functions and *only one* of the labor forces to make the prediction.

Bibliography

[Ahidar-Coutrix-Le Gouic] A. Ahidar-Coutrix, T. Le Gouic, Q. Paris, Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics, *Probability Theory and Related Fields* volume 177, pages323–368, (2020).

[Almgren-Taylor-Wang] F. Almgren, J. E. Taylor, and L. Wang, Curvature-driven flows: a variational approach. *SIAM Journal on Control and Optimization* 31, 2 (1993), 387–438.

[Ambrosio-Gigli-Savare] L. Ambrosio, N. Gigli, G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics ETH Zürich, 2005 Birkhäuser Verlag.

[Ambrosio-Gigli-Savare2] L. Ambrosio, N. Gigli, A user's guide to optimal transport, https://doi.org/10.1007/978-3-642-32160-3, Springer, Berlin, Heidelberg.

[Amos-Cohen-Luise-Redko] B. Amos, S. Cohen, G. Luise, I. Redko, Meta Optimal Transport, *ICML*, 2023.

[Arous-Guionnet] G. Arous, A. Guionnet, A. Large deviations for Wigner's law and Voiculescu's non-commutative entropy, *Probab Theory Relat Fields*, 108, 517–542, 1997. https://doi.org/10.1007/s004400050119

[Bakry-Gentil-Ledoux] D. Bakry and I. Gentil and M. Ledoux, *Book*, Analysis and Geometry of Markov Diffusion Operators, Grundlehren der mathematischen Wissenschaften, Springler, Switzerland 2014.

[Ball] K. Ball, An Elementary Introduction to Monotone Transportation, *Geometric Aspects of Functional Analysis*,Part of the Lecture Notes in Mathematics book series (LNM,volume 1850), pp 41–52.

[Baudoin] Fabrice Baudoin, Bakry meets Villani, *Journal of Functional Analysis*, 273 (2017) 2275–2291.

[Baudoin-Gordina-Mariano] F. Baudoin , M. Gordina, P. Mariano , Gradient bounds for Kolmogorov type diffusions, *Annales de l'Institut Henri Poincaré* - Probabilités et Statistiques, Vol. 56, No. 1, 2020, , 612–636, https://doi.org/10.1214/19-AIHP975

[Bishop] C. Bishop, *Book*, Pattern Recognition and Machine Learning (Information Science and Statistics),Springer-Verlag, Berlin, Heidelberg ISBN:978-0-387-31073-2, (2006).

[Benoit et al.] Benoît Fabrèges, Hélène Hivert, Kévin Le Balc'h, Sofiane Martel, François Delarue, *Numerical schemes for the aggregation equation with pointy potentials.* ESAIM: Proceedings and Surveys, 2019, ff10.1051/proc/201965384ff. ffhal-01788050

[Bertozzi-Laurent-Rosado] A. L. Bertozzi, T. Laurent, and J. Rosado. Lp theory for the multidimensional aggregation equation. *Communications on Pure and Applied Mathematics*, 64(1):45–83, 201

[Billingsley] P. Billingsley, *Book*,Convergence of Probability Measures, Germany, Wiley, 2013.

[Bismut] J-M. Bismut, *Book*, Hypoelliptic Laplacian and Orbital Integrals, AM-177, Ukraine, Princeton University Press,2011.

[Benamou-Brenier] J.-D. Benamou, Y. Brenier, K. Guittet, The Monge-Kantorovitch mass transfer and its Computational Fluid Mechanics formulation, *International Journal For Numerical Methods In Fluids Int. J. Numer. Meth. Fluids* 2000;

[Bogachev] V.I. Bogachev, *Book*, Measure Theory, Springer Berlin, Heidelberg, 2006.

[Bonaschi-Carrillo-DiFrancesco-Peletier] G.A. Bonaschi, J.A. Carrillo, M. Di Francesco and M.A. Peletier, Equivalence of gradient flows and entropy solutions for singular nonlocal interaction equations in 1D. ESAIM: *Control, Optimisation and Calculus of Variations*, 21(2),(2015), 414-441.

[Bourbaki] N. Bourbaki, *Elements of Mathematics, General Topology*, Part 2, Herman , Paris.

[Braides] A. Braides, Gamma-convergence for beginners, Book, *Oxford lecture series in mathematics and its applications*, 22, Oxford University Press, 2002.

[Braun] Mathias Braun. Renyi's entropy on Lorentzian spaces. Timelike curvature dimension conditions. To appear in J. Math. Pures Appl., arXiv:2206.13005.

[Brown] B.M. Brown, A General Three-series Theorem, *Proceedings of the American Mathematical Society*,Volume 28, Number 2, 1971.

[Cavalletti-Mondino] F. Cavalletti,A. Mondino, A review of Lorentzian synthetic theory of timelike Ricci curvature bounds *General Relativity and Gravitation* ,2022,doi: 10.1007/s10714-022-03004-4 , http://cvgmt.sns.it/paper/5527/

[Carrillo-Figalli] J.A. Carrillo, A. Figalli, F. S. Patacchini, Geometry of minimizers for the interaction energy with mildly repulsive potentials, *Annales De l'Institut Henri Poincare (C) Non Linear Analysis*, 34(5), 1299–1308.

[Carrillo-Figalli2011] J. A. Carrillo, M. Di Francesco, A. Figalli, T. Laurent, and D. Slepcev. Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Math. J.*, 156(2):229–271, 2011.

[Carrillo-James-Lagoutiere-Vauchelet] J.A. Carrillo, F. James, F. Lagoutière, and N, Vauchelet, The Filippov characteristic flow for the aggregation equation with mildly singular potentials. *Journal of Differential Equations*, (2016), 260(1), 304-338.

[Chaganti] N. Chaganti, Large Deviations for Joint Distributions and Statistical Applications, *The Indian Journal of Statistics*, Series A (1961-2002) , Jun., 1997, Vol. 59, No. 2 (Jun., 1997), pp. 147-166

[Chevalier-Debbasch] F. Debbasch, K. Mallick, J.P. Rivet, .: Relativistic Ornstein–Uhlenbeck *Process. J. Stat. Phys.*, 88, 945–966, 1997.

[Chiarini-Conforti-Greco] A. Chiarini, G. Conforti, G. Greco, Z. Ren., Entropic turnpike estimates for the kinetic Schrödinger problem., *Electronic Journal of Probability*, 27(none) 1-32 2022. https://doi.org/10.1214/22-EJP850

[Chavel] I. Chavel *Riemannian Geometry A Modern Introduction*, Cambridge Studies in Advanced Mathematics, Second Edition, Cambridge University Press, 2006

[Cordero-Erausquin-McCann-Schmuckenschläger] D. Cordero-Erausquin , R. McCann and M. Schmuckenschläger, A Riemannian interpolation inequality à la Borell, Brascamp and Lieb, *Invent. math.*, 146, 219–257 (2001), (DOI) 10.1007/s002220100160.

[Courty-Flamary] N. Courty, R. Flamary, D. Tuia, A. Rakotomamonjy, Optimal Transport for Domain Adaptation, *IEEE Trans Pattern Anal Mach Intell.*, 2017 Sep;39(9):1853-1865. doi: 10.1109/TPAMI.2016.2615921. Epub 2016 Oct 7

[Arjovsky et. al] Martin Arjovsky and Soumith Chintala and Léon Bottou, Wasserstein Generative Adversarial Networks,*Proceedings of the 34th International Conference on Machine Learning*, 214-224,2017.

[Courty et al.] A. Rakotomamonjy, R. Flamary, G. Gasso, M. El Alaya, M. Berar, N. Courty , Optimal Transport for Conditional Domain Matching and Label Shift, *Machine Learning*, Vol. 111, pages1651–1670, 2022

[Criscitiello-Boumal] C. Criscitiello, N. Boumal, N. An Accelerated First-Order Method for Non-convex Optimization on Manifolds, *Found Comput Math* 23, 1433–1509 (2023). https://doi.org/10.1007/s10208-022-09573-9

[Cuturi-Doucet] M. Cuturi, A. Doucet, Fast Computation of Wasserstein Barycenters, *Proceedings of the 31st International Conference on Machine Learning*, PMLR 32(2):685-693, 2014.

[DeLellis] De Lellis, Camillo. Ordinary differential equations with rough coefficients and the renormalization theorem of Ambrosio (after Ambrosio, DiPerna, Lions), *Séminaire Bourbaki*, Volume 2006/2007 - Exposés 967-981, Astérisque, no. 317, 2008, Talk no. 972, 29 p.

[Dembo-Zeitouni] A. Dembo, O. Zeitouni, *Book*, Large Deviations Techniques and Applications, Springer Berlin, Heidelberg, 2nd Ed, 2009, XVI, 396.

[Devroye] L. Devroye, L, On arbitrarily slow rates of global convergence in density estimation. *Z. Wahrscheinlichkeitstheorie verw Gebiete* 62, 475–483, 1983. https://doi.org/10.1007/BF00534199

[Devroye-Giorfi] L. Devroye, L. Gyorfi No empirical probability measure can converge in the total variation sense for all distributions, *The Annals of Statistics*, Vol. 18, No. 3, 1990, pp. 1496-1499

[De Giorgi] E. De Giorgi,, New problems on minimizing movements. *Selected Papers* (1993), 699–713.

[Do Carmo] M. Do Carmo, *Riemannian Geometry*, Birkhauser, Second edition, 1992,USA.

[Dudley1966] R.M., Dudley, Lorentz-invariant Markov processes in relativistic phase space, *Ark. Mat.* 6, 241–268, 1966. https://doi.org/10.1007/BF02592032

[Dudley1967] R.M. Dudley, A note on Lorentz-invariant Markov processes. *Arkiv Mat.*, 6, 575–581, 1967.

[Dudley1973] R.M. Dudley, Asymptotics of some relativistic Markov processes. *Proc. Natl. Acad. Sci.*, USA 70, 3551–3555, 1973

[Dudley2002] R. M. Dudley, Real Analysis and Probability, Book, Cambridge University Press, 2nd Edition, 2002, Cambridge.

[Dunkel-Hanggi] J. Dunkel, P. Hänggi, Theory of relativistic Brownian motion: The (1+1)-dimensional case, *Phys. Rev.* E 71, 016124, 2005

[Durrett] R. Durrett, Probability Theory, Theory and Examples, *Book*, 2nd Ed, Duxbury Press, 1995.

[Eckstein-Miller] M. Eckstein, T. Miller, Causality for Nonlocal Phenomena, *Ann. Henri Poincaré* 18, 3049–3096 (2017). https://doi.org/10.1007/s00023-017-0566-1

[Franchi] J. Franchi, From Riemannian to Relativistic Diffusions, *From Riemann to Differential Geometry and Relativity*, Springer, Switzerland 2017.

[Franchi-LeJan2012] J. Franchi, Y. Le Jan, Hyperbolic Dynamics and Brownian Motions: an introduction, *Book*, Oxford University Press, USA, 2012.

[Franchi-LeJan2007] J. Franchi, Y. Le Jan, Relativistic Diffusions and Schwarzschild Geometry, *Communications on Pure and Applied Mathematics*, 2007, LX, pp.187-251. ⟨10.1002/cpa.20140⟩. ⟨hal-00089071⟩

[Figalli-Villani-Rifford] A Figalli, L Rifford, C Villani, On the Ma-Trudinger-Wang condition on surfaces, *Calculus of Variations and Partial Differential Equations*, Springer Verlag, 2010, 39 (3-4), pp.307.

[Fitzsimmon-Pitman-Yor] P. Fitzsimmons , J. Pitman , M. Yor, Markovian bridges: construction, Palm interpretation, and splicing, textit Progress in Probability , Volume 33, 1992.

[Feinberg-Kasyanov-Liang] Feinberg P. O. Kasyanov, Y. Liang, Fatou's Lemma for Weakly Converging Probabilities,*Theory of Probability and Its Applications* ,Vol. 58, Iss. 4, 2014.

[Fetecau-Park-Patacchini] R. C. Fetecau, H. Park, F. S. Patacchini, Well-posedness and asymptotic behaviour of an aggregation model with intrinsic interactions on sphere and other manifolds. arXiv:2004.06951

[Fetecau-Patacchini] R. C. Fetecau and F. S. Patacchini, Well-posedness of an interaction model on Riemannian manifolds, preprint, http://arxiv.org/abs/2109.03959, *Comm. Pure Appl. Anal.*, (accepted 2021).

[Gigli-Tamanini] N. Gigli, L. Tamanini, Benamou-Brenier and duality formulas for the entropic cost on $RCD^*(K,N)$ spaces *Probab. Theory Related Fields*, 2020.

[Hakim] Hakim, R. Relativistic Stochastic Processes, *J. Math. Phys.*, 1968, 9, 1805–1818.

[Hawking-Ellis] S. Hawking,, G. Ellis, G. *Book*, The Large Scale Structure of Space-Time, Cambridge Monographs on Mathematical Physics, 1973 Cambridge: Cambridge University Press. doi:10.1017/CBO9780511524646.

[Heckman-Honore] J. J. Heckman and B. E. Honoré,The Empirical Content of the Roy Model *Econometrica*, Vol. 58, No. 5 (Sep., 1990), pp. 1121-1149 (29 pages).

[Hsu] P. Hsu, *Book*, Stochastic Analysis on Manifolds, Graduate Studies in Mathematics, Volume: 38; 2002; 281 pp.

[Hsu1990] P. Hsu, P. Brownian bridges on Riemannian manifolds. *Probab. Th. Rel. Fields*, 84, 103–118 (1990). https://doi.org/10.1007/BF01288561

[Ji-Egerstedt] M. Ji and M. Egerstedt. *Distributed coordination control of multi-agent systems while preserving connectedness.* IEEE Trans. Robot., 23(4):693–703, 2007

[Jordan-Kinderlehrer-Otto] R. Jordan, D. Kinderlehrer, F. Otto ,The Variational Formulation of the Fokker–Planck Equation, *SIAM Journal on Mathematical Analysis*, Volume 29, Issue 1.

[J] J. Sánchez García, Interplay between curvature and isoperimetry: A relationship through functional inequalities, Master's Thesis, Masterarbeit : Rheinische Friedrich-Wilhelms-Universität Bonn, 2018.

[Kallenberg] O. Kallenberg, *Book*, Foundations of Modern Probability, Probability Theory and Stochastic Modelling, Springer, 3rd edition, 2021.

[Kunziger-Saemann] N. Kunzinger,C. Sämann, Lorentzian length spaces, *Ann Glob Anal Geom*, 54, 399–447 (2018). https://doi.org/10.1007/s10455-018-9633-1

[Leonard2014] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 2014, 34(4): 1533-1574. doi: 10.3934/dcds.2014.34.1533

[Leonard2012] C. Léonard, From the Schrödinger problem to the Monge–Kantorovich problem, *Journal of Functional Analysis* 262, 2012, 1879–1920.

[Leonard2001] C. Léonard, Minimizers of energy functionals, *Acta Mathematica Hungarica*, 93, 281–325 (2001).

[Leonard2001b] C. Léonard, Minimization of energy functionals applied to some inverse problems. *J. Appl. Math. Optim.*, 44:273–297, 2001.

[Leonard] C. Léonard. Stochastic derivatives and generalized h-transforms of Markov processes. Preprint, arXiv:1102.3172.

[Leonard2014b] Some properties of path measures, *Séminaire de probabilités*, 46. Lecture Notes in Mathematics 2123 (2014), pp 207-230.

[Loeper] Loeper, G. Regularity of Optimal Maps on the Sphere: the Quadratic Cost and the Reflector Antenna. *Arch Rational Mech Anal*, 199, 269–289 (2011). https://doi.org/10.1007/s00205-010-0330-x

[Ma-Trudinger-Wang] X-N Ma, N S. Trudinger, X-J Wang Regularity of Potential Functions of the Optimal Transportation Problem, ¸*Archive for Rational Mechanics and Analysis* volume 177, pages 151–183 (2005)

[Mauri-Giona] R. Mauri, M.A. Giona, Covariant Non-Local Model of Bohm's Quantum Potential. *Entropy* 2023, 25, 915. https://doi.org/10.3390/e25060915

[Manupriya-Keerti-Biswas-Chandhok-Jagarlapudi] P. Manupriya, R. Keerti, S. Biswas, S. Chandhok, S. Nath Jagarlapudi, Empirical Optimal Transport between Conditional Distributions, arxiv: arXiv:2305.15901

[Matsumoto-Ikeda] N. Ikeda, H. Matsumoto, The Kolmogorov Operator and Classical Mechanics, *In Memoriam Marc Yor Séminaire de Probabilités XLVII*, Lecture Notes in Mathematics, Springer, Switzerland, 2015.

[Milne] T. Milne, Optimal Transport, Congested Transport, and Wasserstein Generative Adversarial Networks, University of Toronto, 2022.

[McCann-Saemann] R. McCann, C.Saemann, A Lorentzian analog for Hausdorff dimension and measure, *Pure and Applied Analysis*, 2022, 367-400. arXiv:2110.04386

[McCann2023] R. McCann, (Preprint), A synthetic null energy condition, arXiv 2304.14341

[McCann2019] R. McCann, Displacement convexity of Boltzmann's entropy characterizes the strong energy condition from general relativity, arXiv1808.1536v2. *Camb. J. Math.*, 8:3, 2020, 609-681.

[McCann2020] R.J. McCann, Displacement convexity of Boltzmann's entropy characterizes the strong energy condition from general relativity, *Camb. J. Math.* 8:3 (2020) 609-681.

[McCann] R.J. McCann, Polar factorization of maps on Riemannian manifolds., *Geom. Funct. Anal.* 11 (2001) 589-608

[McCann-Davies-Lim] R.J. McCann C. Davies and T. Lim, Classifying minimum energy states for interacting particles: spherical shells, arXiv:2107.11718

[McCann-Figalli] R.J McCann, A. Figalli, Y-H Kim, Regularity of optimal transport maps on multiple products of spheres, *J. Eur. Math. Soc. (JEMS)*, 15 (2013) 1131-1166.

[McCann-Kim] R.J. McCann, Y-H Kim Curvature and the continuity of optimal transport, *Oberwolfach Rep.*, 4 (2007) 2060-2062

[McCann-Guillen] R. McCann and N. Guillen, Five lectures on optimal transportation: geometry, regularity and applications, *Analysis and Geometry of Metric Measure Spaces: Lecture Notes of the Seminaire de Mathematiques Superieure (SMS) Montreal 2011.* G. Dafni et al, eds. Providence: Amer. Math. Soc. (2013) 145-180.

[McCann-Trokhimtchouk] Robert J. McCann , Maxim Trokhimtchouk, Optimal partition of a large labor force into working pairs, *Economic Theory* Vol. 42, No. 2 (Feb., 2010), pp. 375-395 (21 pages)

[Miller] T. Miller, Polish spaces of causal curves, *Journal of Geometry and Physics*, Volume 116, 2017, 295-315, https://doi.org/10.1016/j.geomphys.2017.02.006.

[Mogilner-Edelstein.Keshet] A. Mogilner and L. Edelstein-Keshet. *A non-local model for a swarm.* J. Math. Biol., 38:534– 570, 1999.

[Panaretos-Zemel] V.M. Panaretos, Y. Zemel, An Invitation to Statistics in Wasserstein Space, Book, *SpringerBriefs in Probability and Mathematical Statistics*, Springer, 2020.

[Parszen] E. Parzen, On Estimation of a Probability Density Function and Mode, *The Annals of Mathematical Statistics*, Vol. 33, No. 3, pp. 1065-1076, 1962.

[Patacchini-Slepcev] F.S. Patacchini, D. Slepčev, The nonlocal-interaction equation near attracting manifolds, *Discrete and Continuous Dynamical Systems*, 2022, 42 (2) : 903-929. doi: 10.3934/dcds.2021142

[Peyre-Cuturi] G. Peyre, M. Cuturi, *Computational Optimal Transport*, *Foundations and Trends in Machine Learning*, Vol. 11: No. 5-6, pp 355-607. http://dx.doi.org/10.1561/2200000073

[Pouget-Abadie et al.] I. J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks,*Advances in neural information processing systems*, Volume 27, (2014).

[RassoulAgha-Seppäläinen] F. Rassoul-Agha, T. Seppäläinen *Book*, A Course on Large Deviations with an Introduction to Gibbs Measures, Graduate Studies in Mathematics, Vol. 162, American Mathematical Society, 2015.

[Roth-Marilda-Sotomayor] A. Roth, A. Marilda, O. Sotomayor, *Two-sided matching*, Econometric society monographs, vol 18, Cambridge University Press.

[Redko-Vayer-Flamary-Courty] I. Redko, T. Vayer, R. Flamary, N. Courty, CO-Optimal Transport, *Neural Information Processing Systems (NeurIPS)*, 2020.

[Rosenblatt] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, *Ann. Math. Statist.*, 27 832-837, 1956.

[Roy] A. D. Roy, Some Thoughts on the Distribution of Earnings,*Oxford Economic Papers*, New Series, Vol. 3, No. 2 (Jun., 1951), pp. 135-146 (12 pages).

[Royden] H.L. Royden, P.M. Fitzpatrich, Real Analysis, Book, Fourth Edition, Prentice Hall, 2010.

[Rupert-Woolgar] M Rupert and E Woolgar, Bakry-Emery black holes, *Class Quantum Gravit* 31. (2014) 025008 [arxiv:1310.3894].

[Villani2003] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, Graduate Studies in Mathematics Volume 58, 2003.

[Santambrogio] F. Santambrogio, *Optimal Transport for Applied Mathematicians, Calculus of Variations, PDEs, and Modeling*, *Progress in Nonlinear Differential Equations and Their Applications*, Volume 87, Birkhäuser, Springer International Publishing Switzerland 2015

[Siow-Mak2016] (*Draft*) A. Siow and E. Mak, Occupational Choice and Matching in the Labor Market [to appear].

[Galichon] A. Galichon, *Optimal Transport Methods in Economics*, 2016, Princeton University Press, 184 pp.

[Siow-Mak] Aloysius Siow & Eric Mak *Occupational Choice and Matching in the Labor Market*, 2017 Meeting Papers 30, Society for Economic Dynamics.

[Sriperumbudur] B. Sriperumbudur, K. Fukumizu, A. Gretton, B. Scholkopf, and G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6, 2012.

[Tamanini] Luca Tamanini, Analyse et géométrie des espaces RCD par le biais du problème de Schrödinger, PhD. Thesis, 2017.

[Trigila-Tabak] Trigila, G. and Tabak, E.G., Data-Driven Optimal Transport. *Commun. Pur. Appl. Math.*, (2016), 69: 613-648https://doi.org/10.1002/cpa.21588.

[Varadhan] S. Varadhan, Large deviations, *Book*, Courant Institute of Mathematical Sciences, American Mathematical Society;New York, New York, 2016.

[Wald] Wald, Robert M., *Book*, General Relativity. Chicago, University of Chicago Press, 1984.

[Villani2009] C. Villani, *Optimal Transport: Old and New*, Springer, A series of Comprehensive Studies in Mathematics, 2009, Berlin.