Dynamical Programming for off-the-grid dynamic Inverse Problems

Robert Tovey, Vincent Duval

Dynamical Programming for off-the-grid dynamic Inverse Problems

Vincent Duval Robert Tovey

CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University
75016 Paris, France
INRIA-Paris, MOKAPLAN, 75012 Paris, France

Tuesday 21st December, 2021

Abstract

In this work we consider algorithms for reconstructing time-varying data into a finite sum of discrete trajectories, alternatively, an off-the-grid sparse-spikes decomposition which is continuous in time. Recent work showed that this decomposition was possible by minimising a convex variational model which combined a quadratic data fidelity with dynamical Optimal Transport. We generalise this framework and propose new numerical methods which leverage efficient classical algorithms for computing shortest paths on directed acyclic graphs. Our theoretical analysis confirms that these methods converge to globally optimal reconstructions which represent a finite number of discrete trajectories. Numerically, we show new examples for unbalanced Optimal Transport penalties, and for balanced examples we are 100 times faster in comparison to the previously known method.

1 Introduction

The signal-processing task of dynamical super-resolution involves retrieving fine-scale features, in space and time, from a signal which evolves over time. A convex variational model was recently proposed for such tasks using Optimal Transport (OT) to regularise the associated inverse problem [Bredies et al., 2021, Bredies and Fanzon, 2020]. This new approach allows the decomposition of a signal into a finite sum of smooth curves, for example to track the centers of multiple particles in time with smooth trajectories. Similar ideas were explored for a specific example in Alberti et al. [2019] where the shape of curves is built into the model, and without appealing to OT.

In this work we focus on dynamical super-resolution problems regularised by OT. We can consider potential models to be partitioned into two classes, depending on whether the particles have constant mass/brightness in time, or if mass is allowed to vary. These classes are referred to as balanced or unbalanced problems respectively, mathematically encoded in the choice of OT cost. Current literature provides analysis for the balanced Benamou-Brenier (BB) [Bredies et al., 2021] and the unbalanced Wasserstein-Fisher-Rao (WFR) [Bredies et al., 2020a] energies, both are shown to reconstruct data into a finite number of smooth curves with constant or smoothly varying mass. Initial numerical experiments for the Benamou-Brenier model have also been carried out showing great promise [Bredies et al., 2020b], however current methods are too slow for large-scale applications.

Let Ω be an open bounded spatial domain. At the heart of the analysis of Bredies et. al. is the interplay between measures $\rho(t,x)$ defined on the time-space cylinder $[0,1]\times\overline{\Omega}$, and measures $\sigma(\gamma)$ defined on the space of curves $\gamma=(h,\xi)$ with mass $h\in\mathscr{C}([0,1])$ and trajectory $\xi\in\mathscr{C}([0,1];\overline{\Omega})$. We can think of ρ as representing the evolving physical volume that can be observed with a microscope, whereas σ more efficiently represents a collection of (trajectories

of) particles, that we wish to reconstruct. In the present work, building upon the above-mentioned interplay, we focus on variational problems in the space of measures σ on paths. Such a point of view is fruitful, as the problem more closely resembles the Beurling-LASSO which is now well-understood [Bredies and Pikkarainen, 2013, Azaïs et al., 2015, Duval and Peyré, 2015, Poon et al., 2021]. But its main advantage is that it paves the way for "off-the-grid" numerical methods when solving such dynamical inverse problems. Recent works in the field of sparse spike recovery [Bredies and Pikkarainen, 2013, Boyd et al., 2017, Denoyelle et al., 2019] have demonstrated that it is possible to design efficient numerical solvers without reconstructing the unknown on a grid, by exploiting a conditional gradient descent / Frank-Wolfe approach together with a good knowledge of the regularising term. Indeed, the Frank-Wolfe minimisation algorithm and its variants (see the review Jaggi [2013]) build iterates that are convex combinations of the extreme points of level sets of the regulariser; being able to easily encode and handle such extreme points makes it possible to solve variational problems in a continuous (or up to floating point) setting. Moreover, having iterates that are convex combinations of a few extreme points of the level sets of the regulariser is particularly relevant, as it is known in inverse problems that some solutions have precisely that structure when the observation consists of a finite number of linear measurements [Unser et al., 2017, Bredies and Carioni, 2019, Boyer et al., 2019].

In the particular case of dynamical super-resolution using optimal transport regularisation, those extreme points are described by the key results of Bredies et al. [2021, 2020a] (see also Ambrosio et al. [2008]) as measures supported on the above-mentioned curves $\gamma = (h, \xi)\colon [0,1] \to \mathbb{R} \times \overline{\Omega}$. Even though only a small subset of $L^\infty([0,1]\,;\mathbb{R} \times \overline{\Omega})$ is involved, optimising over such a set is a challenging task that is addressed in Bredies et al. [2020b] in the case of balanced optimal transport. Yet, the computation times are hardly compatible with a practical applications.

The aim of the present paper is to propose an alternative optimisation method which keeps the key properties of Bredies et al. [2020b] (i.e. iterates made of a combination of atoms in a continuous setting) while drastically improving the computation time.

1.1 Outline of the paper

In Section 2 we gather several results on the probabilistic representation of the solutions of the continuity equation that are useful to understand the perspective of measures on paths. Section 3 studies the case of the Benamou-Brenier regularisation: we explain how the original problem is equivalent to a variational problem on the space of paths, and we describe a new algorithm to solve it, which is a particular instance of the framework introduced in Section 5. Our main numerical contribution is to reformulate the main Frank-Wolfe step as a shortest-path problem on an ordered, weighted, directed, acyclic graph. This is demonstrated in detail in Section 3.4. Section 4 describes the unbalanced case, corresponding to the Wasserstein-Fisher-Rao regularisation: for clarity, we have decided to present both cases separately.

The main contributions of this work are presented in Section 5 where we introduce our general variational framework, numerical algorithm, and convergence theorem. The proofs of convergence for inexact and stochastic Frank-Wolfe are given in Section 6 where we confirm that our proposed algorithm converges almost-surely to an optimal reconstruction of our variational model. The stochastic algorithm is similar to that considered by Silveti-Falls et al. [2020], although we compute the step-size adaptively rather than fixing it a priori. Our novel reformulation to the shortest-path problem is explained in Section 7, the dynamical programming approach is standard but presented in the current context for clarity.

Finally, in Section 8 we present thorough numerical results demonstrating the properties of stochastic and deterministic variants of our proposed algorithm. Examples with the BB energy are compared with the algorithm of Bredies et al. [2020b], and new unbalanced examples are reconstructed with a WFR penalty.

1.2 Notation

Convex sets and extreme points. Let V be a linear space. For all $\sigma_0, \sigma_1 \in V$, we define the closed line segment between σ_0 and σ_1 as $[\sigma_0, \sigma_1] \stackrel{\text{def.}}{=} \{\, \lambda\sigma_0 + (1-\lambda)\sigma_1 \mid 0 \leq \lambda \leq 1 \,\}$. Similarly, we define the open line segment $]\sigma_0, \sigma_1[\stackrel{\text{def.}}{=} [\sigma_0, \sigma_1] \setminus \{\sigma_0, \sigma_1\}$.

Definition 1.1. *Let $D \subseteq V$. D is called* convex *if $]\sigma_0, \sigma_1[\subset D$ for all $\sigma_0, \sigma_1 \in D$. We say $\sigma \in \operatorname{Ext}(D)$ is an* extreme point *of D if $\sigma \in D$ and there are no points $\sigma_0, \sigma_1 \in D$ such that $\sigma \in]\sigma_0, \sigma_1[$. In other words,*

$$\forall \lambda \in]0,1[\,, \forall \sigma_0, \sigma_1 \in D, \quad (\sigma = \lambda\sigma_0 + (1-\lambda)\sigma_1) \quad \Longrightarrow \quad (\sigma_0 = \sigma_1 = \sigma)\,. \tag{1.1}$$

Furthermore, it is possible to define the notion of face (and elementary face) of a convex set, which extends the notion of extreme point to higher-dimensional sets. We refere to Dubins [1962] for more detail.

Measure spaces. For a separable metric space Γ and Banach space X, we define $\mathscr{C}_b(\Gamma; X)$ to be the set of continuous bounded functions from Γ to X. When $X = \mathbb{R}$, we simply write $\mathscr{C}_b(\Gamma)$. Recall that for any Borel measure σ on Γ, we can define the unique non-negative Borel measure $|\sigma| \in \mathcal{M}^+(\Gamma)$ by

$$|\sigma|(A) \stackrel{\text{def.}}{=} \sup \left\{ \sum_{i=1}^{n} |\sigma(A_i)| \mid n \in \mathbb{N}^*, \{A_1, \ldots, A_n\} \text{ Borel partition of } A \right\} \tag{1.2}$$

for all Borel measurable sets $A \subset \Gamma$. We denote by $\mathcal{M}(\Gamma)$ the space of signed Borel measures σ with finite total variation, i.e. $\|\sigma\| \stackrel{\text{def.}}{=} \int_\Gamma \mathrm{d}|\sigma| < +\infty$. The total variation $\|\cdot\|$ defines a norm on $\mathcal{M}(\Gamma)$, but it is sometimes more convenient to use the *narrow topology*, i.e. the weakest topology on $\mathcal{M}(\Gamma)$ which makes the integration against continuous bounded functions a continuous linear form. The narrow topology is equivalent to the weak-* topology on $(\mathscr{C}_b(\Gamma))'$, in particular, a sequence $\{\sigma^n\}_{n\in\mathbb{N}} \subset \mathcal{M}(\Gamma)$ converges to $\sigma^* \in \mathcal{M}(\Gamma)$ in the narrow topology (denoted $\sigma^n \stackrel{*}{\rightharpoonup} \sigma$) if

$$\forall \phi \in \mathscr{C}_b(\Gamma), \quad \lim_{n\to+\infty} \int_\Gamma \phi \,\mathrm{d}\sigma^n = \int_\Gamma \phi \,\mathrm{d}\sigma^*. \tag{1.3}$$

The support of $\sigma \in \mathcal{M}^+(\Gamma)$ is defined as

$$\operatorname{supp}(\sigma) \stackrel{\text{def.}}{=} \left(\bigcup \{\, U \mid \sigma(U) = 0,\ U \text{ is open} \,\} \right)^c. \tag{1.4}$$

This is a closed set satisfying $\sigma(\operatorname{supp}(\sigma)) = \sigma(\Gamma)$.

Function domains. In this work we consider two measure domains, the time-space cylinder

$$[0,1] \times \overline{\Omega} \quad \text{for an open, bounded, convex domain} \quad \Omega \subseteq \mathbb{R}^d,\ d \geq 1 \tag{1.5}$$

and a closed set Γ of continuous (weighted) curves which are viewed as pairs $\gamma = (h, \xi)$ where $h(t)$ is the mass at time t and $\xi(t)$ is the location. A formal definition will be given in Lemma 2.1. In the time-space cylinder we will denote measures $\rho \in \mathcal{M}([0,1] \times \overline{\Omega})$ with test functions $\psi \in \mathscr{C}([0,1] \times \overline{\Omega})$, and similarly on the space of curves $\sigma \in \mathcal{M}(\Gamma)$ and $\phi \in \mathscr{C}_b(\Gamma)$.

Narrowly continuous measures. An important subspace of $\mathcal{M}([0,1] \times \overline{\Omega})$ is the space of narrowly continuous measures. With a slight abuse of the standard notation, we will say $\rho \in \mathscr{C}_w([0,1]\,; \mathcal{M}(\overline{\Omega})) \subset \mathcal{M}([0,1] \times \overline{\Omega})$ if there exists a map $t \mapsto \rho_t \in \mathcal{M}(\overline{\Omega})$ (informally, $\rho_t = \rho(t, \cdot)$ is a "time slice" of ρ) such that

$$\forall \psi \in \mathscr{C}(\overline{\Omega}), \qquad \left[t \mapsto \int_{\overline{\Omega}} \psi(x) \,\mathrm{d}\rho_t(x) \right] \in \mathscr{C}([0,1]) \tag{1.6}$$

and

$$\forall \psi \in L^1_\rho([0,1]\times\overline{\Omega}), \qquad \int_{[0,1]\times\overline{\Omega}} \psi(t,x)\,\mathrm{d}\rho(t,x) = \int_0^1 \left(\int_{\overline{\Omega}} \psi(t,x)\,\mathrm{d}\rho_t(x) \right) \mathrm{d}t. \tag{1.7}$$

Given a measure on paths $\sigma \in \mathcal{M}(\Gamma)$ such that $\int_\Gamma \|h\|_\infty \,\mathrm{d}\sigma(h,\xi) < +\infty$, one may define the family of measures $(e_t)_\sharp\sigma \in \mathcal{M}(\overline{\Omega})$, for $t \in [0,1]$, by

$$\forall \psi \in \mathscr{C}(\overline{\Omega}), \qquad \int_{\overline{\Omega}} \psi(x)\,\mathrm{d}[(e_t)_\sharp\sigma](x) \overset{\text{def.}}{=} \int_\Gamma h(t)\psi(\xi(t))\,\mathrm{d}\sigma(h,\xi). \tag{1.8}$$

Formally, $(e_t)_\sharp\sigma$ is the image measure of σ by the evaluation at time t. That family is narrowly continuous, and as we explain in Theorem 2.2 below, it is the evaluation (disintegration) of some measure $\Theta(\sigma) \in \mathscr{C}_w([0,1]\,;\mathcal{M}(\overline{\Omega}))$.

The continuity equation.

Definition 1.2. *Let $\rho \in \mathcal{M}([0,1]\times\overline{\Omega})$ be a measure. We say that ρ satisfies the* continuity equation *if there exists $v \in L^1_{|\rho|}([0,1]\times\overline{\Omega};\mathbb{R}^d)$, $g \in L^1_{|\rho|}([0,1]\times\overline{\Omega})$ such that*

$$\forall \psi \in \mathscr{C}^1_c(]0,1[\times\overline{\Omega}), \qquad \int [\partial_t\psi + \nabla\psi\cdot v + \psi g]\,\mathrm{d}\rho = 0. \tag{1.9}$$

Lower semi-continuity. A function $w\colon \Gamma \to \mathbb{R}\cup\{+\infty\}$ is called *lower semi-continuous* if the sets $\{\,\gamma\in\Gamma \mid w(\gamma)\le\lambda\,\}$ are closed for each $\lambda\in\mathbb{R}$. As Γ is a metric space, it is equivalent to say that for all $\gamma\in\Gamma$ and sequences $\gamma_n\to\gamma$, $w(\gamma)\le \liminf_{n\to+\infty} w(\gamma_n)$.

2 Preliminary results

In this work we define Γ_0 to be the set of continuous weighted paths in $\overline{\Omega}$, modelling particles with (varying) mass $h(t)$ at location $\xi(t)$, defined by

$$\Gamma_0 = \left\{\, \gamma=(h,\xi) \mid h\in\mathscr{C}([0,1]),\ \xi\colon [0,1]\to\overline{\Omega},\ \xi|_{\{h\neq0\}} \text{ is continuous} \,\right\}. \tag{2.1}$$

For technical reasons we permit curves ξ which may not be continuous at points t where $h(t)=0$. Heuristically, the location $\xi(t)$ is not necessarily important if the particle has no mass and cannot be observed. The other technical requirements for the set Γ_0 is that it must be a complete separable metric space, as explored in the following lemma. We follow the suggestion of [Bredies et al., 2020a, Prop. 3.6] who used the flat metric on the space $\left\{\, h\delta_\xi \in \mathcal{M}^+([0,1]\times\overline{\Omega}) \mid (h,\xi)\in\Gamma,\ h\ge 0 \,\right\}$.

Lemma 2.1. *Define $\mathrm{d}_\Gamma\colon \Gamma_0\times\Gamma_0 \to [0,+\infty[$ by*

$$\mathrm{d}_\Gamma((h_1,\xi_1),(h_2,\xi_2)) \overset{\text{def.}}{=} \sup_{t\in[0,1]} \mathrm{d}_F((h_1(t),\xi_1(t)),(h_2(t),\xi_2(t))) \qquad \textit{where} \tag{2.2}$$

$$\mathrm{d}_F((r_1,x_1),(r_2,x_2)) \overset{\text{def.}}{=} \begin{cases} |r_1|+|r_2| & r_1r_2\le 0 \text{ or } |x_1-x_2|\ge 2 \\ |r_1-r_2|+\min(|r_1|,|r_2|)|x_1-x_2| & \textit{else,} \end{cases} \tag{2.3}$$

then $(\Gamma_0/\!\sim, \mathrm{d}_\Gamma)$ is a complete separable metric space where

$$(h_1,\xi_1)\sim(h_2,\xi_2) \iff h_1=h_2 \quad \textit{and} \quad \forall t\in\{h_1\neq0\},\quad \xi_1(t)=\xi_2(t). \tag{2.4}$$

Convergence of a sequence $\gamma_n=(h_n,\xi_n)\in\Gamma_0$ in the metric d_Γ can equivalently be stated as:

$$\left[\gamma_n \overset{\mathrm{d}_\Gamma}{\to} (h,\xi)\right] \iff \left[h_n\to h \text{ in } \mathscr{C}([0,1]) \text{ and for all } \varepsilon>0,\ \xi_n\to\xi \text{ in } \mathscr{C}(\{|h|\ge\varepsilon\})\right] \tag{2.5}$$

Furthermore, for any $\psi\in\mathscr{C}([0,1]\times\overline{\Omega})$, we have $\Psi\in\mathscr{C}([0,1]\times\Gamma_0)$ where

$$\forall t\in[0,1]\,,\ (h,\xi)\in\Gamma_0, \qquad \Psi(t,h,\xi)\overset{\text{def.}}{=} h(t)\psi(t,\xi(t)). \tag{2.6}$$

Proof. Metric properties. To see that $(\Gamma_0/\sim, \mathrm{d}_\Gamma)$ is a metric space, it is sufficient to show that $(\mathbb{R}\times\overline{\Omega}/\sim, \mathrm{d}_F)$ is a metric space. Following the argument of [Bredies et al., 2020a, Lemma 3.3] in the case $h \geq 0$, observe that for all $r_1, r_2 \in \mathbb{R}$, $x_1, x_2 \in \overline{\Omega}$

$$\mathrm{d}_F((r_1,x_1),(r_2,x_2)) = \sup_{c_1,c_2\in\mathbb{R}} \{\, r_1c_1 - r_2c_2 \mid |c_1|,|c_2| \leq 1,\ |c_1-c_2| \leq |x_1-x_2| \,\}. \tag{2.7}$$

For $\min(r_1,r_2) \geq 0$, this has already been shown in Bredies et al. [2020a]. The case $\max(r_1,r_2) \leq 0$ is equivalent, therefore it is true for all $r_1r_2 \geq 0$. For the final case, without loss of generality $r_1 \geq 0 \geq r_2$, so the supremum is achieved by $c_1 = c_2 = 1$.

It is clear that $\mathrm{d}_\Gamma \geq 0$. To verify the triangle inequality, let $(r_1,x_1),(r_2,x_2),(r_3,x_3) \in \mathbb{R}\times\overline{\Omega}$. Let $c_1, c_3 \in [-1,1]$ such that $|c_1-c_3| \leq |x_1-x_3|$. By symmetry of d_F, we may assume without loss of generality that $c_1 \leq c_3$. Then

$$c_3 - c_1 \leq |x_1-x_3| \leq |x_2-x_1| + |x_3-x_2|, \tag{2.8}$$

hence there exists $c_2 \in [c_1,c_3]$ such that $|c_2-c_1| \leq |x_2-x_1|$ and $|c_2-c_3| \leq |x_2-x_3|$. Thus,

$$r_1c_1 - r_3c_3 = (r_1c_1 - r_2c_2) + (r_2c_2 - r_3c_3) \leq \mathrm{d}_F((r_1,x_1),(r_2,x_2)) + \mathrm{d}_F((r_2,x_2),(r_3,x_3)). \tag{2.9}$$

Taking the supremum over c_1, c_3 yields

$$\mathrm{d}_F((r_1,x_1),(r_3,x_3)) \leq \mathrm{d}_F((r_1,x_1),(r_2,x_2)) + \mathrm{d}_F((r_2,x_2),(r_3,x_3). \tag{2.10}$$

Hence d_F satisfies the triangle inequality and so does d_Γ.

Finally we must check $\mathrm{d}_\Gamma(\gamma_1,\gamma_2) = 0 \iff \gamma_1 \sim \gamma_2$. To see $h_1 = h_2$, observe for any $t \in [0,1]$ either $|h_1(t)| + |h_2(t)| = 0$, or $|h_1(t) - h_2(t)| = 0$, but in any case, $h_1(t) = h_2(t)$. Now whenever $h_1(t) \neq 0$, we must be in the case $|\xi_1(t) - \xi_2(t)| \leq 2$, therefore $|h_1(t)||\xi_1(t) - \xi_2(t)| = 0$, hence $\xi_1(t) = \xi_2(t)$ as required.

Convergence. Fix a sequence $\gamma_n = (h_n,\xi_n) \in \Gamma_0$ and point $\gamma = (h,\xi) \in \Gamma_0$. Note that for all $r_i \in \mathbb{R}$, $x_i \in \overline{\Omega}$,

$$|r_1 - r_2| \leq \mathrm{d}_F((r_1,x_1),(r_2,x_2)) \leq |r_1| + |r_2|. \tag{2.11}$$

Suppose $\mathrm{d}_\Gamma(\gamma_n,\gamma) \to 0$, then we immediately have $\|h_n - h\|_\infty \to 0$. Also, for any $\varepsilon > 0$ choose $N_\varepsilon \in \mathbb{N}$ such that

$$\forall n \geq N_\varepsilon, \qquad \|h_n - h\|_\infty \leq \frac{\varepsilon}{2} \quad \text{and} \quad \mathrm{d}_\Gamma(\gamma_n,\gamma) \leq \frac{\varepsilon}{2}. \tag{2.12}$$

Observe that for all $t \in \{|h| \geq \varepsilon\}$ and $n \geq N_\varepsilon$, if $|\xi_n(t) - \xi(t)| \geq 2$, then $\mathrm{d}_\Gamma(\gamma_n,\gamma) \geq |h_n(t)| + |h(t)| > \varepsilon$, contradicting the choice of N_ε. Therefore we have the uniform bound

$$\sup_{|h(t)|\geq\varepsilon} |\xi_n(t) - \xi(t)| \leq \sup_{|h(t)|\geq\varepsilon} \frac{2}{\varepsilon}\min(|h_n(t)|,|h(t)|)|\xi_n(t) - \xi(t)| \leq \frac{2}{\varepsilon}\,\mathrm{d}_\Gamma(\gamma_n,\gamma). \tag{2.13}$$

This concludes the " $\Longrightarrow$ " direction of (2.5). Conversely, suppose $\|h_n - h\|_\infty \to 0$ and for any $\varepsilon > 0$, $\|\xi_n - \xi\|_{L^\infty(\{|h|\geq\varepsilon\})} \to 0$. Fix $\varepsilon > 0$ and $N_\varepsilon \in \mathbb{N}$ such that for all

$$\forall n \geq N_\varepsilon, \qquad \|h_n - h\|_\infty \leq \frac{\varepsilon}{2} \quad \text{and} \quad \|\xi_n - \xi\|_{L^\infty(\{|h|\geq\varepsilon\})} \leq \min(2,\varepsilon). \tag{2.14}$$

Then, from (2.11), for all $n \geq N_\varepsilon$ we have

$$\mathrm{d}_\Gamma(\gamma_n,\gamma) \leq \sup_{t\in[0,1]} \begin{cases} |h_n(t)| + |h(t)| & \text{if } |h(t)|, \\ \mathrm{d}_F((h_n(t),\xi_n(t)),(h(t),\xi(t))) & \text{else} \end{cases} \tag{2.15}$$

$$\leq \sup_{t\in[0,1]} \begin{cases} 5\varepsilon/2 & \text{if } |h(t)|, \\ \varepsilon/2 + \|h\|_\infty\varepsilon & \text{else.} \end{cases} \tag{2.16}$$

In either case we have $\limsup_{n\to+\infty} \mathrm{d}_\Gamma(\gamma_n,\gamma) \le O(\varepsilon)$ for all $\varepsilon > 0$, therefore $\mathrm{d}_\Gamma(\gamma_n,\gamma) \to 0$ as required.

Separability. Recall by the Weierstrass approximation theorem that $\mathscr{C}([0,1]\,;[0,1]\times\overline{\Omega})$ is separable, let $\{\gamma_n = (h_n,\xi_n)\}_{n\in\mathbb{N}}$ be a dense set (in the $\mathscr{C}$ sense). Fix $\gamma = (h,\xi) \in \Gamma_0$. Denote the closed domains $U_k = \{k|h| \ge 1\}$. By the Urysohn-Brouwer lemma, for each $k \in \mathbb{N}$ there exists an extension of ξ, say $\widehat{\xi}_k \in \mathscr{C}([0,1]\,;\overline{\Omega})$, such that $\xi = \widehat{\xi}_k$ on U_k. Therefore there exists a sequence n_k such that

$$\forall k \in \mathbb{N}, \qquad \|h_{n_k} - h\|_\infty \le \frac{1}{k} \quad \text{and} \quad \|\xi_{n_k} - \xi\|_{L^\infty(\{k|h|\ge 1\})} \le \frac{1}{k}. \tag{2.17}$$

From (2.5), it is clear that $\gamma_{n_k} \xrightarrow{\mathrm{d}_\Gamma} \gamma$, therefore $\{\gamma_n\}_{n\in\mathbb{N}}$ is also countable and dense in Γ_0.

Completeness. Let $\gamma_n = (h_n,\xi_n)$ be an arbitrary d_Γ-Cauchy sequence, we are required to show that there exists $\gamma \in \Gamma_0$ such that $\mathrm{d}_\Gamma(\gamma_n,\gamma) \to 0$. By the lower bound in (2.11), we have

$$\|h_n - h_m\|_\infty \le \mathrm{d}_\Gamma(\gamma_n,\gamma_m), \tag{2.18}$$

so h_n is Cauchy in $\mathscr{C}([0,1])$. As $\mathscr{C}([0,1])$ is complete, there must be a limit point $h \in \mathscr{C}([0,1])$ such that $h_n \xrightarrow{\mathscr{C}} h$. Repeating the argument in (2.13) confirms that ξ_n is also Cauchy in $\mathscr{C}(\{|h| \ge \varepsilon\};\overline{\Omega})$ for each $\varepsilon > 0$, therefore there is a (unique) curve $\xi \in \mathscr{C}(\{h \ne 0\};\overline{\Omega})$ such that $\xi_n \to \xi$ in each $\mathscr{C}(\{|h| \ge \varepsilon\};\overline{\Omega})$, $\varepsilon > 0$. We may extend ξ by a constant $x_0 \in \Omega$ on $\{h = 0\}$. Due to (2.5), we obtain that $\gamma_n \to (h,\xi)$.

Continuity. Fix $\psi \in \mathscr{C}([0,1]\times\overline{\Omega})$ and define Ψ as in (2.6). As we have now confirmed that $({}^{\Gamma_0}\!/\!\sim, \mathrm{d}_\Gamma)$ is a metric space, we can use the sequential definitions of continuity. Suppose $\tau_n \in [0,1]$, $\gamma_n = (h_n,\xi_n) \in \Gamma_0$ and $\tau_n \to t$, $\gamma_n \xrightarrow{\mathrm{d}_\Gamma} \gamma$, observe for each n we have

$$\begin{aligned}
|\Psi(\tau_n,\gamma_n) - \Psi(t,\gamma)| &= |h_n(\tau_n)\psi(\tau_n,\xi_n(\tau_n)) - h(t)\psi(t,\xi(t))| && (2.19)\\
&= |(h_n(\tau_n) - h(t) + h(t))\psi(\tau_n,\xi_n(\tau_n)) - h(t)\psi(t,\xi(t))| && (2.20)\\
&\le |h_n(\tau_n) - h(t)|\|\psi\|_\infty + |h(t)||\psi(\tau_n,\xi_n(\tau_n)) - \psi(t,\xi(t))|. && (2.21)\\
&\le [\|h_n - h\|_\infty + |h(\tau_n) - h(t)|]\|\psi\|_\infty + |h(t)||\psi(\tau_n,\xi_n(\tau_n)) - \psi(t,\xi(t))|. && (2.22)
\end{aligned}$$

Now consider the limit of $n \to +\infty$. As $h \in \mathscr{C}([0,1])$, $h(\tau_n) \to h(t)$. The characterisation of convergence in d_Γ in (2.5) also confirms $\|h_n - h\|_\infty \to 0$ and either $h(t) = 0$, or

$$|\xi_n(\tau_n) - \xi(t)| \le |\xi_n(\tau_n) - \xi(\tau_n)| + |\xi(\tau_n) - \xi(t)| \to 0. \tag{2.23}$$

In either case, we see that

$$\limsup_{n\to+\infty} |\Psi(\tau_n,\gamma_n) - \Psi(t,\gamma)| \le \limsup_{n\to+\infty} |h(t)||\psi(\tau_n,\xi_n(\tau_n)) - \psi(t,\xi(t))| = 0, \tag{2.24}$$

therefore Ψ is continuous at (t,γ). □

A key analytical tool in related prior works (c.f. Bredies et al. [2021], Bredies and Fanzon [2020], Bredies et al. [2020a,b]) is a mapping between measures $\rho \in \mathcal{M}([0,1]\times\overline{\Omega})$ which satisfy the continuity equation, and measures on (weighted) paths $\sigma \in \mathcal{M}(\Gamma_0)$, making some structures become more apparent. We will now recap and expand on previous results. The first theorem collects results from Ambrosio et al. [2008], Bredies et al. [2020a] and provides minor extensions for the scope of the current work.

Theorem 2.2. *Let $\sigma \in \mathcal{M}(\Gamma_0)$. If $\int_{\Gamma_0} \|h\|_1 \,\mathrm{d}|\sigma|(h,\xi) < +\infty$, then there is a unique finite Borel measure $\Theta(\sigma) \in \mathcal{M}([0,1]\times\overline{\Omega})$ such that*

$$\forall \psi \in \mathscr{C}([0,1]\times\overline{\Omega}), \qquad \int_{[0,1]\times\overline{\Omega}} \psi(t,x)\,\mathrm{d}\Theta(\sigma)(t,x) = \int_{\Gamma_0} \left(\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t \right) \mathrm{d}\sigma(h,\xi). \tag{2.25}$$

Moreover,

1. *The mapping* $\Theta\colon \left\{ \sigma \in \mathcal{M}(\Gamma_0) \mid \int_{\Gamma_0} \|h\|_1 \,\mathrm{d}|\sigma| < +\infty \right\} \to \mathcal{M}([0,1]\times\overline{\Omega})$ *is linear.*
2. *Equality* (2.25) *holds for all* $\psi \in L^1_{|\Theta(\sigma)|}([0,1]\times\overline{\Omega})$.
3. *If* $\int_{\Gamma_0} \|h\|_\infty \,\mathrm{d}|\sigma| < +\infty$, *then* $\Theta(\sigma) \in \mathscr{C}_w([0,1]\,;\mathcal{M}(\overline{\Omega}))$.
4. *If there exist Borel measurable functions* $v\colon [0,1]\times\overline{\Omega} \to \mathbb{R}^d$ *and* $g : [0,1]\times\overline{\Omega} \to \mathbb{R}$ *such that*

$$h'(t) = g(t,\xi(t))h(t) \text{ for } \sigma\text{-a.e. } (h,\xi) \text{ and a.e. } t \in]0,1[, \tag{2.26}$$
$$\xi'(t) = v(t,\xi(t)) \text{ for } \sigma\text{-a.e. } (h,\xi) \text{ and a.e. } t \text{ such that } h(t) > 0, \tag{2.27}$$
$$\text{and } \int_{\Gamma_0}\int_0^1 (1 + |v(t,\xi(t))| + |g(t,\xi(t))|)\,|h(t)|\,\mathrm{d}t\,\mathrm{d}|\sigma|(h,\xi) < +\infty, \tag{2.28}$$

then $\int_{\Gamma_0} \|h\|_\infty \,\mathrm{d}|\sigma| < +\infty$ *and* $\Theta(\sigma)$ *satisfies the continuity equation* (1.9).

Conversely, given $\rho \in \mathcal{M}([0,1]\times\overline{\Omega})$, *if* $\rho \geq 0$ *satisfies the continuity equation* (1.9) *and*

$$\int_{[0,1]\times\overline{\Omega}} (1 + |v(t,x)|^2 + |g(t,x)|^2)\,\mathrm{d}\rho(t,x) < +\infty, \tag{2.29}$$

then $\rho = \Theta(\sigma)$ *for some* $\sigma \in \mathcal{M}^+(\Gamma_0)$ *such that* (2.26)-(2.28) *hold and* $\int_{\Gamma_0} \|h\|_\infty \,\mathrm{d}\sigma < +\infty$.

Proof. By Lemma 2.1, for any $\psi \in \mathscr{C}([0,1]\times\overline{\Omega})$, the map

$$(h,\xi) \mapsto \int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t = \int_0^1 \Psi(t,h,\xi)\,\mathrm{d}t \tag{2.30}$$

is continuous (hence Borel) in Γ_0 and dominated by $\|h\|_1\|\psi\|_\infty$, therefore the right-hand side of (2.25) is well-defined. This induces a linear form on $\mathscr{C}([0,1]\times\overline{\Omega})$ which is moreover bounded, since

$$\left| \int_{\Gamma_0}\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\sigma(h,\xi) \right| \leq \|\psi\|_\infty \int_{\Gamma_0} \|h\|_1 \,\mathrm{d}|\sigma|. \tag{2.31}$$

By the Riesz representation theorem, that linear form is represented by a unique Radon measure $\Theta(\sigma) \in \mathcal{M}([0,1]\times\overline{\Omega})$. This confirms that Θ satisfies the required properties for point (1).

Points (2)-(4) have been proved in Bredies et al. [2020a] under the additional assumptions $\sigma \geq 0$ and $\inf_{t\in[0,1]} h(t) \geq 0$. To apply these results, we need the following modified Hahn-Jordan decomposition.

Claim: For any $\sigma \in \mathcal{M}(\Gamma_0)$ with $\int_{\Gamma_0} \|h\|_1 \,\mathrm{d}|\sigma| < +\infty$ there exists $\sigma^+, \sigma^- \in \mathcal{M}^+(\Gamma_0)$ such that

$$\sigma^\pm\left(\left\{ (h,\xi) \in \Gamma_0 \mid \inf_{t\in[0,1]} h(t) < 0 \right\} \right) = 0 \tag{2.32}$$

with $\int_{\Gamma_0} \|h\|_1 \,\mathrm{d}\sigma^\pm < +\infty$ and $\Theta(\sigma) = \Theta(\sigma^+) - \Theta(\sigma^-)$.

Proof of claim. The Hahn-Jordan decomposition gives $\sigma = \max(0,\sigma) - \max(0,-\sigma)$, therefore it is sufficient to consider the case $\sigma \geq 0$. Define the maps $T^\pm\colon \Gamma_0 \to \Gamma_0$ by

$$\forall (h,\xi) \in \Gamma_0, \qquad T^\pm(h,\xi) \overset{\text{def.}}{=} (\max(0,\pm h),\xi). \tag{2.33}$$

Because $\mathrm{d}_\Gamma(T^\pm(\gamma_1), T^\pm(\gamma_2)) \leq \mathrm{d}_\Gamma(\gamma_1,\gamma_2)$, $T^\pm$ are continuous (therefore Borel), and we can define the image measures $\sigma^\pm \overset{\text{def.}}{=} (T^\pm)_\sharp\sigma$. Since the push-forward operation

does not increase the total variation, we have $\sigma^+, \sigma^- \in \mathcal{M}^+(\Gamma_0)$. Moreover, (2.32) holds, and by construction of the image measure

$$\int_{\Gamma_0} \phi(h,\xi)\,\mathrm{d}\sigma^{\pm} = \int_{\Gamma_0} \phi(\max(0,\pm h),\xi)\,\mathrm{d}\sigma \tag{2.34}$$

for all $\phi \in \mathscr{C}_b(\Gamma_0)$, and by monotone convergence

$$\int_{\Gamma_0} \|h\|_1\,\mathrm{d}\sigma^{\pm} = \int_{\Gamma_0} \|\max(0,\pm h)\|_1\,\mathrm{d}\sigma \leq \int_{\Gamma_0} \|h\|_1\,\mathrm{d}\sigma. \tag{2.35}$$

Finally, we confirm that for all $\psi \in \mathscr{C}([0,1]\times\overline{\Omega})$,

$$\begin{aligned}
\int_{[0,1]\times\overline{\Omega}} \psi(t,x)\,\mathrm{d}\Theta(\sigma) &= \int_{\Gamma_0}\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\sigma && (2.36)\\
&= \int_{\Gamma_0}\int_0^1 [\max(0,h(t)) - \max(0,-h(t))]\psi(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\sigma && (2.37)\\
&= \int_{\Gamma_0}\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}(\sigma^+ - \sigma^-) && (2.38)\\
&= \int_{[0,1]\times\overline{\Omega}} \psi(t,x)\,\mathrm{d}(\Theta(\sigma^+) - \Theta(\sigma^-)) && (2.39)
\end{aligned}$$

as required. □

With this decomposition, we can apply the results of Bredies et al. [2020a] to each $\sigma^{\pm}$. Point (2) is a direct consequence of [Bredies et al., 2020a, Lemma 4.4]. Under the assumptions of point (3), the same lemma also guarantees the existence of a disintegration $\rho_t^{\pm}$ of $\Theta(\sigma^{\pm})$, in the sense of (1.7). In particular,

$$\forall \psi \in \mathscr{C}(\overline{\Omega}),\ t \in [0,1]\,, \qquad \int_{\overline{\Omega}} \psi(x)\,\mathrm{d}\rho_t^{\pm}(x) = \int_{\Gamma_0} h(t)\psi(\xi(t))\,\mathrm{d}\sigma^{\pm}(h,\xi). \tag{2.40}$$

The same property holds for $\Theta(\sigma)$ by linearity. Point (3) requires $t \mapsto \int_{\overline{\Omega}} \psi\,\mathrm{d}(\rho_t^+ - \rho_t^-)$ to be continuous for all $\psi \in \mathscr{C}(\overline{\Omega})$. By Lemma 2.1, for each $t \in [0,1]$ the function $\Psi(t,h,\xi) \overset{\text{def.}}{=} h(t)\psi(\xi(t))$ is continuous on $[0,1]\times\Gamma_0$ and

$$\left|\int_{\overline{\Omega}} \psi\,\mathrm{d}(\rho_\tau - \rho_t)\right| \leq \int_{\Gamma_0} |h(\tau)\psi(\xi(\tau)) - h(t)\psi(\xi(t))|\,\mathrm{d}|\sigma|(h,\xi) = \int_{\Gamma_0} |\Psi(\tau,\gamma) - \Psi(t,\gamma)|\,\mathrm{d}|\sigma|(\gamma). \tag{2.41}$$

The integrand is pointwise bounded by $2\|h\|_\infty\|\psi\|_\infty$, therefore we conclude that the limit of the integral as $\tau \to t$ is 0 by dominated convergence.

Point (4) and its converse are addressed by [Bredies et al., 2020a, Thm. 4.2]. For the forward direction, we again consider the modified Hahn-Jordan decomposition. The assumptions (2.26)-(2.28) are only assumed to hold for almost every t and h, therefore they also hold for the choice of measures $\sigma^{\pm}$ given in the claim replacing h with $\max(0,\pm h)$. We can then apply [Thm. 4.2 Bredies et al., 2020a] to each component and sum for the result. The converse is exactly the statement of [Thm. 4.2 Bredies et al., 2020a] because we assume $\rho \geq 0$. □

The previous theorem makes precise statements about when $\Theta(\sigma)$ satisfies particular smoothness properties, the assumption $\|\sigma\| < +\infty$ is too weak for any useful properties. In this work we would like Θ to be narrowly continuous and only to consider measures σ such that $\Theta(\sigma) \in \mathscr{C}_w([0,1]\,;\mathcal{M}(\overline{\Omega}))$. The following lemma shows that we can choose a subset of Γ_0 which only removes un-desired elements from the image of Θ. Recall $\|\sigma\| < +\infty$ for all $\sigma \in \mathcal{M}(\Gamma_0)$, if $\Gamma \subset \Gamma_0$ is sufficiently "small", then this implicit assumption is sufficient for $\Theta|_{\mathcal{M}(\Gamma)}$ to have the desired properties.

Lemma 2.3. *Define the set*

$$\text{for some} \quad p \in [1, +\infty], \qquad \Gamma_p \stackrel{\text{def.}}{=} \left\{ \gamma = (h, \xi) \in \Gamma_0 \mid \|h\|_p \leq 1 \right\}, \tag{2.42}$$

then

$$\left\{ \Theta(\sigma) \mid \sigma \in \mathcal{M}(\Gamma_0),\ \int_{\Gamma_0} \|h\|_p \, \mathrm{d}|\sigma| < +\infty \right\} = \{ \Theta(\widehat{\sigma}) \mid \widehat{\sigma} \in \mathcal{M}(\Gamma_p) \} \tag{2.43}$$

and $\Theta \colon \mathcal{M}(\Gamma_p) \to \mathcal{M}([0,1] \times \overline{\Omega})$ *is narrowly continuous.*

Furthermore, if $p = +\infty$, *then* $\forall t \in [0,1]$, $(e_t)_\sharp \colon \mathcal{M}(\Gamma_\infty) \to \mathcal{M}(\overline{\Omega})$ *is also narrowly continuous.*

In particular, the sequentially we have that, for any sequence $\sigma^n \overset{*}{\rightharpoonup} \sigma$ *narrowly in* $\mathcal{M}(\Gamma_p)$*:*

$$\text{for all } p \in [1, +\infty], \qquad \Theta(\sigma^n) \overset{*}{\rightharpoonup} \Theta(\sigma) \text{ narrowly in } \mathcal{M}([0,1] \times \overline{\Omega}), \tag{2.44}$$

$$\text{if } p = +\infty,\ \forall t \in [0,1], \qquad (e_t)_\sharp \sigma^n \overset{*}{\rightharpoonup} (e_t)_\sharp \sigma \text{ narrowly in } \mathcal{M}(\overline{\Omega}). \tag{2.45}$$

Proof. The "$\supset$" inclusion is clear: if $\widehat{\sigma} \in \mathcal{M}(\Gamma_p)$, extend $\widehat{\sigma}$ by 0 such that $\widehat{\sigma} \in \mathcal{M}(\Gamma_0)$, then

$$\int_{\Gamma_0} \|h\|_p \, \mathrm{d}|\widehat{\sigma}| \leq \int_{\Gamma_p} 1 \, \mathrm{d}|\widehat{\sigma}| = \|\widehat{\sigma}\| < +\infty \tag{2.46}$$

as required. Conversely, suppose $\int_{\Gamma_0} \|h\|_p \, \mathrm{d}|\sigma| < +\infty$. Note as $h \mapsto \max(1, \|h\|_p)$ and $h \mapsto \frac{h}{\max(1, \|h\|_p)}$ are continuous in $\mathscr{C}([0,1])$ and $\mathrm{d}_\Gamma(\gamma_1, \gamma_2) \geq \|h_1 - h_2\|_\infty$, both $(h, \xi) \mapsto \max(1, \|h\|_p)$ and $(h, \xi) \mapsto (\frac{h}{\max(1, \|h\|_p)}, \xi)$ are continuous w.r.t. d_Γ. Define $\widehat{\sigma}$ through the (rescaled) push-forward $\widehat{\sigma}(h, \xi) = \max(1, \|h\|_p) \cdot T_\sharp \sigma(h, \xi)$ where $T \colon \Gamma_0 \to \Gamma_p$ is defined by the map $T(h, \xi) \stackrel{\text{def.}}{=} \left(\frac{h}{\max(1, \|h\|_p)}, \xi \right)$. T is Borel measurable, therefore $\widehat{\sigma}$ is a Borel measure which satisfies

$$\forall \phi \in \mathscr{C}_b(\Gamma_p), \qquad \int_{\Gamma_p} \phi \, \mathrm{d}\widehat{\sigma} = \int_{\Gamma_0} \max(1, \|h\|_p) \phi \left(\frac{h}{\max(1, \|h\|_p)}, \xi \right) \mathrm{d}\sigma(h, \xi). \tag{2.47}$$

Furthermore, $\widehat{\sigma} \in \mathcal{M}(\Gamma_p)$ as

$$\left| \int_{\Gamma_p} \phi \, \mathrm{d}\widehat{\sigma} \right| \leq \|\phi\|_\infty \int_{\Gamma_0} (1 + \|h\|_p) \, \mathrm{d}|\sigma|(h, \xi), \tag{2.48}$$

so $\|\widehat{\sigma}\| < +\infty$. The last step is to confirm $\Theta(\widehat{\sigma}) = \Theta(\sigma)$. For any $\psi \in \mathscr{C}([0,1] \times \overline{\Omega})$ observe

$$\int_{[0,1] \times \overline{\Omega}} \psi(t, x) \, \mathrm{d}\Theta(\widehat{\sigma})(t, x) = \int_{\Gamma_p} \left(\int_0^1 h(t) \psi(t, \xi(t)) \, \mathrm{d}t \right) \mathrm{d}\widehat{\sigma}(h, \xi) \tag{2.49}$$

$$= \int_{\Gamma_0} \max(1, \|h\|_p) \left(\int_0^1 \frac{h(t)}{\max(1, \|h\|_p)} \psi(t, \xi(t)) \, \mathrm{d}t \right) \mathrm{d}\sigma(h, \xi) \tag{2.50}$$

$$= \int_{[0,1] \times \overline{\Omega}} \psi(t, x) \, \mathrm{d}\Theta(\sigma)(t, x). \tag{2.51}$$

This shows the "$\subset$" direction, therefore we conclude the equality in (2.43). For continuity of Θ, in the $p \geq 1$ context [Thm. 1, Sec. 1.6 Yosida, 1980] states that Θ is continuous in $(\mathscr{C}([0,1] \times \overline{\Omega}))'$ if and only if for every $\psi \in \mathscr{C}([0,1] \times \overline{\Omega})$ there exists a finite collection $\Phi \subset \mathscr{C}_b(\Gamma_p)$ such that

$$\forall \sigma \in \mathcal{M}(\Gamma_p), \qquad \left| \int_{[0,1] \times \overline{\Omega}} \psi(t, x) \, \mathrm{d}\Theta(\sigma) \right| \leq \max_{\phi \in \Phi} \left| \int_{\Gamma_p} \phi(h, \xi) \, \mathrm{d}\sigma(h, \xi) \right|. \tag{2.52}$$

From Lemma 2.1, define $\Psi(t,h,\xi) = h(t)\psi(t,\xi(t))$, then $\phi(h,\xi) \overset{\text{def.}}{=} \int_0^1 \Psi(t,h,\xi)\,\mathrm{d}t$ is continuous and bounded as $|\phi(h,\xi)| \leq \|\psi\|_\infty \|h\|_1 \leq \|\psi\|_\infty$ for each $(h,\xi) \in \Gamma_p$, $p \geq 1$. This confirms the continuity of Θ because

$$\int_{[0,1]\times\overline{\Omega}} \psi(t,x)\,\mathrm{d}\Theta(\sigma) = \int_{\Gamma_p} \left(\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t \right) \mathrm{d}\sigma(h,\xi) = \int_{\Gamma_p} \phi(h,\xi)\,\mathrm{d}\sigma(h,\xi) \tag{2.53}$$

for each $\sigma \in \mathcal{M}(\Gamma_p)$, therefore $\Phi = \{\phi\}$ is sufficient. The $p = +\infty$ case is special because for each $t \in [0,1]$, the map $\phi(h,\xi) \overset{\text{def.}}{=} \Psi(t,h,\xi)$ is continuous and bounded by $\|\psi\|_\infty$. This leads to

$$\int_{\overline{\Omega}} \psi(t,x)\,\mathrm{d}(e_t)_\sharp\sigma = \int_{\Gamma_\infty} (h(t)\psi(t,\xi(t)))\,\mathrm{d}\sigma(h,\xi) = \int_{\Gamma_\infty} \phi(h,\xi)\,\mathrm{d}\sigma(h,\xi), \tag{2.54}$$

so we conclude similarly that $(e_t)_\sharp$ is continuous in $(\mathscr{C}(\overline{\Omega}))'$ as required. □

In the rest of this work we want to work on a domain $D \subset \mathcal{M}^+(\Gamma_0)$ such that $\Theta|_D$ has nice properties with respect to the space $\mathscr{C}_w([0,1]\,;\mathcal{M}(\overline{\Omega}))$. In particular, Theorem 2.2 shows that we may choose $D \subset \left\{ \sigma \in \mathcal{M}^+(\Gamma_0) \mid \int_{\Gamma_0} \|h\|_\infty \,\mathrm{d}\sigma < \infty \right\}$. Lemma 2.3 shows that it is equivalent to consider $D \subset \mathcal{M}^+(\Gamma_\infty)$, from a numerical perspective there is very little difference. Analytically we will always choose $D \subset \mathcal{M}^+(\Gamma)$ for Γ a closed subset of Γ_∞ in order to have the additional continuity properties.

Remark 2.1. *Note that the definition of Γ_p is consistent with the relation $\sim$, so $(\Gamma_p/\sim, \mathrm{d}_\Gamma)$ is also a metric space. In the rest of the paper, we require that Γ si a closed subset of $\Gamma_\infty/\sim$ in order for it to be a complete separable metric space.*

3 The balanced Benamou-Brenier example

The purpose of this section is to motivate our framework using a specific example which has been studied in the seminal work of Bredies et al. [2021, 2020b]. Following Bredies et al. [2021, 2020b], we define for $\alpha, \beta > 0$ the Benamou-Brenier penalty $\mathcal{W}\colon \mathcal{M}^+([0,1]\times\overline{\Omega}) \to \mathbb{R}\cup\{+\infty\}$ by

$$\mathcal{W}(\rho) \overset{\text{def.}}{=} \inf_{v \in L^2_\rho([0,1]\times\overline{\Omega};\mathbb{R}^d)} \left\{ \int_{[0,1]\times\overline{\Omega}} \left[\alpha + \frac{\beta}{2}|v|^2 \right] \mathrm{d}\rho \quad \text{s. t.} \quad \partial_t\rho + \operatorname{div}(v\rho) = 0 \right\}, \tag{3.1}$$

where the continuity equation is satisfied in the sense of (1.9). The function with $\alpha = 0$, $\beta = 1$ is referred to as the Benamou-Brenier energy whose main properties can be found in [Santambrogio, 2015, Sec. 5.3.1]. By Theorem 2.2, if $\mathcal{W}(\rho) < +\infty$, then $\rho \in \mathscr{C}_w([0,1]\,;\mathcal{M}(\overline{\Omega}))$. In particular, there is a curve $t \mapsto \rho_t$ in $\mathcal{M}^+(\overline{\Omega})$, $0 \leq t \leq 1$, which is continuous in the narrow topology and such that

$$\forall \psi \in L^1([0,1]\times\overline{\Omega};\rho), \quad \int_{[0,1]\times\overline{\Omega}} \psi(t,x)\,\mathrm{d}\rho(t,x) = \int_0^1 \left(\int_{\overline{\Omega}} \psi(t,x)\,\mathrm{d}\rho_t(x) \right) \mathrm{d}t. \tag{3.2}$$

Thanks to this property, given times $t_0 = 0 < \ldots < t_T = 1$, it is justified to consider "slices" ρ_{t_j} and to assume that we are given data $b_j \in \mathbb{R}^m$ (for simplicity assume m does not change with j) at time t_j, corresponding to linear observations of ρ_{t_j}, given by linear operators $A_j\colon \mathcal{M}(\overline{\Omega}) \to \mathbb{R}^m$. To solve the corresponding inverse problem, [Bredies et al., 2021, (43)] proposes the minimisation of

$$\mathcal{E}(\rho) = \frac{1}{2}\sum_{j=0}^{T} \left\| A_j\rho_{t_j} - b_j \right\|_2^2 + \mathcal{W}(\rho). \tag{3.3}$$

We assume that each A_j is narrowly continuous (that is, in the topology induced by $\mathscr{C}_b(\overline{\Omega})$). The more involved case of continuous-time observations is handled in Bredies and Fanzon [2020].

Minimising such a model comes with very clear reconstruction properties.

Theorem 3.1 ([Bredies et al., 2021, Thms. 7, 10])**.** *Let $\alpha, \beta > 0$ and*

$$\Gamma \stackrel{\text{def.}}{=} \left\{ (h, \xi) \mid h \equiv 1,\ \xi \in \mathrm{AC}^2([0,1]; \overline{\Omega}) \right\}. \tag{3.4}$$

Then for all $\rho \in \mathcal{M}^+([0,1] \times \overline{\Omega})$ with $\mathcal{W}(\rho) < +\infty$, there exists $\sigma \in \mathcal{M}^+(\Gamma)$ such that $\rho = \Theta(\sigma)$, as defined in Theorem 2.2, and $v \in L^2_\rho([0,1] \times \overline{\Omega}; \mathbb{R}^d)$ such that

$$\mathcal{W}(\rho) = \int_{[0,1]\times\overline{\Omega}} \left[\alpha + \frac{\beta}{2}|v|^2 \right] \mathrm{d}\rho, \qquad \partial_t \rho + \mathrm{div}(v\rho) = 0, \tag{3.5}$$

and

$$\text{for a.e. } t \in [0,1] \text{ and } \sigma\text{-a.e. } \xi, \qquad \xi'(t) = v(t, \xi(t)). \tag{3.6}$$

Moreover, there exists a minimiser $\rho^* \in \operatorname*{argmin}_{\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})} \mathcal{E}(\rho)$ such that

$$\text{for some } a_i \geq 0,\ \xi^i \in \mathrm{AC}^2([0,1]; \overline{\Omega}), \qquad \forall t \in [0,1], \qquad \rho^*_t = \sum_{i=1}^{m(T+1)} a_i \delta_{\xi^i(t)}. \tag{3.7}$$

Remark 3.1. *The formal definition of AC^2 is given in [Ambrosio et al., 2008, Sec. 1.1], but here it can be thought of as H^1, the space of L^2 functions with weak derivative in L^2. See Lemma A.4 for more details.*

Proof of Theorem 3.1. Fix $\rho \in \mathcal{M}^+([0,1] \times \overline{\Omega})$ with $\mathcal{W}(\rho) < +\infty$. To apply [Thm. 7 Bredies et al., 2021], we need to show that there exists $v \in L^2_\rho([0,1] \times \overline{\Omega}; \mathbb{R}^d)$ which achieves the minimum of (3.1). To do this, note that the set

$$\left\{ v \in L^2_\rho([0,1] \times \overline{\Omega}; \mathbb{R}^d) \mid \int_{[0,1]\times\overline{\Omega}} \left[\alpha + \frac{\beta}{2}|v|^2 \right] \mathrm{d}\rho \leq \mathcal{W}(\rho) + 1,\ \partial_t \rho + \mathrm{div}(v\rho) = 0 \right\} \tag{3.8}$$

is bounded, therefore there is a weak limit-point $v \in L^2_\rho([0,1] \times \overline{\Omega}; \mathbb{R}^d)$ with $\mathcal{W}(\rho) = \int [\alpha + \frac{\beta}{2}|v|^2] \,\mathrm{d}\rho$. The weak form of the continuity equation is preserved under weak limits in v, therefore the triplet $(\rho, v, 0)$ also satisfies (1.9). Then, [Bredies et al., 2021, Thm. 7] (or Theorem 2.2(4)) confirms the existence of $\sigma \in \mathcal{M}^+(\Gamma)$ with the desired properties. The second statement, concerning the minimiser ρ^*, is exactly the result of [Bredies et al., 2021, Thm. 10]. □

These results highlight a very close relationship between the representations ρ and σ, i.e. dynamical measures and measures on paths. In the following subsections we first show that minimisation of (3.3) over $\rho \in \mathcal{M}^+([0,1] \times \overline{\Omega})$ can be expressed as an equivalent minimisation over $\sigma \in \mathcal{M}^+(\Gamma)$. This reformulation will then allow us to propose new and more efficient numerical schemes to compute reconstructions.

3.1 Reformulation in space of measures on paths

Here we propose a new energy $\mathrm{E}\colon \mathcal{M}^+(\Gamma) \to \mathbb{R}$ whose minimisers coincide with those of $\mathcal{E}\colon \mathcal{M}^+([0,1] \times \overline{\Omega}) \to \mathbb{R}$. The main problem is non-injectivity of Θ, so we do not necessarily expect to have $\mathrm{E}(\sigma) = \mathcal{E}(\Theta(\sigma))$ for all σ. On the other hand, the value of the data-fidelity is unaffected by the non-uniqueness. For any $\rho \in \mathscr{C}_w([0,1]; \mathcal{M}(\overline{\Omega}))$ and $\sigma \in \mathcal{M}^+(\Gamma)$ such that $\rho = \Theta(\sigma)$, we have

$$\forall t \in [0,1], \qquad \rho_t = \Theta(\sigma)_t = (e_t)_\sharp \sigma, \tag{3.9}$$

(see (1.8) and Theorem 2.2). Motivated by this equivalence of "time slice" measures, we propose the energy

$$\forall \sigma \in \mathcal{M}^+(\Gamma), \qquad \mathrm{E}(\sigma) \stackrel{\text{def.}}{=} \frac{1}{2} \sum_{j=0}^{T} \left\| A_j (e_{t_j})_\sharp \sigma - b_j \right\|_2^2 + \mathrm{W}(\sigma), \tag{3.10}$$

$$\mathrm{W}(\sigma) \stackrel{\text{def.}}{=} \int_\Gamma \int_0^1 \left[\alpha + \frac{\beta}{2}|\xi'|^2 \right] \mathrm{d}t \,\mathrm{d}\sigma(1, \xi). \tag{3.11}$$

Since the mass h is constant in this example, we will write ξ or $(1,\xi) \in \Gamma$ interchangeably for the remainder of this section. Key results from Bredies et al. [2021] show that $\mathrm{W}\colon \mathcal{M}^+(\Gamma) \to \mathbb{R}$ is equivalent to $\mathcal{W}\colon \mathcal{M}^+([0,1]\times\overline{\Omega}) \to \mathbb{R}$ in the following way.

Lemma 3.2. *The energy* E *is convex and lower semi-continuous. Also, for any* $\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})$,

$$\mathcal{W}(\rho) = \min_{\sigma\in\mathcal{M}^+(\Gamma)} \left\{ \mathrm{W}(\sigma) \mid \Theta(\sigma) = \rho \right\}. \tag{3.12}$$

We conclude that

$$\min\left\{ \mathrm{E}(\sigma) \mid \sigma \in \mathcal{M}^+(\Gamma) \right\} = \min\left\{ \mathcal{E}(\rho) \mid \rho \in \mathcal{M}^+([0,1]\times\overline{\Omega}) \right\}, \tag{3.13}$$

and there is a minimising $\sigma^* \in \operatorname{argmin}_{\sigma\in\mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma)$ *of the form*

$$\text{for some } a_i \geq 0,\ (1,\xi^i) \in \Gamma, \qquad \sigma^* = \sum_{i=1}^{m(T+1)} a_i \delta_{(1,\xi^i)}. \tag{3.14}$$

Proof. As $\Gamma \subset \Gamma_\infty$, Lemma 2.3 confirms that $A_j(e_{t_j})_\sharp\colon \mathcal{M}^+(\Gamma) \to \mathbb{R}^m$ is linear and narrowly continuous for each $j = 0,\dots,T$, so E is still convex and lower semi-continuous.

Secondly, fix $\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})$. If $\mathcal{W}(\rho) = +\infty$, then $\mathcal{W}(\rho) \geq \min\{ \mathrm{W}(\sigma) \mid \Theta(\sigma) = \rho \}$ is clear. Otherwise, let $\widehat{\sigma} \in \mathcal{M}^+(\Gamma)$ be the measure given by Theorem 3.1 satisfying $\rho = \Theta(\widehat{\sigma})$, then

$$\begin{aligned}
\mathcal{W}(\rho) &= \int_{[0,1]\times\overline{\Omega}} \left[\alpha + \frac{\beta}{2}|v|^2\right] \mathrm{d}\rho && \text{Theorem 3.1} && (3.15)\\
&= \int_\Gamma \int_0^1 \left[\alpha + \frac{\beta}{2}|v(t,\xi(t))|^2\right] \mathrm{d}t\,\mathrm{d}\widehat{\sigma}(\xi) && \text{Theorem 2.2(2)} && (3.16)\\
&= \int_\Gamma \int_0^1 \left[\alpha + \frac{\beta}{2}|\xi'(t)|^2\right] \mathrm{d}t\,\mathrm{d}\widehat{\sigma}(\xi) = \mathrm{W}(\widehat{\sigma}). && (3.6) && (3.17)
\end{aligned}$$

This confirms $\mathcal{W}(\rho) \geq \min\{ \mathrm{W}(\sigma) \mid \Theta(\sigma) = \rho \}$ for all $\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})$.

The converse holds by Jensen's inequality. In particular, because $\mathcal{W}$ is convex, proper, and lower semi-continuous ([Bredies et al., 2021, Lemma 2]), by [Thm. 5 Rockafellar, 1974] there exists a collection $\{(c_i,\psi_i)\}_{i\in I} \subset \mathbb{R}\times\mathscr{C}([0,1]\times\overline{\Omega})$ such that

$$\forall \rho \in \mathcal{M}^+([0,1]\times\overline{\Omega}), \qquad \mathcal{W}(\rho) = \sup_{i\in I} \left(c_i + \int_{[0,1]\times\overline{\Omega}} \psi_i \,\mathrm{d}\rho \right). \tag{3.18}$$

Because $\mathcal{W}$ is positively homogeneous, this simplifies with $c_i = 0$. For any $\widehat{\sigma} \in \mathcal{M}^+(\Gamma)$, $i \in I$, we have $\Theta(\widehat{\sigma}) \in \mathcal{M}^+([0,1]\times\overline{\Omega})$ and

$$\begin{aligned}
\int_{[0,1]\times\overline{\Omega}} \psi_i \,\mathrm{d}\Theta(\widehat{\sigma}) = \int_\Gamma \int_0^1 \psi_i(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\widehat{\sigma} &\leq \int_\Gamma \sup_{j\in I} \int_0^1 \psi_j(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\widehat{\sigma}\\
&= \int_\Gamma \left[\sup_{j\in I} \int_{[0,1]\times\overline{\Omega}} \psi_j \,\mathrm{d}\Theta(\delta_\xi) \right] \mathrm{d}\widehat{\sigma} = \int_\Gamma \mathcal{W}(\Theta(\delta_\xi))\,\mathrm{d}\widehat{\sigma}. \qquad (3.19)
\end{aligned}$$

It is shown in [Bredies et al., 2021, (19)] that $\mathcal{W}(\Theta(\delta_\xi)) = \int_0^1 \alpha + \frac{\beta}{2}|\xi'(t)|^2\,\mathrm{d}t$. Now taking a supremum over $i \in I$ gives

$$\mathcal{W}(\Theta(\widehat{\sigma})) \leq \int_\Gamma \mathcal{W}(\Theta(\delta_\xi))\,\mathrm{d}\widehat{\sigma} = \int_\Gamma \int_0^1 \alpha + \frac{\beta}{2}|\xi'(t)|^2\,\mathrm{d}t\,\mathrm{d}\widehat{\sigma} = \mathrm{W}(\widehat{\sigma}). \tag{3.20}$$

This shows $\mathcal{W}(\rho) \leq \min\{ \mathrm{W}(\sigma) \mid \Theta(\sigma) = \rho \}$ for all $\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})$, which becomes equality when combined with the first result.

Finally we consider the equivalence of minimums and minimisers. With a_i and ξ^i given by ρ^* in (3.7), define σ^* as in (3.14). One can check that $\Theta(\sigma^*) = \rho^*$. Observe each measure has equal energy:

$$\mathcal{W}(\rho^*) = \sum_{i=1}^{m(T+1)} a_i \left[\int_0^1 \alpha + \frac{\beta}{2} |(\xi^i)'(t)|^2 \,\mathrm{d}t \right] \qquad \text{[Bredies et al., 2021, Thm. 10]} \tag{3.21}$$

$$= \mathrm{W}\left(\sum_{i=1}^{m(T+1)} a_i \delta_{\xi^i} \right) = \mathrm{W}(\sigma^*), \qquad \text{(linearity of W)} \tag{3.22}$$

therefore

$$\min_{\sigma \in \mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma) \leq \frac{1}{2} \sum_{j=0}^{T} \left\| A_j (e_{t_j})_\sharp \sigma^* - b_j \right\|_2^2 + \mathrm{W}(\sigma^*) = \mathcal{E}(\rho^*) = \min_{\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})} \mathcal{E}(\rho). \tag{3.23}$$

Conversely, for any $\sigma \in \mathcal{M}^+(\Gamma)$ we have

$$\min_{\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})} \mathcal{E}(\rho) \leq \mathcal{E}(\Theta(\sigma)) \leq \frac{1}{2} \sum_{j=0}^{T} \left\| A_j (e_{t_j})_\sharp \sigma - b_j \right\|_2^2 + \mathrm{W}(\sigma) = \mathrm{E}(\sigma). \tag{3.24}$$

We conclude that $\min_{\sigma \in \mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma) = \mathrm{E}(\sigma^*) = \mathcal{E}(\rho^*) = \min_{\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})} \mathcal{E}(\rho)$ as required. □

This theorem justifies that Θ can be thought of as a reparametrisation; the two optimisation problems are equivalent. Although different measures σ might represent the same mass ρ (see for instance [Bredies et al., 2020b, Rem. 4.2]), the point of view of measures on paths σ is still interesting as it provides plausible trajectories for physical particles.

3.2 Minimisation algorithm

A major advantage of Frank-Wolfe / Generalised Conditional Gradient methods is their ability to minimise energies like (3.3) while constructing iterates which share the same sparse structure (3.7) as the minimisers (see for instance Bredies and Pikkarainen [2013], Boyd et al. [2017], Denoyelle et al. [2019] in the case of point sources recovery). The method proposed in Bredies et al. [2020b] follows that approach, and the core of the algorithm is the minimisation of a linear functional on the level sets of $\mathcal{W}$ (see [Bredies et al., 2020b, Prop. 3.6]). The description of the extreme points of those level sets in Bredies et al. [2021] makes it possible to reformulate that problem as a variational problem on curves.

In the present section, we point out that, while the algorithm derived in Bredies et al. [2020b] is quite natural, there are many possible ways to define a Frank-Wolfe–based algorithm minimising (3.3).

Relying on the linear map Θ, we note that the minimisation of (3.10) is equivalent to

$$\min_{\sigma \in \mathcal{M}^+(\Gamma)} \frac{1}{2} \sum_{j=0}^{T} \left\| A_j (e_{t_j})_\sharp \sigma - b_j \right\|_2^2 + \int_\Gamma w \,\mathrm{d}\sigma \tag{3.25}$$

$$\text{where} \quad \forall \xi \in \Gamma, \quad w(\xi) \overset{\text{def.}}{=} \alpha + \frac{\beta}{2} \int_0^1 |\xi'(t)|^2 \,\mathrm{d}t. \tag{3.26}$$

The standard Frank-Wolfe procedure consists in iteratively minimising the linearisation of (3.25) on a compact set (for some suitable topology) to yield descent directions. At iteration n, let $\eta_j \overset{\text{def.}}{=} A_j^* \left(A_j (e_{t_j})_\sharp \sigma^n - b_j \right)$ denote the linearisation of each summand of the fidelity term. We are led to minimise

$$\min_{\sigma \in C} \int_\Gamma \left[\sum_{j=0}^{T} \eta_j(\xi(t_j)) + w(\xi) \right] \mathrm{d}\sigma, \tag{3.27}$$

where $C \subset \mathcal{M}^+(\Gamma)$ is a suitably chosen compact convex set which contains the minimisers of E on $\mathcal{M}^+(\Gamma)$. A first choice is to consider the set

$$C_1 \stackrel{\text{def.}}{=} \left\{ \sigma \in \mathcal{M}^+(\Gamma) \mid \int_\Gamma w \,\mathrm{d}\sigma \leq \mathrm{E}(0) \right\}, \tag{3.28}$$

which does contain the minimisers of (3.25) since they must satisfy $\int_\Gamma w \,\mathrm{d}\sigma \leq \mathrm{E}(\sigma) \leq \mathrm{E}(0)$. By Lemma A.5, the set of extreme points of C_1 is $\{0\} \cup \left\{ \frac{\mathrm{E}(0)}{w(\xi)} \delta_\xi \mid \xi \in \Gamma \right\}$, which yields the descent direction

$$\xi^{n+1} \in \operatorname*{argmin}_{\xi \in \Gamma} \frac{\sum_{j=0}^T \eta_j(\xi_j) + w(\xi)}{w(\xi)} = \operatorname*{argmin}_{\xi \in \Gamma} \frac{\sum_{j=0}^T \eta_j(\xi_j)}{w(\xi)}. \tag{3.29}$$

That is precisely the descent direction used in [Bredies et al., 2020b, (4.30)], although in that context (3.27) is written as a minimisation over $\rho \in \mathcal{M}^+([0,1] \times \overline{\Omega})$.

However, alternative choices for C are possible. We propose to use

$$C_2 \stackrel{\text{def.}}{=} \left\{ \sigma \in \mathcal{M}^+(\Gamma) \mid \int_\Gamma \alpha \,\mathrm{d}\sigma \leq \mathrm{E}(0) \right\}, \tag{3.30}$$

which contains the minimisers of (3.25) too, since they must satisfy $\int_\Gamma \alpha \,\mathrm{d}\sigma \leq \int_\Gamma w \,\mathrm{d}\sigma \leq \mathrm{E}(\sigma) \leq \mathrm{E}(0)$. Applying Lemma A.5, we see that the set of extreme points of C_2 is $\{0\} \cup \left\{ \frac{\mathrm{E}(0)}{\alpha} \delta_\xi \mid \xi \in \Gamma \right\}$, which yields the linear minimisation step

$$\xi^{n+1} \in \operatorname*{argmin}_{\xi \in \Gamma} \frac{\sum_{j=0}^T \eta_j(\xi_j) + w(\xi)}{1} = \operatorname*{argmin}_{\xi \in \Gamma} \sum_{j=0}^T \eta_j(\xi_j) + w(\xi). \tag{3.31}$$

We found (3.31) more convenient than (3.29) as it is amenable to dynamic programming techniques (see Section 3.4). The whole minimisation algorithm is summarised in Algorithm 1.

Remark 3.2. *In measure spaces, compactness is equivalent to boundedness and tightness. In both cases, boundedness comes from* $\inf_{\gamma \in \Gamma} \varphi > 0$*, and tightness comes from the fact that* w *has compact sub-levelsets. The set* C_1 *is compact by Lemma A.3. On the other hand,* C_2 *is only bounded, but sub-levelsets of* (3.27) *are still compact, as argued in Lemma 5.4.*

Algorithm 1 Frank-Wolfe algorithm for the Benamou-Brenier example

1: Choose $\sigma^0 = 0$, $n \leftarrow 0$
2: **repeat**
3: $\quad \eta_j \leftarrow A_j^*(A_j(e_{t_j})_\sharp \sigma^n - b_j) \in \mathscr{C}(\overline{\Omega})$ ▷ Gradient of data term
4: $\quad$ Compute ξ^{n+1} according to (3.29) or (3.31) ▷ Linear oracle step
5: $\quad$ Choose $\mu^{n+1} \propto \delta_{\xi^{n+1}}$ such that $\mu^{n+1} \in \operatorname{Ext}(C)$ ▷ Choice of scaling
6: $\quad \sigma^{n+1} \leftarrow (1 - \lambda_n)\sigma^n + \lambda_n \mu^{n+1}$ ▷ Some choice of $\lambda_n \in [0,1]$
7: $\quad n \leftarrow n + 1$
8: **until** converged

3.3 Discrete-time formulation

In the previous two sections (Lemma 3.2 and Section 3.2) we have shown that minimisers σ^* of (3.10) can be approximated whilst only considering spatially discrete measures. Because of the discrete-time data, we can further simplify the numerical computations by considering

discrete-time measures. Observe

$$\mathrm{E}(\sigma^*) = \min_{\sigma\in\mathcal{M}^+(\Gamma)} \left\{ \frac{1}{2}\sum_{j=0}^{T} \left\| A_j(e_{t_j})_\sharp\sigma - b_j \right\|_2^2 + \mathrm{W}(\sigma) \right\} \tag{3.32}$$

$$= \min_{\substack{\rho_j\in\mathcal{M}^+(\overline{\Omega})\\ j=0,\dots,T}} \min_{\substack{\sigma\in\mathcal{M}^+(\Gamma)\\ (e_{t_j})_\sharp\sigma=\rho_j}} \left\{ \frac{1}{2}\sum_{j=0}^{T} \|A_j\rho_j - b_j\|_2^2 + \mathrm{W}(\sigma) \right\} \tag{3.33}$$

$$= \min_{\substack{\rho_j\in\mathcal{M}^+(\overline{\Omega})\\ j=0,\dots,T}} \left\{ \frac{1}{2}\sum_{j=0}^{T} \|A_j\rho_j - b_j\|_2^2 + \min_{\substack{\sigma\in\mathcal{M}^+(\Gamma)\\ (e_{t_j})_\sharp\sigma=\rho_j}} \mathrm{W}(\sigma) \right\}. \tag{3.34}$$

The minimal ρ_j^* will inherit the same sparse structure from ρ^*. This highlights that the minimal energy can be computed exactly whilst only considering measures which are discrete in both time and space, then the continuous-time minimiser is recovered as a W-minimising interpolation. In the σ measures-on-paths representation, this continuous-time interpolation can be written path-wise. We simply replace any discrete path $(\xi^i(t_0),\dots,\xi^i(t_T))$ with the Benamou-Brenier geodesic curve interpolating those points.

Lemma 3.3. *Assume that Ω is convex and $\sigma^* \in \operatorname{argmin}_{\sigma\in\mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma)$, then*

$$\text{for } \sigma^*\text{-a.e. } \xi, \qquad \xi \in \operatorname*{argmin}_{\widetilde{\xi}\in\Gamma} \left\{ \int_0^1 \left[\alpha + \frac{\beta}{2}|\widetilde{\xi}'|^2 \right] \mathrm{d}t \quad \text{s.t.} \quad \widetilde{\xi}(t_j) = \xi(t_j) \text{ for all } j \right\}. \tag{3.35}$$

This is an example of Lemma B.1. In this specific case, as commented in [Bredies et al., 2020b, Rem. 4.10], paths are piecewise linear. If we restrict Γ to this set of geodesics

$$\Gamma \overset{\text{def.}}{=} \left\{ (h,\xi) \in \{1\} \times \mathrm{AC}^2([0,1];\overline{\Omega}) \mid \xi' \text{ is constant on }]t_{j-1},t_j[,\ 1\le j\le T \right\}, \tag{3.36}$$

then we can write W in its explicit discrete-time form:

$$\mathrm{W}(\sigma) = \int_\Gamma \left[\alpha + \frac{\beta}{2}\sum_{j=1}^{T} \frac{|\xi(t_j)-\xi(t_{j-1})|^2}{t_j - t_{j-1}} \right] \mathrm{d}\sigma(\xi). \tag{3.37}$$

3.4 Linear oracle computation

The most challenging step of Algorithm 1 for our choice of C is the minimisation (3.31), the linear oracle computation. Fix $\sigma^n \in \mathcal{M}^+(\Gamma)$. Utilising the discrete-time form of W, (3.31) becomes

$$\xi^{n+1} \in \operatorname*{argmin}_{\xi\in\overline{\Omega}^{T+1}} \widetilde{\mathrm{E}}_T(\xi), \tag{3.38}$$

$$\text{where } \forall J = 0,\dots,T,\ \xi\in\overline{\Omega}^{T+1}, \qquad \widetilde{\mathrm{E}}_J(\xi) \overset{\text{def.}}{=} \sum_{j=0}^{J} \eta_j(\xi_j) + \alpha + \frac{\beta}{2}\sum_{j=1}^{J} \frac{|\xi_j-\xi_{j-1}|^2}{t_j-t_{j-1}} \tag{3.39}$$

$$\text{and } \forall j = 0,\dots,T, \qquad \eta_j \overset{\text{def.}}{=} A_j^*(A_j(e_{t_j})_\sharp\sigma^n - b_j). \tag{3.40}$$

Unless η_j has some special structure, computing ξ^{n+1} is a challenging non-convex optimisation. As Ω is a non-discrete set, it cannot be computed exactly in finite time. Even if Ω is discretised, the search space is still of size $|\Omega|^{T+1}$. In Bredies et al. [2020b] they use a stochastic search of $\overline{\Omega}^{T+1}$ to approximate ξ^{n+1}. We propose taking a discrete subset, and then use the discrete minimiser for ξ^{n+1}.

Assume that we are given a family of finite sets $\Lambda_j \subseteq \overline{\Omega}$ ("finite grids" which discretise $\overline{\Omega}$), for $j = 0,\dots,T$. For each $J = 0,\dots,T$ and $y\in\Lambda_J$ define the optimal curves

$$Y^*[y,J] \in \operatorname*{argmin}_{Y\in\overline{\Omega}^{J+1}} \left\{ \widetilde{\mathrm{E}}_J(Y) \mid Y_J = y,\ Y_j\in\Lambda_j \text{ for } j=0,\dots,J-1 \right\}. \tag{3.41}$$

Observe that for all $J = 1, \dots, T$ and $y \in \Lambda_J$,

$$\min_{\substack{Y \in \prod_{j=0}^{J} \Lambda_j \\ Y_J = y}} \widetilde{\mathrm{E}}_J(Y) = \min_{Y \in \prod_{j=0}^{J} \Lambda_j} \left\{ \widetilde{\mathrm{E}}_{J-1}(Y) + \eta_J(y) + \alpha + \frac{\beta}{2} \frac{|y - Y_{J-1}|^2}{t_J - t_{J-1}} \right\} \tag{3.42}$$

$$= \eta_J(y) + \alpha + \min_{y' \in \Lambda_{J-1}} \left\{ \frac{\beta}{2} \frac{|y - y'|^2}{t_J - t_{J-1}} + \min_{\substack{Y \in \prod_{j=0}^{J-1} \Lambda_j \\ Y_{J-1} = y'}} \widetilde{\mathrm{E}}_{J-1}(Y) \right\}. \tag{3.43}$$

In particular, we can choose $Y^*[y, J]$ inductively to be

$$Y^*[y, J] = \begin{pmatrix} Y^*[y', J-1] & y \end{pmatrix} \quad \text{for} \quad y' \in \operatorname*{argmin}_{y' \in \Lambda_{J-1}} \left[\frac{\beta}{2} \frac{|y - y'|^2}{t_J - t_{J-1}} + \widetilde{\mathrm{E}}_{J-1}(Y^*[y', J-1]) \right]. \tag{3.44}$$

The key complexity result is that we can compute this minimiser in linear time in T, rather than the naive exponential scaling.

4 The unbalanced Wasserstein-Fisher-Rao example

In the last section we showed that optimisation of the Benamou-Brenier energy can be reformulated as an optimisation over the space of measures on paths. In this section we will justify that this is also true for the Wasserstein-Fisher-Rao (WFR) example considered in Bredies et al. [2020a], where the transportation cost allows for changes of mass. The two arguments are very similar and so we will only emphasise the differences and reference the different published results.

The regularisation penalty considered in [Bredies et al., 2020a, (33) and (152)] is $\mathcal{W}\colon \mathcal{M}^+([0,1] \times \overline{\Omega}) \to \mathbb{R} \cup \{+\infty\}$ such that for all $\rho \in \mathcal{M}^+([0,1] \times \overline{\Omega})$,

$$\mathcal{W}(\rho) \stackrel{\text{def.}}{=} \inf_{\substack{v \in L^2_\rho([0,1] \times \overline{\Omega}; \mathbb{R}^d) \\ g \in L^2_\rho([0,1] \times \overline{\Omega})}} \left\{ \int_{[0,1] \times \overline{\Omega}} \left[\alpha + \frac{\beta}{2} |v|^2 + \frac{\beta \delta^2}{2} g^2 \right] \mathrm{d}\rho \quad \text{s.t.} \quad \partial_t \rho + \operatorname{div}(v\rho) = g\rho \right\} \tag{4.1}$$

for some $\alpha, \beta, \delta > 0$, and the continuity equation is satisfied in the sense of (1.9). As in (3.3), we choose the energy

$$\mathcal{E}(\rho) = \frac{1}{2} \sum_{j=0}^{T} \left\| A_j \rho_{t_j} - b_j \right\|_2^2 + \mathcal{W}(\rho). \tag{4.2}$$

Much is already known about the structure of minimisers of this energy.

Theorem 4.1 ([Bredies et al., 2020a, Thms. 4.2,6.4])**.** *Let $\alpha, \beta, \delta > 0$ and*

$$\Gamma \stackrel{\text{def.}}{=} \left\{ (h, \xi) \colon [0,1] \to [0,1] \times \overline{\Omega} \mid \sqrt{h} \in \mathrm{AC}^2([0,1]), \ \sqrt{h}\xi \in \mathrm{AC}^2([0,1]; \mathbb{R}^d) \right\}. \tag{4.3}$$

Then for all $\rho \in \mathcal{M}^+([0,1] \times \overline{\Omega})$ with $\mathcal{W}(\rho) < +\infty$, there exists $\sigma \in \mathcal{M}^+(\Gamma)$, $v \in L^2_\rho([0,1] \times \overline{\Omega}; \mathbb{R}^d)$, $g \in L^2_\rho([0,1] \times \overline{\Omega})$ such that $\rho = \Theta(\sigma)$,

$$\mathcal{W}(\rho) = \int_{[0,1] \times \overline{\Omega}} \left[\alpha + \frac{\beta}{2} |v|^2 + \frac{\beta \delta^2}{2} g^2 \right] \mathrm{d}\rho, \qquad \partial_t \rho + \operatorname{div}(v\rho) = g\rho, \tag{4.4}$$

and

$$\begin{aligned} \xi'(t) &= v(t, \xi(t)) \quad \textit{for} \quad \textit{a.e. } t \in \{h > 0\} \textit{ and } \sigma\textit{-a.e. } (h, \xi), \\ h'(t) &= g(t, \xi(t)) h(t) \quad \textit{for} \quad \textit{a.e. } t \in [0,1] \textit{ and } \sigma\textit{-a.e. } (h, \xi). \end{aligned} \tag{4.5}$$

Moreover, there exists a minimiser $\rho^* \in \operatorname*{argmin}_{\rho \in \mathcal{M}^+([0,1] \times \overline{\Omega})} \mathcal{E}(\rho)$ such that

$$\textit{for some } a_i \geq 0, \ (h^i, \xi^i) \in \Gamma, \qquad \forall t \in [0,1], \qquad \rho_t^* = \sum_{i=1}^{m(T+1)} a_i h^i(t) \delta_{\xi^i(t)}. \tag{4.6}$$

Sketch proof. If $\mathcal{W}(\rho) < +\infty$, the existence of optimal v and g for (4.1) is again given by weak compactness in $L^2_\rho([0,1]\times\overline{\Omega})$. The result of [Bredies et al., 2020a, Thm. 4.2] (or Theorem 2.2) shows that $\rho = \Theta(\sigma)$ for some $\sigma \in \mathcal{M}^+(\Gamma)$ satisfying (4.5). Finally, the structure of ρ^* is given by [Bredies et al., 2020a, Thm. 6.4]. □

4.1 Reformulation in space of measures on paths

The function $\mathcal{E}$ can be reformulated into an equivalent energy $\mathrm{E}\colon \mathcal{M}^+(\Gamma) \to \mathbb{R}\cup\{+\infty\}$ defined by

$$\forall \sigma \in \mathcal{M}^+(\Gamma), \qquad \mathrm{E}(\sigma) \overset{\text{def.}}{=} \frac{1}{2}\sum_{j=0}^{T} \left\| A_j(e_{t_j})_\sharp \sigma - b_j \right\|_2^2 + \mathrm{W}(\sigma), \tag{4.7}$$

$$\mathrm{W}(\sigma) \overset{\text{def.}}{=} \int_\Gamma \int_0^1 \left[\alpha + \frac{\beta}{2}|\xi'|^2 + \frac{\beta\delta^2}{2}\left(\frac{h'}{h}\right)^2 \right] h \,\mathrm{d}t \,\mathrm{d}\sigma(h,\xi). \tag{4.8}$$

Remark 4.1. *Since for all $(h,\xi)\in\Gamma$, $\sqrt{h}$ and $\sqrt{h}\xi$ are absolutely continuous, they are differentiable almost everywhere in $[0,1]$, hence ξ is differentiable at a.e. t such that $h(t) > 0$. Hence, the integrand in (4.8) makes sense when regarded as*

$$\int_{\{h>0\}} \left[\alpha + \frac{\beta}{2}\left| \left(\frac{\sqrt{h}\xi}{\sqrt{h}}\right)'(t) \right|^2 + 2\beta\delta^2 \left| (\sqrt{h})'(t) \right|^2 \right] h \,\mathrm{d}t. \tag{4.9}$$

We can now replicate the extension performed in Lemma 3.2

Lemma 4.2. *The energy* E *is convex and lower semi-continuous. Also, for any $\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})$,*

$$\mathcal{W}(\rho) = \min_{\sigma\in\mathcal{M}^+(\Gamma)} \left\{ \mathrm{W}(\sigma) \mid \Theta(\sigma) = \rho \right\}. \tag{4.10}$$

We conclude that

$$\min\left\{ \mathrm{E}(\sigma) \mid \sigma \in \mathcal{M}^+(\Gamma) \right\} = \min\left\{ \mathcal{E}(\rho) \mid \rho \in \mathcal{M}^+([0,1]\times\overline{\Omega}) \right\}, \tag{4.11}$$

and there is a minimising $\sigma^* \in \operatorname{argmin}_{\sigma\in\mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma)$ of the form

$$\text{for some } a_i \geq 0,\ (h^i,\xi^i)\in\Gamma, \qquad \sigma^* = \sum_{i=1}^{m(T+1)} a_i \delta_{(h^i,\xi^i)}. \tag{4.12}$$

Sketch proof. As before, Lemma 2.3 confirms that E is convex and lower semi-continuous and the existence results of Theorem 4.1 show that $\mathcal{W}(\rho) \geq \min_{\sigma\in\mathcal{M}^+(\Gamma)} \{ \mathrm{W}(\sigma) \mid \Theta(\sigma) = \rho \}$ for all $\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})$.

Conversely, [Bredies et al., 2020a, Lemma A.6] shows that $\mathcal{W}$ is convex, proper, lower semi-continuous, and positively homogeneous. Therefore the argument of Lemma 3.2 by Jensen's inequality holds, giving us $\mathcal{W}(\rho) \leq \int_\Gamma \mathcal{W}(\delta_\gamma)\,\mathrm{d}\sigma$ for any $\sigma \in \mathcal{M}^+(\Gamma)$ such that $\Theta(\sigma) = \rho$. It is the result of [Bredies et al., 2020a, Prop. 3.9] that confirms $\int_\Gamma \mathcal{W}(\delta_\gamma)\,\mathrm{d}\sigma = \mathrm{W}(\sigma)$.

Finally we consider the equivalence of minimums and minimisers. Define ρ^* as in (4.6) and σ^* as in (4.12), again we can confirm $\Theta(\sigma^*) = \rho^*$. The required equality $\mathcal{W}(\rho^*) = \sum_{i=1}^{m(T+1)} a_i \,\mathrm{W}(\delta_{\gamma^i})$ is given by [Bredies et al., 2020a, Thm. 6.4]. The rest of the proof follows as before. □

4.2 Discrete-time formulation

Problems of the form (4.7) also have a discrete-time structure inherited from discrete-time data. The same argument from Section 3.3 (namely the use of (3.34) and Lemma B.1) is valid for unbalanced examples. In the context of the WFR penalty, this is stated as follows.

Lemma 4.3. *Assume that Ω is convex and $\sigma^* \in \operatorname{argmin}_{\sigma\in\mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma)$, define*

$$d_{\alpha,\beta,\delta}(\gamma_0,t_0,\gamma_1,t_1) = \inf_{\gamma\in\Gamma} \left\{ \int_{t_0}^{t_1} \left[\alpha + \frac{\beta}{2}|\xi'(t)|^2 + \frac{\beta\delta^2}{2}\left(\frac{h'(t)}{h(t)}\right)^2\right] h(t)\,\mathrm{d}t \right.$$
$$\left. \text{such that } \gamma=(h,\xi),\ \gamma(t_0)=\gamma_0,\ \gamma(t_1)=\gamma_1 \right\}. \quad (4.13)$$

Then σ^*-a.e. $\gamma=(h,\xi)$ achieves minimal energy. In particular

$$\int_0^1 \left[\alpha + \frac{\beta}{2}|\xi'|^2 + \frac{\beta\delta^2}{2}\left(\frac{h'}{h}\right)^2\right] h\,\mathrm{d}t = \sum_{j=1}^{T} d_{\alpha,\beta,\delta}(\gamma(t_{j-1}),t_{j-1},\gamma(t_j),t_j). \quad (4.14)$$

The analytical form of these WFR-geodesics is not so convenient as for the Benamou-Brenier penalty, although for $\alpha=0$ it is known ([Chizat et al., 2018a, Thm. 5.6])

$$d_{0,\beta,\delta}((h_0,\xi_0),t_0,(h_1,\xi_1),t_1) = \frac{4\beta\delta^2}{t_1-t_0}\left[\tfrac{h_1+h_0}{2} - \sqrt{h_1h_0}\cos\left(\min\left(\tfrac{|\xi_1-\xi_0|}{2\delta},\pi\right)\right)\right]. \quad (4.15)$$

4.3 Linear oracle computation

Similarly to Section 3.4, the linear oracle computation is to find

$$\gamma^{n+1} \in \operatorname*{argmin}_{\gamma\in([0,1]\times\overline{\Omega})^{T+1}} \widetilde{\mathrm{E}}_T(\gamma), \quad (4.16)$$

where, for all $\gamma=(h,\xi)$, $j,J=0,\dots,T$ we define

$$\widetilde{\mathrm{E}}_J(\gamma) \overset{\text{def.}}{=} \sum_{j=0}^{J} h_j\eta_j(\xi_j) + \sum_{j=1}^{J} d_{\alpha,\beta,\delta}(\gamma_{j-1},t_{j-1},\gamma_j,t_j) \quad \text{and} \quad \eta_j \overset{\text{def.}}{=} A_j^*(A_j(e_{t_j})_\sharp\sigma^n - b_j). \quad (4.17)$$

This formulation is much less explicit than in Section 3.4, but in Section 5.3 we will see that this is sufficient to be covered by our general framework. In particular, the numerical method explained in Section 7 is also applicable to the WFR example.

5 Proposed framework and analytical results

In this section we will introduce the framework which we use to generalise the examples of Sections 3 and 4, and state the main results which will be proven in later sections.

5.1 Variational formulation

In this work we consider dynamical inverse problems with a discrete-time structure. In particular, there exist observation times $t_j\in[0,1]$, $j=0,\dots,T$ and narrowly continuous linear operators $A_j\colon\mathcal{M}(\overline{\Omega})\to\mathbb{R}^m$. The operators A_j are described by $a_i^j\in\mathscr{C}(\overline{\Omega})$ such that

$$\forall\rho\in\mathcal{M}(\overline{\Omega}),\ i=1,\dots,m,\ j=0,\dots,T, \qquad (A_j\rho)_i \overset{\text{def.}}{=} \int_{\overline{\Omega}} a_i^j(x)\,\mathrm{d}\rho(x). \quad (5.1)$$

As stated at the end of Section 2, we work with a closed set of curves

$$\Gamma\subset\Gamma_\infty \overset{\text{def.}}{=} \left\{\,\gamma=(h,\xi)\mid h\in\mathscr{C}([0,1];[-1,1]),\ \xi\colon[0,1]\to\overline{\Omega},\ \xi|_{\{h\neq0\}} \text{ is continuous}\,\right\}, \quad (5.2)$$

so that $\forall t\in[0,1]$, the map $(e_t)_\sharp\colon\mathcal{M}(\Gamma)\to\mathcal{M}(\overline{\Omega})$ is narrowly continuous where

$$\forall\sigma\in\mathcal{M}(\Gamma),\ \psi\in\mathscr{C}(\overline{\Omega}), \qquad \int_{\overline{\Omega}}\psi(x)\,\mathrm{d}[(e_t)_\sharp\sigma](x) \overset{\text{def.}}{=} \int_\Gamma h(t)\psi(\xi(t))\,\mathrm{d}\sigma(h,\xi). \quad (5.3)$$

We can therefore define $\mathrm{F}\colon \mathcal{M}(\Gamma) \to \mathbb{R}$ by

$$\text{for some convex } \mathrm{F}_j \in C^1(\mathbb{R}^m;\mathbb{R}), \qquad \mathrm{F}(\sigma) \stackrel{\text{def.}}{=} \sum_{j=0}^{T} \mathrm{F}_j(A_j\left[(e_{t_j})_\sharp \sigma\right]). \tag{5.4}$$

For lower semi-continuous $w, \varphi\colon \Gamma \to [0,+\infty]$ define $\mathrm{W}\colon \mathcal{M}^+(\Gamma) \to]-\infty,+\infty]$, $D \subset \mathcal{M}^+(\Gamma)$ by

$$\forall \sigma \in D, \qquad \mathrm{W}(\sigma) \stackrel{\text{def.}}{=} \int_\Gamma w(\gamma)\,\mathrm{d}\sigma(\gamma) \tag{5.5}$$

$$\text{where} \qquad D \stackrel{\text{def.}}{=} \left\{ \sigma \in \mathcal{M}(\Gamma) \quad | \quad \sigma \geq 0, \quad \int_\Gamma \varphi(\gamma)\,\mathrm{d}\sigma \leq 1 \right\}. \tag{5.6}$$

Our general energy is of the form $\mathrm{E}\colon D \to]-\infty,+\infty]$ defined by

$$\forall \sigma \in D, \qquad \mathrm{E}(\sigma) \stackrel{\text{def.}}{=} \mathrm{F}(\sigma) + \mathrm{W}(\sigma). \tag{5.7}$$

Similarly to the examples in Sections 3 and 4, we could use the map $\Theta\colon \mathcal{M}(\Gamma) \to \mathscr{C}_w([0,1]\,;\mathcal{M}(\overline{\Omega}))$ to represent an energy $\mathcal{E}\colon \mathscr{C}_w([0,1]\,;\mathcal{M}(\overline{\Omega})) \to]-\infty,+\infty]$ via the formula

$$\mathcal{E}(\rho) = \sum_{j=0}^{T} \mathrm{F}_j(A_j\rho_j) + \inf\left\{\, \mathrm{W}(\sigma) \mid \sigma \in D,\ \Theta(\sigma) = \rho \,\right\}. \tag{5.8}$$

For instance, the example of Section 3 is recovered by choosing $\Gamma = \{\, (h,\xi) \in \Gamma_0 \mid h \equiv 1 \,\}$,

$$\mathrm{F}_j(A_j\rho) = \frac{1}{2}\|A_j\rho - b_j\|_2^2, \quad \text{and} \quad w(h,\xi) = \int_0^1 \alpha + \frac{\beta}{2}|\xi'(t)|^2\,\mathrm{d}t. \tag{5.9}$$

The choice of φ in Bredies et al. [2020b] was equivalent to $\varphi(\gamma) = \frac{w(\gamma)}{\mathrm{E}(0)}$, whereas we suggest the default of $\varphi(\gamma) = \frac{\alpha}{\mathrm{E}(0)}$. Both functions are strictly positive but sufficiently small not to change the minimisers of E or $\mathcal{E}$ (i.e. $\operatorname{argmin}_{\sigma\in D} \mathrm{E}(\sigma) = \operatorname{argmin}_{\sigma\in\mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma)$, see Section 3.2).

The generality of Γ allows us to treat different applications with the same analysis, for example:

$\boldsymbol{\Gamma \subset \{\, (\mathbf{h},\xi) \in \Gamma_0 \mid \mathbf{h} \equiv \mathbf{1} \,\}}$ This accounts for the example of Section 3. If $\sigma \in \mathcal{M}^+(\Gamma)$, then $\Theta(\sigma) \geq 0$ and mass is preserved on paths (e.g. $t \mapsto \int_\Omega \mathrm{d}\Theta(\sigma)_t$ is constant).

$\boldsymbol{\Gamma \subset \{\, (\mathbf{h},\xi) \in \Gamma_0 \mid \mathbf{h} \in \mathscr{C}([\mathbf{0},\mathbf{1}]\,;[\mathbf{0},\mathbf{1}]) \,\}}$ This accounts for the example of Section 4. We still have $\Theta(\sigma) \geq 0$, but $t \mapsto \int_\Omega \mathrm{d}\Theta(\sigma)_t$ is not (necessarily) constant. In particular, mass can be created or destroyed (continuously) at any time.

$\boldsymbol{\Gamma \subset \Gamma_\infty = \{\, (\mathbf{h},\xi) \in \Gamma_0 \mid \|\mathbf{h}\|_\infty \leq \mathbf{1} \,\}}$ In the general case $\Theta(\sigma)$ is a general signed measure, even when $\sigma \geq 0$. In words, σ can give positive weight to negative mass curves. The only constraint is that $\Theta(\sigma) \in \mathscr{C}_w([0,1]\,;\mathcal{M}(\overline{\Omega}))$ is continuous in time.

Any choice of energy in this framework leads to a sparse reconstruction in the space of curves. The following theorem is equivalent to that of Theorems 3.1 and 4.1.

Theorem 5.1. *If A_j are given by (5.1) and $\varphi, w\colon \Gamma \to [0,+\infty]$ are lower semi-continuous, then* F *and* E *are lower semi-continuous. Furthermore, if* F *is bounded below, w or φ have compact sub-levelsets, and $\inf_{\gamma\in\Gamma} \varphi(\gamma) > 0$, then $\mathrm{E}\,|_D$ has compact sub-levelsets. There exists a choice of minimiser $\sigma^* \in \operatorname{argmin}_{\sigma\in D} \mathrm{E}(\sigma)$ such that*

$$\text{for some } a_i \geq 0,\ \gamma^i \in \Gamma, \qquad \sigma^* = \sum_{i=1}^{s} a_i \delta_{\gamma^i} \tag{5.10}$$

for some $s \leq m(T+1)+1$. If in addition $\int_\Gamma \varphi\,\mathrm{d}\sigma^* < 1$, then $s \leq m(T+1)$.

The proof is given in Appendix A.

5.2 Inexact sliding Frank-Wolfe algorithm

As previously stated, we will use a variant of the Frank-Wolfe algorithm to take advantage of the sparse structure of reconstructions. We will also allow for inexact computation of the linear oracle, as is necessary in this application. With a slight abuse of notation, define $\mathrm{F}' \colon D \to \mathscr{C}_b(\Gamma)$ so

$$\forall \sigma \in D, \qquad \mathrm{F}'(\sigma) = \left[(h, \xi) \mapsto \sum_{j=0}^{T} h(t_j) \eta_j(\xi(t_j)) \right] \quad \text{where} \quad \eta_j \overset{\text{def.}}{=} A_j^* \nabla \mathrm{F}_j(A_j(e_{t_j})_\sharp \sigma). \tag{5.11}$$

It can be verified that F' is a representation of the Gateaux derivative of F. For all $\sigma \in D$ define the exact linear oracle $\Phi \colon D \to D$ by a choice

$$\Phi(\sigma) \in \operatorname*{argmin}_{\widetilde{\sigma} \in D} \sum_{j=0}^{T} \int_\Gamma \big(\mathrm{F}'(\sigma)[\gamma] + w(\gamma) \big) \, \mathrm{d}\widetilde{\sigma}(\gamma). \tag{5.12}$$

Lemma 5.4 will justify that this minimiser is well-defined. The following result is our main convergence proof for Algorithm 2, which clarifies the sense of approximation $\mu^{n+1} \approx \Phi(\sigma^n)$.

Theorem 5.2. *Suppose $\{\sigma^n, \mu^n, \lambda_n\}_{n \in \mathbb{N}}$ are generated by Algorithm 2 and there exists $\sigma^* \in \operatorname{argmin}_{\sigma \in D} \mathrm{E}(\sigma)$. If* F *has bounded curvature, i.e.*

$$C \overset{\text{def.}}{=} 2 \sup_{\substack{\sigma, \widetilde{\sigma} \in D \\ \lambda \in]0,1[}} \frac{\mathrm{F}((1-\lambda)\sigma + \lambda\widetilde{\sigma}) - \mathrm{F}(\sigma) - \lambda \int_\Gamma \mathrm{F}'(\sigma) \, \mathrm{d}[\widetilde{\sigma} - \sigma]}{\lambda^2} < +\infty, \tag{5.13}$$

and

$$\liminf_{n \to +\infty} \mathrm{gap}_n \leq 0 \quad \textit{where} \quad \mathrm{gap}_n \overset{\text{def.}}{=} \int_\Gamma \big(\mathrm{F}'(\sigma^n) + w \big) \, \mathrm{d}[\mu^{n+1} - \sigma^*], \tag{5.14}$$

then $\lim_{n \to +\infty} \mathrm{E}(\sigma^n) = \mathrm{E}(\sigma^*)$.

The proof of this theorem is in Section 6.1. Note that gap_n is equivalent to the standard linear duality gap, c.f. Jaggi [2013], informally written as $\mathrm{gap}_n = \langle \nabla \mathrm{E}(\sigma^n), \Phi(\sigma^n) - \sigma^* \rangle$. The motivation for this formulation is to enable the proof of almost-sure convergence for a stochastic variant.

Theorem 5.3. *Let σ^* and $\{\sigma^n, \mu^n, \lambda_n\}_{n \in \mathbb{N}}$ be as in Theorem 5.2 where μ^n is chosen stochastically. Let $\mathcal{S}^N$ be the filtration generated by $\{\sigma^n, \mu^n\}_{0 \leq n \leq N}$, and let $\delta \colon [0, +\infty[\to [0, 1]$ such that*

$$\forall n \in \mathbb{N}, \ \varepsilon \geq 0, \qquad \mathbb{P}(\mathrm{gap}_n \leq \varepsilon | \mathcal{S}^n) \geq \delta(\varepsilon) \quad \mathcal{S}^n\textit{-almost surely.} \tag{5.15}$$

If F *has bounded curvature and $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$, then $\lim_{n \to +\infty} \mathrm{E}(\sigma^n) = \mathrm{E}(\sigma^*)$ almost-surely.*

The proof of this theorem is in Section 6.2. The key idea is that for every n there is uniformly positive chance of finding a good approximation of σ^* or $\Phi(\sigma^{n-1})$.

Silveti-Falls et al. [2020] proves a similar convergence result in the stochastic setting, our random gap_n is an equivalent error estimate to their "λ_k^s". Their proof requires (assumption P.8) something similar to $\sum_{n=0}^{\infty} \lambda_n \, \mathrm{gap}_n < +\infty$ where λ_n is chosen deterministically, which necessitates the random error decaying at an a priori known rate. This is unrealistic in our application, and so we allow λ_n to be adaptive to stabilise the convergence.

5.3 Inexact linear oracle computation

As in Section 3.4, the most challenging computational component of Algorithm 2 is evaluating the (inexact) linear oracle, i.e. (5.12). Similarly to Section 3.4, we can show that it is equivalent to optimise directly over Γ.

Algorithm 2 Inexact sliding Frank-Wolfe algorithm

Choose $\sigma^0 \in D$ with $\mathrm{E}(\sigma^0) < +\infty$, $n \leftarrow 0$,
repeat
  $\mu^{n+1} \approx \Phi(\sigma^n)$ ▷ inexact linear oracle
  $\lambda_n \leftarrow \operatorname{argmin}_{\lambda\in[0,1]} \mathrm{E}((1-\lambda)\sigma^n + \lambda\mu^{n+1})$ ▷ exact line search
  Choose σ^{n+1} such that $\mathrm{E}(\sigma^{n+1}) \leq \mathrm{E}((1-\lambda_n)\sigma^n + \lambda_n\mu^{n+1})$ ▷ Sliding step
  $n \leftarrow n+1$
until converged

Lemma 5.4. *Suppose $\varphi, w\colon \Gamma \to [0,+\infty]$ are lower semi-continuous. If $\inf_{\gamma\in\Gamma}\varphi(\gamma) > 0$ and either φ or w have compact sub-levelsets, then there exists a choice of $\Phi(\widetilde{\sigma})$ such that*

$$\Phi(\widetilde{\sigma}) \in \{0\} \cup \left\{ \varphi(\gamma^*)^{-1}\delta_{\gamma^*} \mid \gamma^* \in \operatorname*{argmin}_{\gamma\in\Gamma} \frac{\eta(\gamma)+w(\gamma)}{\varphi(\gamma)} \right\}. \tag{5.16}$$

where $\eta(h,\xi) \stackrel{\text{def.}}{=} \sum_{j=0}^{T} h(t_j)\eta_j(\xi(t_j)) \in \mathscr{C}_b(\Gamma)$.

Proof. Recall the definition of Φ,

$$\Phi(\widetilde{\sigma}) \in \operatorname*{argmin}_{\sigma\in D} \widetilde{\mathrm{E}}(\sigma) \quad \text{where} \quad \widetilde{\mathrm{E}}(\sigma) \stackrel{\text{def.}}{=} \int_\Gamma (\eta + w)\,\mathrm{d}\sigma. \tag{5.17}$$

The properties of D come from Lemma A.5, in particular D is convex, closed and bounded, so $\inf_{\sigma\in D}\int_\Gamma \eta\,\mathrm{d}\sigma > -\infty$ and $\widetilde{\mathrm{E}}$ is lower semi-continuous (Lemma A.1). It is therefore well-posed to consider minimisers of $\widetilde{\mathrm{E}}$. We show that there is a choice $\Phi(\widetilde{\sigma}) = \sigma^* \in \operatorname{Ext}(D)$.

Case D is compact: If φ has compact sub-levelsets, then D is compact by Lemma A.5. Bauer's principle therefore states that there exists a point

$$\sigma^* \in \operatorname{Ext}(D) \quad \text{such that} \quad \widetilde{\mathrm{E}}(\sigma^*) = \inf_{\sigma\in D} \widetilde{\mathrm{E}}(\sigma). \tag{5.18}$$

Else: Otherwise, w has compact sub-levelsets, so the sub-levelset

$$U \stackrel{\text{def.}}{=} \left\{ \sigma \in D \quad | \quad \widetilde{\mathrm{E}}(\sigma) \leq 1 \right\} \tag{5.19}$$

is compact by Theorem A.6, convex, and non-empty because $0 \in D$, $\widetilde{\mathrm{E}}(0) = 0$. Application of Bauer's principle now gives a point $\sigma^* \in \operatorname{Ext}(U)$, to complete the proof we will confirm $\sigma^* \in \operatorname{Ext}(D)$.

Suppose for contradiction that there exists $\sigma_0, \sigma_1 \in D \setminus \{\sigma^*\}$ with $\sigma^* \in]\sigma_0, \sigma_1[$. Without loss of generality, we assume that $\sigma^* = \frac{1}{2}\sigma_0 + \frac{1}{2}\sigma_1$ and we define

$$\forall \lambda \in [0,1], \qquad \sigma_\lambda \stackrel{\text{def.}}{=} \lambda\sigma_1 + (1-\lambda)\sigma_0 \in D, \qquad \sigma_{\frac{1}{2}} = \sigma^*. \tag{5.20}$$

Consider the function $f(\lambda) = \widetilde{\mathrm{E}}(\sigma_\lambda)$, which is linear and $f(\frac{1}{2}) = \widetilde{\mathrm{E}}(\sigma^*) \leq \widetilde{\mathrm{E}}(0) = 0$. As f is continuous, there exists $\varepsilon > 0$ with $f(\frac{1}{2} \pm \varepsilon) \leq \frac{1}{2}$. We conclude that $\sigma_{\frac{1}{2}\pm\varepsilon} \in U$ and $\sigma^* \in \left]\sigma_{\frac{1}{2}-\varepsilon}, \sigma_{\frac{1}{2}+\varepsilon}\right[$, contradicting the assumption that $\sigma^* \in \operatorname{Ext}(U)$.

In both cases we see there is a choice $\Phi(\widetilde{\sigma}) = \sigma^* \in \operatorname{Ext}(D)$. Finally, by Lemma A.5 we have

$$\operatorname{Ext}(D) = \{0\} \cup \left\{ \varphi(\gamma)^{-1}\delta_\gamma \mid \varphi(\gamma) < +\infty \right\}, \tag{5.21}$$

so (5.16) is just the required statement

$$\sigma^* \in \operatorname*{argmin}_{\sigma\in\operatorname{Ext}(D)} \widetilde{\mathrm{E}}(\sigma) \subset \{0\} \cup \left\{ \varphi(\gamma^*)^{-1}\delta_{\gamma^*} \mid \gamma^* \in \operatorname*{argmin}_{\gamma\in\Gamma} \widetilde{\mathrm{E}}\left(\varphi(\gamma)^{-1}\delta_\gamma\right) \right\}. \tag{5.22}$$

□

We use this formulation numerically because it is easier to optimise over a set of curves than a set of measures on curves. Because of this, we define our discrete minimiser

$$\Phi_d(\sigma) \in \{0\} \cup \left\{ \varphi(\gamma_d^*)^{-1}\delta_{\gamma_d^*} \mid \gamma_d^* \in \operatorname*{argmin}_{\gamma \in \prod_{j=0}^T \Lambda_j} \frac{\eta(\gamma) + w(\gamma)}{\varphi(\gamma)} \right\} \tag{5.23}$$

for some discrete sets $\Lambda_j \subset \mathbb{R} \times \overline{\Omega}$. For us to compute this discrete minimiser efficiently, we must further assume

$$\varphi(\gamma) = \varphi_0 \qquad \text{for some constant } \varphi_0 > 0, \tag{A1}$$

$$w(\gamma) = \sum_{j=1}^{T} \operatorname{step}_j(\gamma(t_{j-1}), \gamma(t_j)) \qquad \text{for some } \operatorname{step}_j \colon (\mathbb{R} \times \overline{\Omega})^2 \to \mathbb{R}. \tag{A2}$$

These properties were clear in Section 3.4 where for all $\gamma = (h, \xi) \in \Gamma$

$$\varphi_0 = \frac{\alpha}{\mathrm{E}(0)}, \qquad \operatorname{step}_j(\gamma(t_{j-1}), \gamma(t_j)) = \alpha(t_j - t_{j-1}) + \frac{\beta}{2}\frac{|\xi(t_j) - \xi(t_{j-1})|^2}{t_j - t_{j-1}}. \tag{5.24}$$

In particular, the step function of the Benamou-Brenier penalty is Wasserstein optimal transport, and the step function of the dynamic Wasserstein-Fisher-Rao penalty is the static Wasserstein-Fisher-Rao penalty.

As before, it is possible to recover the continuous curve γ from the vector $(\gamma(t_0), \dots, \gamma(t_T))$ using geodesics of w. If $\sigma^* \in \operatorname{argmin}_{\sigma \in \mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma)$ and $\int_\Gamma \varphi \,\mathrm{d}\sigma^* < 1$, then σ^* is supported on geodesics of w (Lemma B.1). Otherwise, if $\int_\Gamma \varphi \,\mathrm{d}\sigma^* = 1$, then the continuous-time interpolation of γ must balance the costs of $w(\gamma)$ and $\varphi(\gamma)$.

Under assumptions (A1) and (A2), the problem of evaluating Φ_d can be formulated as finding the minimal path on an ordered, weighted, directed acyclic graph. The vertices are the points (t_j, y) for $y \in \Lambda_j$, the edge weights are given by η_j and step_j, and the time index provides an ordering and prevents cycles in the graph. In Section 7 we will write out the algorithm fully for clarity, but we state here the known complexity bound.

Theorem 5.5 ([Cormen et al., 2009, Sec. 24.2])**.** *Suppose assumptions* (A1) *and* (A2) *hold and* $|\Lambda_j| \leq N$ *for each* j*, then* γ_d^* *can be computed with complexity*

$$\textit{total time} = O\left(N \sum_{j=0}^{T} \operatorname{cost}(\eta_j) + N^2 \sum_{j=1}^{T} \operatorname{cost}(\operatorname{step}_j) \right) \tag{5.25}$$

where $\operatorname{cost}(\eta_j)$ *is the cost of evaluating* η_j *once etc.*

Remark 5.1. *In graph terminology, the two terms of* (5.25) *represent the number of vertices plus the number of edges, asymptotically* $O(NT)$ *and* $O(N^2T)$ *respectively. If, for example, the paths have a maximum velocity* V*, then the number of edges per vertex is reduced from* N *to* $O(N(VT^{-1})^d)$*. This results in a reduced total complexity of* $O(NT + N^2T^{1-d}V^d)$*.*

6 Frank-Wolfe convergence

In this section we analyse Algorithm 2, first proving Theorem 5.2 in Section 6.1, then the stochastic variant Theorem 5.3 is in Section 6.2.

6.1 Deterministic Frank-Wolfe analysis

The proof of Theorem 5.2 closely follows the argument of [Jaggi, 2013, Thm. 1] where $\operatorname{gap}_n = O(n^{-1})$, whereas we only assume $\operatorname{gap}_{n_k} = O(k^{-1})$ on some sub-sequence n_k. This relaxed assumption on gap_n is accounted for by the linesearch in λ. This guarantees both that small

steps are made for bad search directions, and large steps are made for good ones. For example, for all $n \in \mathbb{N}$ and $\mu^{n+1} \in D$, we have a monotone descent,

$$\mathrm{E}(\sigma^{n+1}) \leq \mathrm{E}((1-\lambda_n)\sigma^n + \lambda_n \mu^{n+1}) = \inf_{\lambda \in [0,1]} \mathrm{E}((1-\lambda)\sigma^n + \lambda \mu^{n+1}) \leq \mathrm{E}(\sigma^n). \tag{6.1}$$

The remaining assumption in Theorem 5.2 is the smoothness of F, i.e. $C < +\infty$ where

$$C \stackrel{\text{def.}}{=} \max\left(\frac{\mathrm{E}(\sigma^0) - \mathrm{E}(\sigma^*)}{2},\ 2 \sup_{\substack{\sigma, \widetilde{\sigma} \in D \\ \lambda \in]0,1[}} \frac{\mathrm{F}((1-\lambda)\sigma + \lambda\widetilde{\sigma}) - \mathrm{F}(\sigma) - \lambda \int_\Gamma \mathrm{F}'(\sigma)\,\mathrm{d}[\widetilde{\sigma} - \sigma]}{\lambda^2} \right). \tag{6.2}$$

For all $n \in \mathbb{N}$ and $\lambda \in [0,1]$, by the convexity of F we have

$$\begin{aligned}
\mathrm{F}((1-\lambda)\sigma^n + \lambda\mu^{n+1}) &\leq \mathrm{F}(\sigma^n) + \lambda \int_\Gamma \mathrm{F}'(\sigma^n)\,\mathrm{d}[\mu^{n+1} - \sigma^n] + \frac{C\lambda^2}{2} && (6.3)\\
&= (1-\lambda)\,\mathrm{F}(\sigma^n) + \lambda \left[\mathrm{F}(\sigma^n) + \int_\Gamma \mathrm{F}'(\sigma^n)\,\mathrm{d}[\mu^{n+1} - \sigma^n] \right] + \frac{C\lambda^2}{2} && (6.4)\\
&= (1-\lambda)\,\mathrm{F}(\sigma^n) + \lambda \left[\mathrm{F}(\sigma^n) + \int_\Gamma \mathrm{F}'(\sigma^n)\,\mathrm{d}[\sigma^* - \sigma^n + \mu^{n+1} - \sigma^*] \right] + \frac{C\lambda^2}{2} && (6.5)\\
&\leq (1-\lambda)\,\mathrm{F}(\sigma^n) + \lambda \left[\mathrm{F}(\sigma^*) + \int_\Gamma \mathrm{F}'(\sigma^n)\,\mathrm{d}[\mu^{n+1} - \sigma^*] \right] + \frac{C\lambda^2}{2}. && (6.6)
\end{aligned}$$

The linearity of W gives the same approximation,

$$\begin{aligned}
\mathrm{W}((1-\lambda)\sigma^n + \lambda\mu^{n+1}) &= (1-\lambda)\,\mathrm{W}(\sigma^n) + \lambda\,\mathrm{W}(\mu^{n+1}) && (6.7)\\
&= (1-\lambda)\,\mathrm{W}(\sigma^n) + \lambda \left[\mathrm{W}(\sigma^*) + \mathrm{W}(\mu^{n+1} - \sigma^*) \right]. && (6.8)
\end{aligned}$$

Recalling $\mathrm{gap}_n = \int_\Gamma \left(\mathrm{F}'(\sigma^n) + w \right) \mathrm{d}[\mu^{n+1} - \sigma^*]$, we summarise

$$\mathrm{E}((1-\lambda)\sigma^n + \lambda\mu^{n+1}) \leq (1-\lambda)\,\mathrm{E}(\sigma^n) + \lambda[\mathrm{E}(\sigma^*) + \mathrm{gap}_n] + \frac{C}{2}\lambda^2. \tag{6.9}$$

We can now combine these approximations to complete the proof.

Proof of Theorem 5.2. Since $\liminf_{n\to\infty} \mathrm{gap}_n \leq 0$ there exist a subsequence $(n_k)_{k\in\mathbb{N}}$ such that $\mathrm{gap}_{n_k} \leq \frac{C}{k}$ for all $k \geq 1$. For all $\lambda \in [0,1]$,

$$\begin{aligned}
\mathrm{E}(\sigma^{n_{k+1}}) - \mathrm{E}(\sigma^*) &\leq \mathrm{E}(\sigma^{n_k+1}) - \mathrm{E}(\sigma^*) && (6.10)\\
&\leq \mathrm{E}((1-\lambda)\sigma^{n_k} + \lambda\mu^{n_k+1}) - \mathrm{E}(\sigma^*) && (6.11)\\
&= (1-\lambda)[\mathrm{E}(\sigma^{n_k}) - \mathrm{E}(\sigma^*)] + \lambda\,\mathrm{gap}_{n_k} + \frac{C}{2}\lambda^2. && (6.12)
\end{aligned}$$

Now we proceed by induction. Note that $\mathrm{E}(\sigma^{n_2}) - \mathrm{E}(\sigma^*) \leq \mathrm{E}(\sigma^0) - \mathrm{E}(\sigma^*) \leq 2C$ by (6.1) and (6.2). Next, if for some $k \geq 2$, $\mathrm{E}(\sigma^{n_k}) - \mathrm{E}(\sigma^*) \leq \frac{4C}{k}$, we choose $\lambda = \frac{2}{k} \in [0,1]$, and we observe

$$\begin{aligned}
\mathrm{E}(\sigma^{n_{k+1}}) - \mathrm{E}(\sigma^*) &\leq \left(1 - \frac{2}{k}\right)\frac{4C}{k} + \frac{2}{k}\frac{C}{k} + \frac{C}{2}\frac{4}{k^2} && (6.13)\\
&= \frac{4C}{k}\left(1 - \frac{2}{k} + \frac{1}{k}\right) && (6.14)\\
&\leq \frac{4C}{k}\left(1 - \frac{1}{k+1}\right) && (6.15)\\
&= \frac{4C}{k+1}. && (6.16)
\end{aligned}$$

We conclude by induction that the inequality holds for all $k \geq 2$. By monotonicity, $\lim_{n\to+\infty} \mathrm{E}(\sigma^n) = \lim_{k\to+\infty} \mathrm{E}(\sigma^{n_k}) = \mathrm{E}(\sigma^*)$ as required. □

6.2 Stochastic Frank-Wolfe analysis

The previous proof shows that Algorithm 2 converges whenever $\liminf_{n\to+\infty} \text{gap}_n = 0$. The main aim in this section is to prove Theorem 5.3 which shows this is satisfied almost surely. We will then explore two explicit and practical examples in which it applies.

6.2.1 Convergence proof

In Theorem 5.3 we assume $\mathbb{P}(\text{gap}_n \leq \varepsilon | \mathcal{S}^n) \geq \delta(\varepsilon) > 0$ for all $n \in \mathbb{N}$ and $\varepsilon > 0$. The first task is to translate this assumption into a statement about $\liminf_{n\to+\infty} \text{gap}_n$. For fixed $\varepsilon > 0$, observe

$$\begin{aligned}
\mathbb{P}\left[\liminf_{n\to+\infty} \text{gap}_n > \varepsilon\right] &= \mathbb{P}\left[\bigcup_{N\in\mathbb{N}} \left\{\inf_{n\geq N} \text{gap}_n > \varepsilon\right\}\right] && (6.17)\\
&\leq \sum_{N\in\mathbb{N}} \mathbb{P}\left[\inf_{n\geq N} \text{gap}_n > \varepsilon\right] && \text{(sub-additivity)} \quad (6.18)\\
&\leq \sum_{N\in\mathbb{N}} \mathbb{P}\left[\bigcap_{n\geq N} \{\text{gap}_n > \varepsilon\}\right] && (6.19)\\
&= \sum_{N\in\mathbb{N}} \lim_{M\to+\infty} \mathbb{P}[G_N^M] && \text{(monotone convergence)} \quad (6.20)
\end{aligned}$$

where

$$G_N^M \overset{\text{def.}}{=} \bigcap_{n=N}^{M} \{\text{gap}_n > \varepsilon\}. \tag{6.21}$$

The probability of these events can be bounded inductively for $M \geq N$,

$$\begin{aligned}
\mathbb{P}(G_N^{M+1}) &= \mathbb{E}\left[\mathbb{E}(\mathbb{1}_{G_N^M \cap \{\text{gap}_{M+1} > \varepsilon\}} | \mathcal{S}^M)\right] && (6.22)\\
&= \mathbb{E}\left[\mathbb{E}(\mathbb{1}_{G_N^M} \cdot \mathbb{1}_{\{\text{gap}_{M+1} > \varepsilon\}} | \mathcal{S}^M)\right] && (6.23)\\
&= \mathbb{E}\left[\mathbb{1}_{G_N^M} \cdot \mathbb{E}(\mathbb{1}_{\{\text{gap}_{M+1} > \varepsilon\}} | \mathcal{S}^M)\right] && (G_N^M \in \mathcal{S}^M) \quad (6.24)\\
&= \mathbb{E}\left[\mathbb{1}_{G_N^M} \cdot \mathbb{P}(\text{gap}_{M+1} > \varepsilon | \mathcal{S}^M)\right] && \text{(def. of } \mathbb{P}) \quad (6.25)\\
&\leq \mathbb{E}\left[\mathbb{1}_{G_N^M} \cdot (1 - \delta(\varepsilon))\right] && \text{(def. of } \delta) \quad (6.26)\\
&= (1 - \delta(\varepsilon))\mathbb{P}(G_N^M) \ \ldots && (6.27)\\
&\leq (1 - \delta(\varepsilon))^{M+1-N}. && (6.28)
\end{aligned}$$

We can now complete the proof.

Proof of Theorem 5.3. To see $\lim_{n\to+\infty} \mathrm{E}(\sigma^n) = \mathrm{E}(\sigma^*)$ almost-surely, observe

$$\begin{aligned}
\mathbb{P}(\lim_{n\to+\infty} \mathrm{E}(\sigma^n) = \mathrm{E}(\sigma^*)) &\geq \mathbb{P}(\liminf_{n\to+\infty} \text{gap}_n \leq 0) && \text{(Theorem 5.2)} \quad (6.29)\\
&= 1 - \mathbb{P}(\liminf_{n\to+\infty} \text{gap}_n > 0) && (6.30)\\
&= 1 - \lim_{\varepsilon\to 0} \mathbb{P}(\liminf_{n\to+\infty} \text{gap}_n > \varepsilon) && \text{(monotone convergence)} \quad (6.31)\\
&\geq 1 - \lim_{\varepsilon\to 0} \sum_{N\in\mathbb{N}} \lim_{M\to+\infty} \mathbb{P}\left[G_N^M\right] && (6.20) \quad (6.32)\\
&\geq 1 - \lim_{\varepsilon\to 0} \sum_{N\in\mathbb{N}} \lim_{M\to+\infty} (1 - \delta(\varepsilon))^{M-N} && (6.28) \quad (6.33)\\
&= 1 && (6.34)
\end{aligned}$$

as required. □

6.2.2 Independent random gaps variant

In this section we focus on the case

$$\mu^n = \widetilde{\Phi}(\sigma^{n-1}, z^n) \tag{6.35}$$

for some approximate linear oracle $\widetilde{\Phi}$, every $n \in \mathbb{N}$, where $z^n \sim Z$ is a random instance of an ambient probability density. In the context of Section 5.3, z^n could be a random discretisation of Γ and μ^n is the discrete minimiser. The following corollary shows that, if z^n are independent and have a uniformly positive probability of being close to σ^*, then the conditions of Theorem 5.3 are satisfied.

Corollary 6.1. *Suppose $\mu^n = \widetilde{\Phi}(\sigma^{n-1}, z^n)$ for some $z^n \sim Z$ and define $\widetilde{\delta}\colon [0, +\infty[\to [0, 1]$ by*

$$\widetilde{\delta}(\varepsilon) \stackrel{\text{def.}}{=} \inf_{\sigma \in D} \mathbb{P}_{z \sim Z} \left(\int_\Gamma \mathrm{F}'(\sigma) + w \, \mathrm{d}[\widetilde{\Phi}(\sigma, z) - \sigma^*] \leq \varepsilon \right). \tag{6.36}$$

If z^n is independent of the filtration $\widetilde{\mathcal{S}}^{n-1}$ generated by $\{z^k, \sigma^k\}_{k=0}^{n-1}$ for all $n \in \mathbb{N}^*$, then $\widetilde{\delta}$ is a valid choice for δ in Theorem 5.3.

Lemma 6.2. *Suppose X is an $\mathcal{S}$-measurable random variable and $Z \perp X$ (i.e. Z is independent of X) is another random variable. Then, for all integrable Borel measurable functions g*

$$\mathbb{E}(g(X, Z)|\mathcal{S}) = \mathbb{E}\left[\mathbb{E}_Z(g(X, Z))|\mathcal{S}\right]. \tag{6.37}$$

This is a standard result in probability theory for which a proof is provided in Appendix C.

Proof of Corollary 6.1. We need to link the values of δ from Theorem 5.3 with $\widetilde{\delta}$. Recall that $\mathrm{gap}_n = \int_\Gamma \left(\mathrm{F}'(\sigma^n) + w\right) \mathrm{d}[\widetilde{\Phi}(\sigma^n, z^{n+1}) - \sigma^*]$ is measurable w.r.t. the filtration generated by $\{\widetilde{\mathcal{S}}^n, z^{n+1}\}$, with $\widetilde{\mathcal{S}}^n \perp z^{n+1}$. We therefore have

$$\begin{aligned}
\mathbb{P}(\mathrm{gap}_n \leq \varepsilon | \mathcal{S}^n) &\equiv \mathbb{E}\left[\mathbb{1}_{\mathrm{gap}_n \leq \varepsilon} | \mathcal{S}^n\right] && (6.38)\\
&= \mathbb{E}\left[\mathbb{E}(\mathbb{1}_{\mathrm{gap}_n \leq \varepsilon} | \widetilde{\mathcal{S}}^n) | \mathcal{S}^n\right] && (\widetilde{\mathcal{S}}^n \supset \mathcal{S}^n) \quad (6.39)\\
&= \mathbb{E}\left[\mathbb{E}(\mathbb{E}_{z^{n+1} \sim Z}(\mathbb{1}_{\mathrm{gap}_n \leq \varepsilon}) | \widetilde{\mathcal{S}}^n) | \mathcal{S}^n\right] && (\text{Lemma } 6.2) \quad (6.40)\\
&\geq \mathbb{E}\left[\mathbb{E}(\widetilde{\delta}(\varepsilon) | \widetilde{\mathcal{S}}^n) | \mathcal{S}^n\right] && (\text{def. of } \widetilde{\delta}) \quad (6.41)\\
&= \widetilde{\delta}(\varepsilon). && (6.42)
\end{aligned}$$

This bound is uniform in n, therefore $\widetilde{\delta}$ satisfies the definition of δ. □

A natural choice of $\widetilde{\Phi}$ might be where z^n is a random discretisation and $\widetilde{\Phi}$ is the discrete minimiser from (5.23). In the context of this work, we expect $\mathrm{supp}(\sigma^*)$ to be a discrete set and can choose $z^n \subset ([-1, 1] \times \overline{\Omega})^{T+1}$ (which is compact). If $|z^n| \geq |\mathrm{supp}(\sigma^*)|$, then it should be straight-forward to show $\mathbb{P}_{z^n \sim Z}\left(\int_\Gamma \mathrm{F}'(\sigma^n) + w \, \mathrm{d}[\widetilde{\Phi}(\sigma^{n-1}, z^n) - \sigma^*] \leq \varepsilon\right) > 0$. To show $\widetilde{\delta}(\varepsilon) > 0$, the final challenge is to show uniformity in σ^{n-1}. For example, this follows if $\mathrm{F}'(\sigma) + w \in \mathscr{C}(([-1, 1] \times \overline{\Omega})^{T+1})$ is Lipschitz with Lipschitz constant independent of σ.

Another natural approach would be, for known σ^{n-1}, to choose for z^n with guaranteed approximation properties for $\Phi(\sigma^{n-1})$, rather than the fixed but unknown target of σ^*. This approach is addressed by the following corollary.

Corollary 6.3. *Suppose $\mu^n = \Phi(\sigma^{n-1}, z^n)$ for some $z^n \sim Z$ and define $\widetilde{\delta}\colon [0, +\infty[\to [0, 1]$ by*

$$\widetilde{\delta}(\varepsilon) \stackrel{\text{def.}}{=} \inf_{\sigma \in D} \mathbb{P}_{z \sim Z} \left(\int_\Gamma \left(\mathrm{F}'(\sigma) + w\right) \mathrm{d}[\Phi(\sigma, z) - \Phi(\sigma)] \leq \varepsilon \right). \tag{6.43}$$

If z^n is independent of the filtration $\widetilde{\mathcal{S}}^{n-1}$ generated by $\{z^k, \sigma^k\}_{k=0}^{n-1}$ for all $n \in \mathbb{N}^*$, then $\widetilde{\delta}$ is a valid choice for δ in Theorem 5.3.

Proof. This corollary is a strictly weaker statement than the previous. Fix $\sigma \in D$ and define $\eta \overset{\text{def.}}{=} \mathrm{F}'(\sigma) + w$, then

$$\int_\Gamma \eta \,\mathrm{d}[\Phi(\sigma, z) - \Phi(\sigma)] \leq \varepsilon \iff \int_\Gamma \eta \,\mathrm{d}[\Phi(\sigma, z) - \sigma^*] \leq \int_\Gamma \eta \,\mathrm{d}[\Phi(\sigma) - \sigma^*] + \varepsilon \tag{6.44}$$

$$\iff \int_\Gamma \eta \,\mathrm{d}[\Phi(\sigma, z) - \sigma^*] \leq \inf_{\widetilde{\sigma} \in D} \int_\Gamma \eta \,\mathrm{d}[\widetilde{\sigma} - \sigma^*] + \varepsilon \tag{6.45}$$

$$\implies \int_\Gamma \eta \,\mathrm{d}[\Phi(\sigma, z) - \sigma^*] \leq \varepsilon. \tag{6.46}$$

We conclude that the assumptions of Corollary 6.1 hold, and so the conclusion is the same. □

In this case, we see that $\widetilde{\Phi}$ only needs to approximate the exact linear oracle $\Phi(\sigma^{n-1})$, so we can have convergence by choosing random discretisations $z^n \subset ([-1,1] \times \overline{\Omega})^{T+1}$ with $|z^n| \geq |\operatorname{supp}(\Phi(\sigma^{n-1}))| = 1$. In our setting, the criterion of this second corollary is much easier to verify.

7 General dynamical programming method

As stated at the end of Section 5.3, we compute optimal curves of (5.23) using the standard algorithm for minimal paths on ordered, weighted, directed, acyclic graphs described in [Cormen et al., 2009, Sec. 24.2]. In this section we expand on the precise computations for a fixed $\sigma \in D$. Recall for all $\gamma \in \Gamma \subset \Gamma_\infty$ we have $\gamma(t) \in [-1,1] \times \overline{\Omega}$ for all $t \in [0,1]$. Under assumptions (A1) and (A2), the optimisation problem can therefore be written as

$$\gamma_d^* \in \operatorname*{argmin}_\gamma \left\{ \sum_{j=0}^T \eta_j(\gamma_j) + \sum_{j=1}^T \operatorname{step}_j(\gamma_{j-1}, \gamma_j) \mid \gamma \in \prod_{j=0}^T \Lambda_j \right\} \tag{7.1}$$

$$\text{where} \quad \eta_j(h_j, \xi_j) \overset{\text{def.}}{=} h_j[A_j^* \nabla \mathrm{F}_j(A_j(e_{t_j})_\sharp \sigma)](\xi_j) \tag{7.2}$$

and $\Lambda_j \subseteq [-1,1] \times \overline{\Omega}$ are finite sets ("grids" in the mass-location space), $j = 0, \ldots, T$. Similarly to Section 3.4, we define the functions

$$\widetilde{\mathrm{E}}_J(\gamma) \overset{\text{def.}}{=} \sum_{j=0}^J \eta_j(\gamma_j) + \sum_{j=1}^J \operatorname{step}_j(\gamma_{j-1}, \gamma_j), \tag{7.3}$$

$$Y^*[y, J] \in \operatorname*{argmin}_Y \left\{ \widetilde{\mathrm{E}}_J(Y) \mid Y \in \prod_{j=0}^J \Lambda_j,\ Y_J = y \right\} \tag{7.4}$$

for each $J = 0, \ldots, T$ and $y \in \Lambda_J$. This generates γ_d^* through the computation

$$\gamma_d^* = Y^*[y^*, T], \qquad y^* \in \operatorname*{argmin}_{y \in \Lambda_T} \widetilde{\mathrm{E}}_T(Y^*[y, T]). \tag{7.5}$$

Observe that for all $J = 1, \ldots, T$ and $y \in \Lambda_J$,

$$\min_{\substack{Y \in \prod_{j=0}^J \Lambda_j \\ Y_J = y}} \widetilde{\mathrm{E}}_J(Y) = \min_{Y \in \prod_{j=0}^J \Lambda_j} \left\{ \widetilde{\mathrm{E}}_{J-1}(Y) + \eta_J(y) + \operatorname{step}_J(Y_{J-1}, y) \right\} \tag{7.6}$$

$$= \eta_J(y) + \min_{y' \in \Lambda_{J-1}} \left\{ \operatorname{step}_J(y', y) + \min_{\substack{Y \in \prod_{j=0}^{J-1} \Lambda_j \\ Y_{J-1} = y'}} \widetilde{\mathrm{E}}_{J-1}(Y) \right\}. \tag{7.7}$$

In particular, we can choose $Y^*[y, J]$ inductively to be

$$Y^*[y, J] = \begin{pmatrix} Y^*[y', J-1] & y \end{pmatrix} \quad \text{for} \quad y' \in \operatorname*{argmin}_{y' \in \Lambda_{J-1}} \left[\operatorname{step}_J(y', y) + \widetilde{\mathrm{E}}_{J-1}(Y^*[y', J-1]) \right]. \tag{7.8}$$

8 Numerical results

For numerical experiments we implement variants of Algorithm 2 using the linear-oracle strategy discussed in Section 7. The expanded form of this algorithm is given in Algorithm 3. The choice of A_j, F_j, t_j, and step_j define the optimisation problem, then the final choices of Λ_j and k dictate the variant of the algorithm. We always choose the sliding step to select a local minimum of E (or $\widetilde{\mathrm{E}}$) as computed by an implementation of the L-BFGS algorithm. All code to reproduce the results and figures in this work can be found online[1].

The classical sliding Frank-Wolfe algorithm (Algorithm 1) can be recovered by choosing $\Lambda_j = [-1,1] \times \overline{\Omega}$ and $k = 1$. The potential for $k > 1$ was considered in [Bredies et al., 2020b, Sec 5.1.5] as a "multistart" parameter. The idea is that the approximation of optimal curves (e.g. $Y^*[\cdot, T]$) is very expensive, so some of the other near-optimal curves should also be used to improve efficiency of the algorithm. Suppose Λ_j is chosen to approximate $[-1,1] \times \overline{\Omega}$ with a grid of M masses in $[-1,1]$ and N^d points in $\overline{\Omega}$. From Theorem 5.5 we know that the computational complexity of the linear oracle at every iteration is $O((N^d M)^2)$. If the "true curves' are easy to find, then we can hope to reduce this to $O(\frac{N^{2d}M^2}{k})$ per curve. An alternative motivation is provided by Ding et al. [2020] who propose the "kFW" algorithm. They argue that the inclusion of multiple extreme points at each iteration allows the algorithm to move directly towards an extreme face, rather than zig-zagging between extreme points. Note that a similar observation is made in Flinth et al. [2021] in the context of sparse spike recovery, where a collection of Dirac masses are added at each iteration. The sliding step in line 13 of Algorithm 3 further addresses this property, allowing non–zig-zagging movement.

We consider two families of algorithms implementing Algorithm 3, one stochastic and the other deterministic. Throughout this section we fix $\Omega \overset{\text{def.}}{=}]0,1[^2$ and only consider non-negative curves $h \geq 0$, so the algorithms can be stated as:

(k, N, M)-random mesh The "multistart" parameter is k. For each n and j we generate new independent uniformly random points $H = \{h_i^j \sim \mathcal{U}([[0,1]]\}_{i=1}^M$, $X = \{x_i^j \sim \mathcal{U}([[0,1]^2]\}_{i=1}^{N^2}$ and

$$\Lambda_j = \{\, (h,x) \mid h \in H,\ x \in X \,\}. \tag{8.1}$$

When $M = 0$ we take $H = \{1\}$ (i.e. a balanced mesh).

(k, N, M)-uniform mesh The "multistart" parameter is k. For each n we choose $\Lambda_0 = \ldots = \Lambda_T$ to be of the form

$$\Lambda_j = \Lambda_j(\widetilde{N}) \overset{\text{def.}}{=} \left\{ (h,(x_1,x_2)) \mid h \in \{0, M^{-1}, \ldots, 1\},\ x_1, x_2 \in \{0, \widetilde{N}^{-1}, \ldots, 1\} \right\} \tag{8.2}$$

for some $\widetilde{N} \geq 1$. We start with $\widetilde{N} = 16$ at $n = 0$ then increment $\widetilde{N} \leftarrow 2\widetilde{N}$ whenever $\mathrm{E}(\sigma^{n+1}) = \mathrm{E}(\sigma^n)$ (tested after line 13 of Algorithm 3). This process is terminated once $\widetilde{N} > N$. Again, if $M = 0$ then we only allow the balanced mass $h = 1$.

For all problems, $(1, +\infty, +\infty)$-random and $(1, +\infty, +\infty)$-uniform mesh algorithms are equivalent to the exact Algorithm 1. For balanced problems, such as in Section 3, it is sufficient to use the triplet $(1, \infty, 0)$. The random algorithms with $N, M \geq 1$ are guaranteed to converge (eventually) to the exact minimiser by Theorem 5.3. On the other hand, the uniform path algorithms have nice computational properties which allows $Y^*[\cdot, T]$ to be computed more efficiently at larger N. Whilst all results for random meshes are asymptotic, the uniform path algorithms also provide a clear stopping criterion. The algorithm stops at iteration n when $\mathrm{E}(\sigma^{n+1}) = \mathrm{E}(\sigma^n)$, often this will mean $\sigma^{n+1} = \sigma^n$. Let $D_d \overset{\text{def.}}{=} \left\{ \sigma \in D \mid \mathrm{supp}(\sigma) \subset \prod_{j=0}^T \Lambda_j(N) \right\}$ be the discretisation of D. Expanding on the linesearch computation, due to the convexity and smoothness of E,

$$\forall \lambda \in [0,1], \qquad \mathrm{E}((1-\lambda)\sigma^n + \lambda\mu^{n+1}) - \mathrm{E}(\sigma^n) = \lambda \int_\Gamma (\mathrm{F}'(\sigma^n) + w)\, \mathrm{d}[\mu^{n+1} - \sigma^n] + O(\lambda^2). \tag{8.3}$$

[1] https://gitlab.inria.fr/rtovey/DP-for-dynamic-IPs

If the linesearch terminates with $\lambda = 0$, then clearly $\int_\Gamma (\mathrm{F}'(\sigma^n) + w)\,\mathrm{d}[\mu^{n+1} - \sigma^n] \geq 0$. The optimality given by the dual gap (see Jaggi [2013]), can therefore be stated

$$\mathrm{E}(\sigma^n) - \min_{\sigma \in D_d} \mathrm{E}(\sigma) \leq \sup_{\sigma \in D_d} \int_\Gamma (\mathrm{F}'(\sigma^n) + w)\,\mathrm{d}[\sigma^n - \sigma] = \int_\Gamma (\mathrm{F}'(\sigma^n) + w)\,\mathrm{d}[\sigma^n - \mu^{n+1}] \leq 0. \quad (8.4)$$

We conclude that σ^n is at least optimal up to a spatial resolution of $\frac{1}{N}$.

Algorithm 3 Inexact sliding Frank-Wolfe algorithm

Set $\sigma^0 \leftarrow 0 \in \mathcal{M}^+(\Gamma)$, $n \leftarrow 0$, fix $k \in \mathbb{N}$
repeat
  Let $\eta_j = A_j^* \nabla \mathrm{F}_j(A_j(e_{t_j})_\sharp \sigma^n) \in \mathscr{C}(\overline{\Omega})$ for $j = 0, \ldots, T$,
    $\widetilde{\mathrm{E}}((h,\xi)) = \sum_{j=0}^T \eta_j(\xi_j) h_j + \sum_{j=1}^T \mathrm{step}_j((h_{j-1}, \xi_{j-1}), (h_j, \xi_j))$
  Choose $\Lambda_j \subset [0,1] \times \overline{\Omega}$, $j = 0, \ldots T$ ▷ the discrete mesh
  Compute $Y^*[y, T] \in \prod_{j=0}^T \Lambda_j$ for all $y \in \Lambda_T$, as in (7.8) ▷ discrete optimal paths
  Choose $\widetilde{\gamma}^1, \ldots, \widetilde{\gamma}^k \in \mathrm{Image}(Y^*[\cdot, T])$ with least energy in $\widetilde{\mathrm{E}}$ ▷ select best k endpoints
  Find $\gamma^1, \ldots \gamma^k \in [0,1] \times \overline{\Omega}$ with $\widetilde{\mathrm{E}}(\gamma^i) \leq \widetilde{\mathrm{E}}(\widetilde{\gamma}^i)$ ▷ sliding step on linearised problem
  Set $\sigma \leftarrow \sigma^n$, re-order index i such that $\widetilde{\mathrm{E}}(\gamma^1) \leq \widetilde{\mathrm{E}}(\gamma^2), \ldots$
  for $i = 1, \ldots, k$ **do**
    $\lambda \leftarrow \mathrm{argmin}_{\lambda \in [0,1]} \mathrm{E}((1-\lambda)\sigma + \lambda \varphi(\gamma^i)^{-1} \delta_{\gamma^i})$ ▷ exact line search
    $\sigma \leftarrow (1-\lambda)\sigma + \lambda \varphi(\gamma^i)^{-1} \delta_{\gamma^i}$
  end for
  Choose σ^{n+1} such that $\mathrm{E}(\sigma^{n+1}) \leq \mathrm{E}(\sigma)$ ▷ sliding step on exact problem
  $n \leftarrow n + 1$
until converged

8.1 Benamou-Brenier example

First we compare directly with the numerical results presented in Bredies et al. [2020b] for the model discussed in Section 3. In particular, the energy we seek to minimise is $\mathrm{E}\colon \mathcal{M}^+(\Gamma) \to]-\infty, +\infty]$ defined by

$$\mathrm{E}(\sigma) = \frac{1}{2} \sum_{j=0}^T \left\| A_j(e_{t_j})_\sharp \sigma - b_j \right\|_2^2 + \int_\Gamma \int_0^1 \left(\alpha + \frac{\beta}{2} |\xi'(t)|^2 \right) \mathrm{d}t\, \mathrm{d}\sigma(h, \xi) \quad (8.5)$$

where $t_j = \frac{j}{T}$ for each $j = 0, \ldots, T$,

$$\Gamma \stackrel{\mathrm{def.}}{=} \left\{ (h, \xi) \in \{1\} \times \mathrm{AC}^2([0,1]; \overline{\Omega}) \mid \xi' \text{ is constant on }]t_{j-1}, t_j[,\ 1 \leq j \leq T \right\}, \quad (8.6)$$

and $A_j\colon \mathcal{M}([0,1]^2) \to \mathbb{R}^m$ represents a finite number of smoothed Fourier samples. The precise details are given in [Bredies et al., 2020b, Sec. 6]. In view of the convexity of Ω, (8.5) reformulates as

$$\mathrm{E}(\sigma) = \frac{1}{2} \sum_{j=0}^T \left\| A_j(e_{t_j})_\sharp \sigma - b_j \right\|_2^2 + \int_\Gamma \left(\sum_{j=1}^T (t_j - t_{j-1}) \left(\alpha + \frac{\beta}{2} \frac{|\xi(t_j) - \xi(t_{j-1})|^2}{(t_j - t_{j-1})^2} \right) \right) \mathrm{d}\sigma(h, \xi). \quad (8.7)$$

The two phantoms are also from [Bredies et al., 2020b, Sec. 6]. In the notation of this work, we would say that, for example, phantom 1 is represented by

$$\sigma_1^B = \delta_{(h^1, \xi^1)} + \delta_{(h^2, \xi^2)} \quad \text{where} \quad h^1 = h^2 = 1,\ \xi^1(t) = (0.2 + 0.6t,\ 0.2 + 0.6t), \quad (8.8)$$

$$\xi^2(t) = (0.8 - 0.6t,\ 0.2 + 0.6t). \quad (8.9)$$

Similarly phantom 2 is the sum of 3 Dirac masses, both phantoms are shown here in Figure 1. In the setting of Section 5.3, we choose $\varphi_0 = 0.1$ (as 10 is much larger than 2 or 3 which is the mass of phantoms 1 and 2 respectively), and

$$\mathrm{step}_j((h_0,\xi_0),(h_1,\xi_1)) = \alpha(t_j - t_{j-1}) + \frac{\beta}{2}\frac{|\xi_1-\xi_0|^2}{t_j - t_{j-1}} = \frac{\alpha}{T} + \frac{\beta T}{2}|\xi_1-\xi_0|^2. \tag{8.10}$$

For phantom 1 we have $\alpha = \beta = 0.5$, $T = 21$, and $\alpha = \beta = 0.1$, $T = 51$ for phantom 2. The algorithm of Bredies et al. [2020b] found online[2] is run with original parameters as a baseline, although with a time limit of 5 days when necessary. This is compared with multiple variants of the random and uniform algorithms described at the beginning of the section. The uniform algorithms are run to convergence and the random variants are run for 100 and 10,000 iterations for phantoms 1 and 2 respectively. Figure 2a shows the reconstructions of each algorithm with default parameters, each image is visually equivalent. The convergence behaviour is shown in Figure 2b. The energy plots confirm that each reconstruction has approximately the same energy, although we find that the random mesh algorithm finds the lowest energy, closely followed by the uniform mesh. Similarly, the sparsity of each final reconstruction is equal. The greatest difference between algorithms is run-time. The random and uniform mesh algorithms are over 100 times faster than that of Bredies et al. [2020b] in both examples.

We also replicate the noise-scenarios for phantom 2 as tested in Bredies et al. [2020b]. Our only modification of the baseline algorithm is to remove the early-stopping routine and run the algorithm for the minimum of 21 iterations or 5 days (c.f. [Bredies et al., 2020b, Table 1]). We again use the $(1, 25, 0)$-random mesh and $(1, 256, 0)$-uniform mesh algorithms for comparison in each example. In the three noisy scenarios we add 20%, 60%, and 60% Gaussian white noise to the data respectively. The first two scenarios use $\alpha = \beta = 0.1$ while the third scenario uses $\alpha = \beta = 0.3$. Seen in Figure 3, both the random and uniform mesh algorithms converge with similar rates, while the algorithm of Bredies et. al. is still at least 100 times slower. We see the expected behaviour that the uniform variant converges to a (possibly non-optimal) energy. As predicted by Theorem 5.3, the random variant often finds an even lower energy, despite having a much smaller value of N.

8.2 Wasserstein-Fisher-Rao example

In this section we show numerical results for the unbalanced transport example presented in Section 4, the data fidelity is the same as in the Benamou-Brenier example. Ideally we would use the exact function $d_{\alpha,\beta,\delta}$ from (4.13) to define step_j, but it lacks a closed-form expression, hence for computational reasons we use the approximation

$$\mathrm{step}_j(\gamma_0,\gamma_1) \overset{\text{def.}}{=} \alpha\frac{h_0+h_1}{2}(t_j - t_{j-1}) + d_{0,\beta,\delta}(\gamma_0, t_{j-1}, \gamma_1, t_j). \tag{8.11}$$

where $d_{0,\beta,\delta}$ is given in (4.15).

Our synthetic phantoms are equivalent to those in Section 8.1 but with modified, time-dependent masses. For example, the first phantom is now $\sigma = \delta_{(h^0,\xi^0)} + \delta_{(h^1,\xi^1)}$ where

$$h^0(t) = \tfrac{1}{2}(1+3t^2), \qquad \xi^0(t) = (0.2+0.6t,\ 0.2+0.6t), \tag{8.12}$$

$$h^1(t) = \tfrac{3}{2}\sqrt{1-t}, \qquad \xi^1(t) = (0.8-0.6t,\ 0.2+0.6t). \tag{8.13}$$

The transformation for the second phantom is very similar and the precise formulae can be found in the supplementary code. All curves h have been normalised so $\int_0^1 h\,\mathrm{d}t = 1$, $\|h\|_\infty \leq 2$.

As in Section 8.1 we use $\alpha = \beta = 0.5$ or 0.1 for the first and second phantom respectively, $\varphi = \delta = 0.1$ throughout. This leads to the explicit form of step_j:

$$\mathrm{step}_j(\gamma_0,\gamma_1) = \frac{\alpha}{T}\frac{h_0+h_1}{2} + \frac{\beta T}{25}\left[\frac{h_0+h_1}{2} - \sqrt{h_0 h_1}\cos\left(\min\left(5|\xi_0-\xi_1|,\pi\right)\right)\right]. \tag{8.14}$$

[2] https://github.com/panchoop/DGCG_algorithm/commit/553a564fd8641abcfac6067ebf51a900a6a91d0f

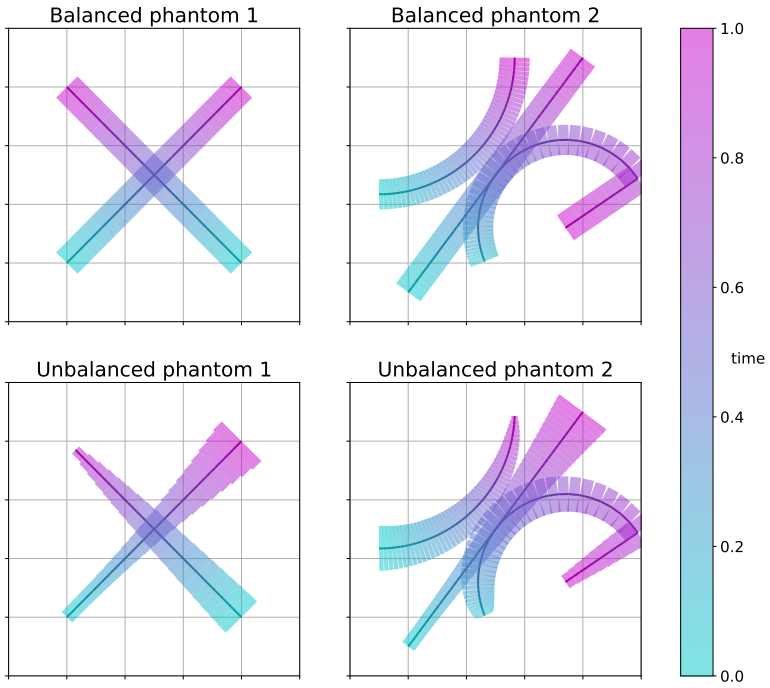


Figure 1: Synthetic phantoms of the form $\sigma = \sum_i \delta_{(h^i, \xi^i)}$ for balanced (top row) and unbalanced (bottom row) examples. Colour indicates the time t, the solid line indicates the positions $\xi^i(t)$, and the width of the overlayed band is proportional to $h^i(t)$.

Numerically we re-parametrise the mass to $\widetilde{h}_j \overset{\text{def.}}{=} \sqrt{h_j}$ for each j so that step_j is a C^1 function of $\widetilde{h}_j$. Note that this only effects the sliding step of the Frank-Wolfe algorithm, the remaining steps are unchanged.

Again, we run the uniform-mesh algorithms to convergence but now the random-mesh is only run for 100 or 1,000 iterations for phantoms 1 and 2 respectively. The reconstructions are shown in Figure 4a with corresponding convergence plots in Figure 4b.

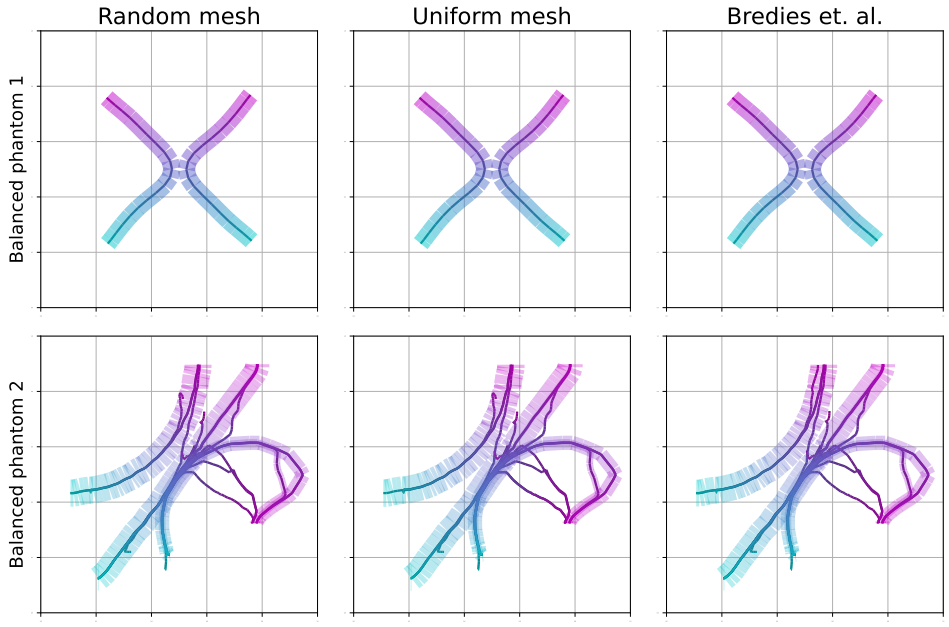
8.3 Observations on parameter choices

Both the random and uniform mesh algorithms have three parameters to choose: the multi-start parameter k, spatial resolution N, and mass resolution M. The first phantom (balanced or unbalanced) nicely highlights characteristics of the reconstructions but is too trivial numerically to compare different algorithm choices, all methods converge within a few iterations. For the second phantom we prioritised the trade-off of between energy and computation time.

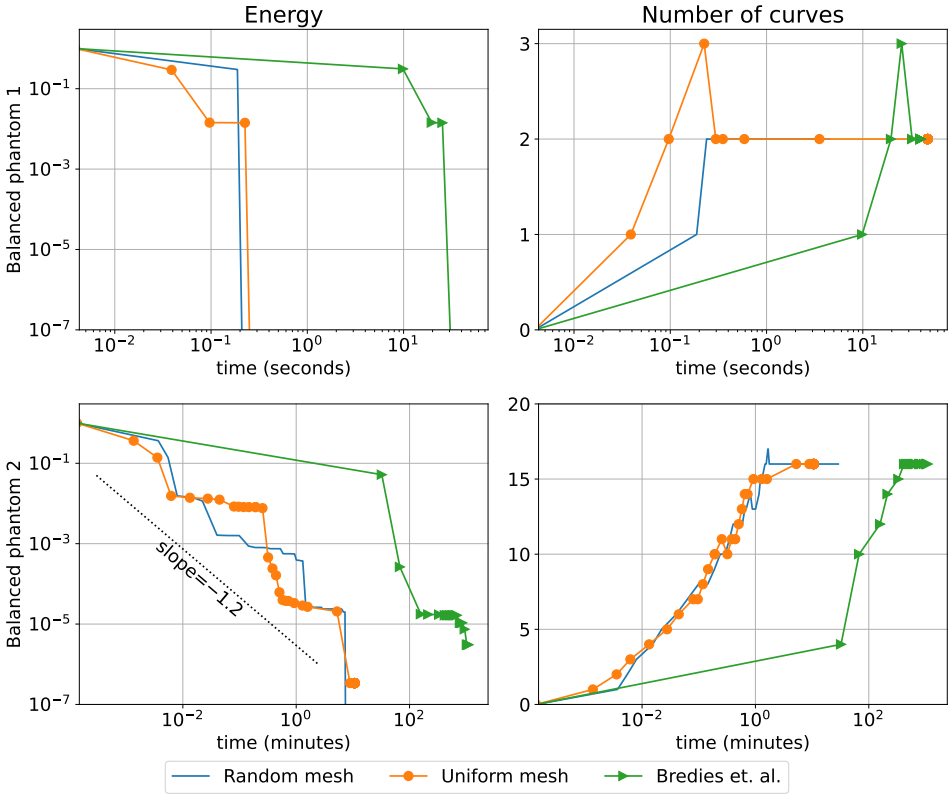
The classical choice of $k = 1$ was always a competitive choice. Large k potentially enables the algorithm to find multiple atoms in one iteration, but slows down the sliding steps. We found that $k \in [1, 5]$ was a reasonable range depending on the sparsity of the signal to be recovered.

Choice of spatial resolution made the largest impact on performance. If the resolution is too small then the algorithm will not find new atoms, but computation time of the linear oracle scales with N^4. For the random mesh, we found that $N = 25$ was a good balance. For the uniform mesh the resolution is also a stopping criterion so we used the more conservative $N = 256$ and 512 for the balanced and unbalanced examples respectively.

In the unbalanced experiments the value of M had little effect. Surprisingly, the value of $M = 0$ was also competitive for our unbalanced phantom 2, achieving energies within $10^{-3}\%$ of the best observed energy. It's unclear whether this will generalise to more complicated



(a) Final reconstructions visualised equivalently to those in Figure 1.



(b) Convergence of energy and sparsity of reconstructions. For each phantom, every energy is translated by the smallest energy found by any method.

Figure 2: Comparison of algorithm from Bredies et al. [2020b], the $(1, 25, 0)$-random mesh algorithm, and $(1, 256, 0)$-uniform mesh algorithm applied to balanced phantoms 1 and 2 (first and second rows respectively).

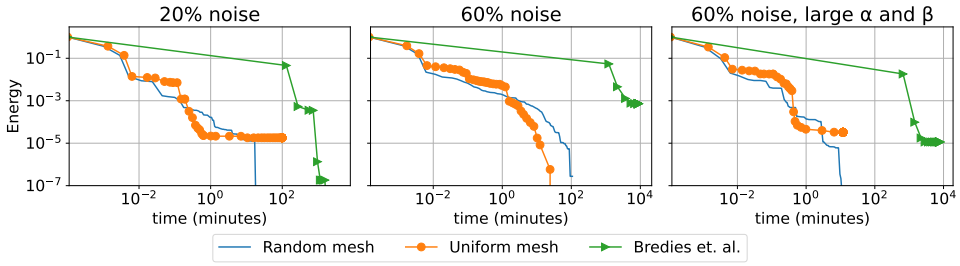


Figure 3: Convergence of three default algorithms (see Figure 2) with different levels of noise and choice of $\alpha = \beta \in \{0.1, 0.3\}$.

examples, we chose $M = 10$ as a default. Computational complexity of the linear oracle also scales with M^2 so it should not be too large.

Both the random and uniform mesh algorithms have advantages over each other. The main advantages of a uniform mesh are computational: one can use a finer resolution (i.e. larger N), and there is a clear stopping criterion. In practice this was a very reliable method without parameter tuning and was as fast as the random algorithm. The random algorithms have analytical guarantees of converging asymptotically to the true minimiser, and indeed it achieved the best observed energy in almost of our experiments, performing noticeably better in the noisy case. The challenge is setting a stopping criterion, in phantom 2 of Figure 4b one sees several plateaux where the algorithm fails to find a descent direction for a number of iterations before continuing to descend.

Figure 5 shows the random variation of the random mesh algorithm with different values of $N \in \{5, 10, 15\}$. Increasing N both decreases the spread and improves the performance of the algorithm. In the balanced example, the three algorithms complete 1000 iterations in the same time, showing that the linear oracle step is a small part of the total time. On the other hand, with larger M the $N = 25$ algorithm is nearly 10 times slower than $N = 10$, indicating that the $O(N^4 M^2)$ cost is starting to dominate the computation time. It's possible that $N = 5$ is slower that $N = 10$ for the unbalanced example because the sliding step has to work much harder to find local minima.

9 Conclusion

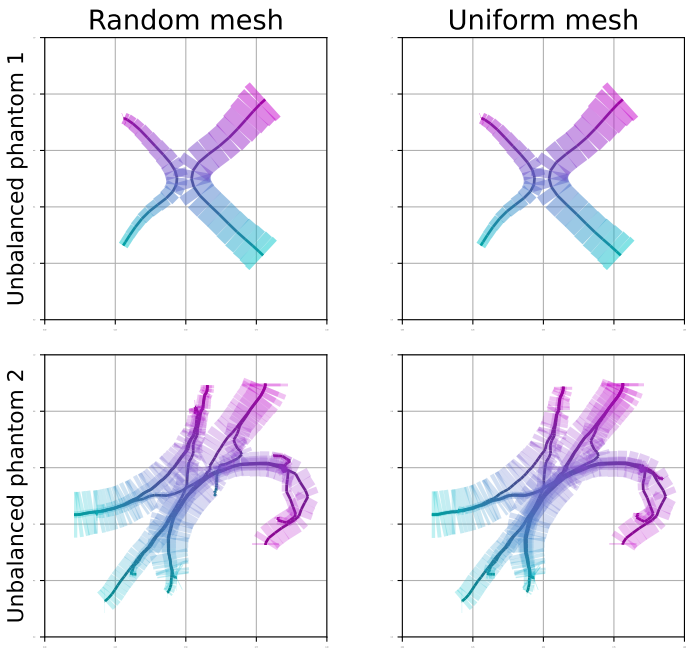
The main contribution of this work was to extend a variational model proposed by Bredies et. al. in order to accelerate the reconstruction of sparse measures in dynamical inverse problems. Using algorithms developed for computing shortest paths on graphs, we can improve the speed by a factor of 100 while still finding lower energy solutions. This allows us to process new unbalanced examples where the mass of curves is not constant in time, our proposed algorithms still recover good reconstructions in a reasonable amount of time.

We also presented new analysis of a stochastic variant of Frank-Wolfe which guarantees the convergence of our algorithm (in energy) to a globally optimal solution. This is supported by our experiments where we see that the random-mesh algorithm achieves the lowest energy in almost every example.

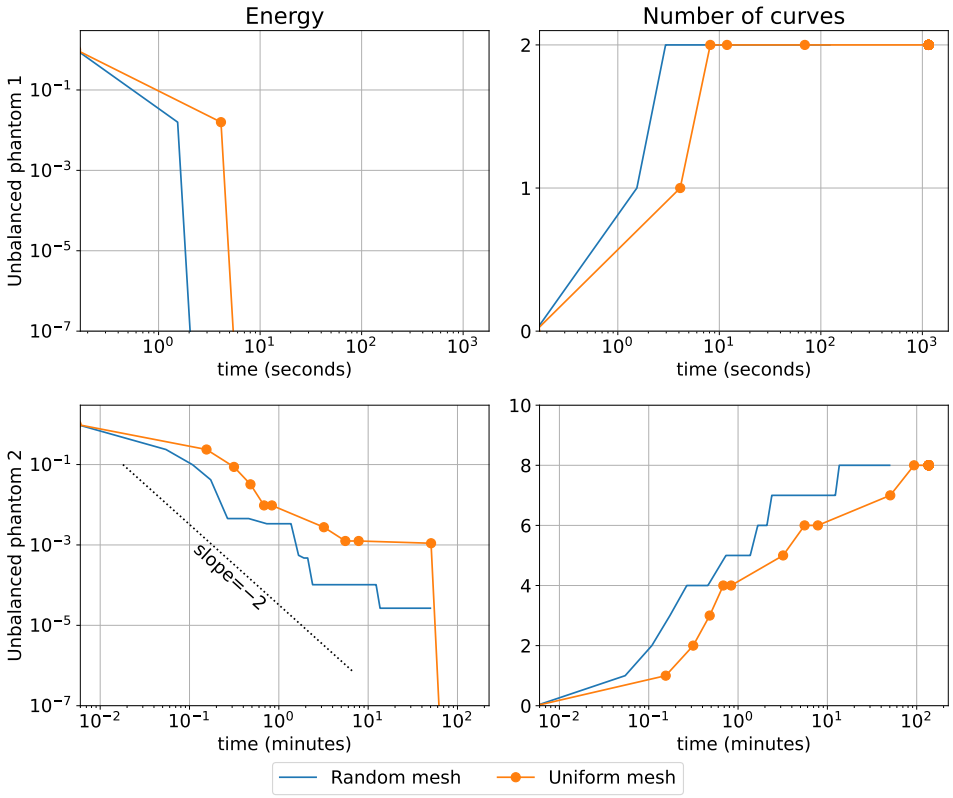
One feature of the algorithm in Bredies et al. [2020b] which we did not take advantage of was the idea of importance sampling. We chose points $y \in \Lambda_j$ uniformly randomly, whereas it is likely to be beneficial to choose y such that $\eta_j(y)$ is small. If the step functions step_j satisfy the triangle inequality, then this bias could be implemented by choosing points y such that y is a local minima of

$$\widetilde{y} \mapsto \eta_j(\widetilde{y}) + \text{step}_{j-1}(y, \widetilde{y}) + \text{step}_j(\widetilde{y}, y). \tag{9.1}$$

In Algorithm 3, implementing such a sliding step on Λ_j between lines 4 and 5 would preserve the analytical properties of the current algorithm, whilst possibly improving practical performance.



(a) Final reconstructions visualised equivalently to those in Figure 1.



(b) Convergence of energy and sparsity of reconstructions. For each phantom, the energies are translated by the smallest energy found by either method.

Figure 4: Comparison of the $(1, 25, 10)$-random mesh and $(1, 128, 10)$-uniform mesh algorithms applied to unbalanced phantoms 1 and 2 (first and second rows respectively).

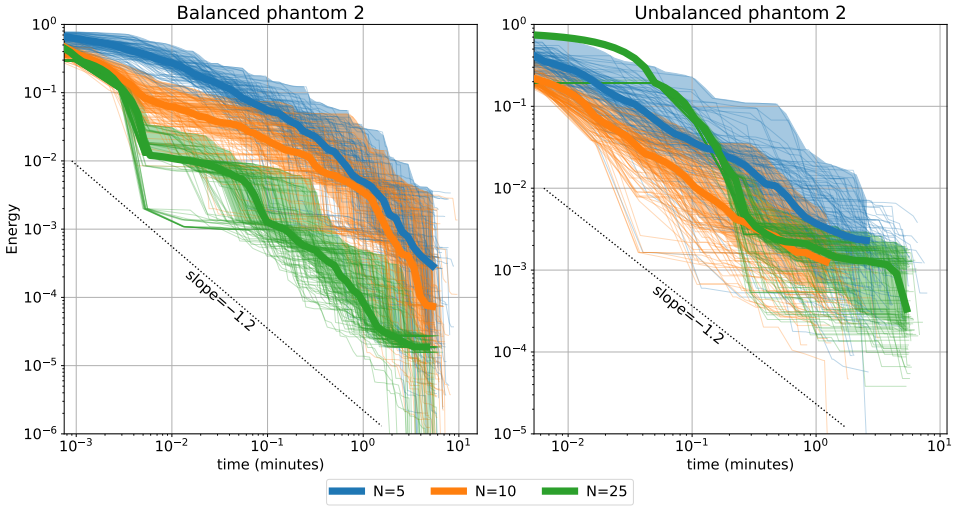


Figure 5: Convergence of different random realisations of the random mesh algorithms applied to the balanced and unbalanced phantom 2. The two algorithms are $(1, N, 0)$- and $(1, N, 10)$-random mesh run for 1000 and 100 iterations respectively. 100 random instances of each algorithm are run, each drawn as a thin line. The median is drawn with a thick line, and the inter-quartile range indicated by the shaded area.

However, it is also possible that the sliding step already present (e.g. line 7) is powerful enough to find these improved mesh points after computing $\widetilde{\gamma}^i$, without the extra help beforehand. It is likely that the benefits depend on the application, particularly on the smoothness of the operators A_j.

10 Acknowledgements

This work was supported by the ANR CIPRESSI project, grant ANR-19-CE48-0017-01 of the French Agence Nationale de la Recherche.

References

G. S. Alberti, H. Ammari, F. Romero, and T. Wintz. Dynamic spike superresolution and applications to ultrafast ultrasound imaging. *SIAM Journal on Imaging Sciences*, 12(3): 1501–1527, 2019.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser Basel, second edition, 2008.

J.-M. Azaïs, Y. de Castro, and F. Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, Mar. 2015. ISSN 10635203. doi: 10.1016/j.acha.2014.03.004.

V. I. Bogachev. *Measure Theory*. Springer-Verlag Berlin Heidelberg, 2007.

N. Boyd, G. Schiebinger, and B. Recht. The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems. *SIAM Journal on Optimization*, 27(2):616–639, Jan. 2017. ISSN 1052-6234. doi: 10.1137/15M1035793.

C. Boyer, A. Chambolle, Y. D. Castro, V. Duval, F. De Gournay, and P. Weiss. On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2):1260–1281, 2019.

K. Bredies and M. Carioni. Sparsity of solutions for variational inverse problems with finite-dimensional data. *Calculus of Variations and Partial Differential Equations*, 59(1):14, Dec. 2019. ISSN 1432-0835. doi: 10.1007/s00526-019-1658-1.

K. Bredies and S. Fanzon. An optimal transport approach for solving dynamic inverse problems in spaces of measures. *ESAIM: M2AN*, 54(6):2351–2382, Nov. 2020. ISSN 0764-583X, 1290-3841. doi: 10.1051/m2an/2020056.

K. Bredies and H. K. Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, Jan. 2013. ISSN 1292-8119, 1262-3377. doi: 10.1051/cocv/2011205.

K. Bredies, M. Carioni, and S. Fanzon. A superposition principle for the inhomogeneous continuity equation with Hellinger-Kantorovich-regular coefficients. *arXiv preprint arXiv:2007.06964*, 2020a.

K. Bredies, M. Carioni, S. Fanzon, and F. Romero. A generalized conditional gradient method for dynamic inverse problems with optimal transport regularization. *arXiv preprint arXiv:2012.11706*, 2020b.

K. Bredies, M. Carioni, S. Fanzon, and F. Romero. On the extremal points of the ball of the Benamou–Brenier energy. *Bulletin of the London Mathematical Society*, 53(5):1436–1452, 2021. ISSN 1469-2120. doi: 10.1112/blms.12509.

H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer New York, New York, NY, 2011. ISBN 978-0-387-70913-0 978-0-387-70914-7. doi: 10.1007/978-0-387-70914-7.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018a.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and Fisher-Rao metrics. *Foundations of Computational Mathematics*, 18 (1):1–44, 2018b.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT press, third edition, 2009.

Q. Denoyelle, V. Duval, G. Peyre, and E. Soubies. The Sliding Frank-Wolfe Algorithm and its application to super-resolution microscopy. *Inverse Problems*, 2019. ISSN 0266-5611. doi: 10.1088/1361-6420/ab2a29.

L. Ding, J. Fan, and M. Udell. kFW: A Frank-Wolfe style algorithm with stronger subproblem oracles. *arXiv preprint arXiv:2006.16142*, 2020.

L. E. Dubins. On extreme points of convex sets. *Journal of Mathematical Analysis and Applications*, 5(2):237–244, Oct. 1962. ISSN 0022247X. doi: 10.1016/S0022-247X(62)80007-9.

V. Duval and G. Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

H. Federer. *Geometric measure theory*. Classics in mathematics. Springer, Berlin ; New York, 1996. ISBN 978-3-540-60656-7.

A. Flinth, F. de Gournay, and P. Weiss. On the linear convergence rates of exchange and continuous methods for total variation minimization. *Mathematical Programming*, 190(1-2): 221–257, Nov. 2021. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-020-01530-0.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

C. Poon, N. Keriven, and G. Peyré. The geometry of off-the-grid compressed sensing. *Foundations of Computational Mathematics*, Oct. 2021. doi: https://doi.org/10.1007/s10208-021-09545-5.

R. T. Rockafellar. *Conjugate duality and optimization*. SIAM, 1974.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015. ISBN 978-3-319-20828-2.

A. Silveti-Falls, C. Molinari, and J. Fadili. Inexact and stochastic generalized conditional gradient with augmented lagrangian and proximal step. *arXiv preprint arXiv:2005.05158*, 2020.

M. Unser, J. Fageot, and J. P. Ward. Splines are Universal Solutions of linear inverse problems with generalized-tv regularization. *SIAM Review*, 59(4):769–793, Jan. 2017. ISSN 0036-1445, 1095-7200. doi: 10.1137/16M1061199.

K. Yosida. *Functional analysis*. Springer Berlin Heidelberg, sixth edition, 1980.

A Results for the structure of E

Here we prove Theorem 5.1 as a result of a sequence of simple lemmas. We start with results for the linear term W. Throughout this section we fix lower semi-continuous $w\colon \Gamma \to [0,+\infty]$ for a complete separable metric space Γ and define $\mathrm{W}\colon \mathcal{M}^+(\Gamma) \to [0,+\infty]$ by

$$\forall \sigma \in \mathcal{M}^+(\Gamma), \qquad \mathrm{W}(\sigma) \overset{\text{def.}}{=} \int_\Gamma w(\gamma)\,\mathrm{d}\sigma(\gamma). \tag{A.1}$$

A.1 Properties of W

The following lower semi-continuity result is well known (see for instance [Santambrogio, 2015, Prop. 7.1], or [Ambrosio et al., 2008, Prop. 5.1.7]).

Lemma A.1 (Fatou's lemma for measures)**.** *If $w\colon \Gamma \to [0,+\infty]$ is lower semi-continuous, then the functional* W *is lower semi-continuous on $\mathcal{M}^+(\Gamma)$ with respect to the narrow topology.*

Compactness in spaces of measure can be characterised by the following lemmas.

Lemma A.2 (Prokhorov's theorem on measure spaces)**.** *If Γ is a complete separable metric space and $\mathcal{A} \subset \mathcal{M}(\Gamma)$ is a family of Borel measures, then*

$$\mathcal{A} \text{ is relatively compact in the narrow topology} \quad \text{if and only if} \quad \mathcal{A} \text{ is tight and norm-bounded.} \tag{A.2}$$

The proof of this is [Bogachev, 2007, Thms. 7.1.7, 8.6.7-8]. Because of this lemma, we can give sufficient conditions for the compactness of sub-levelsets of W.

Lemma A.3. *Let D be a closed subset of $\mathcal{M}^+(\Gamma)$ and denote $U_t = \{\sigma \in D \text{ s.t. } \mathrm{W}(\sigma) \leq t\}$ for $t \in \mathbb{R}$. If $w\colon \Gamma \to [0,+\infty]$ is lower semi-continuous, w has compact sub-levelsets, and U_t is bounded in norm for each t, then $\mathrm{W}|_D$ has compact sub-levelsets in the narrow topology of $\mathcal{M}^+(\Gamma)$.*

Proof. As $w \geq 0$, Lemma A.1 shows that W is lower semi-continuous, therefore U_t is closed. As U_t is bounded, Lemma A.2 equates compactness with tightness. Finally, [Bredies et al., 2020a, Prop. A.2] (adapted from [Ambrosio et al., 2008, Rem. 5.1.5]) states:

> If Γ is a complete separable metric space and w has compact sub-levelsets, then U_t is tight for each $t \in \mathbb{R}$.

We conclude that U_t is narrowly compact for each $t \in \mathbb{R}$. □

We now confirm that the results of this section are applicable to the examples of Sections 3-4.

Lemma A.4 (Lower semi-continuity and coercivity of optimal transport regularisations)**.** *Choose $\alpha, \beta, \delta > 0$, Γ a closed subset of Γ_∞, and let $w : \Gamma \to]0, +\infty]$ be defined by*

$$\forall (h,\xi) \in \Gamma, \qquad w(\gamma) \stackrel{\text{def.}}{=} \begin{cases} \int_{h>0} \left[\alpha + \frac{\beta}{2}|\xi'|^2 + \frac{\beta\delta^2}{2}\left(\frac{h'}{h}\right)^2\right] h \,\mathrm{d}t & \sqrt{h}, \sqrt{h}\xi \in \mathrm{AC}^2 \\ +\infty & \textit{otherwise.} \end{cases} \tag{A.3}$$

Then w is lower semi-continuous and has compact sub-levelsets (in the metric d_Γ).

Proof. The spaces AC^2 *and* H^1*.* First we note, that $\mathrm{AC}^2([0,1]\,;\mathbb{R}^d)$ is exactly the collection of the continuous representatives of functions in $H^1([0,1]\,;\mathbb{R}^d)$. Moreover, the weak (distributional) gradient is represented by the (almost everywhere defined) pointwise derivative. This is alluded to in [Bredies et al., 2020b, Sec. 2.3], but we provide the proof for the sake of completeness.

Assume that $\xi \in \mathrm{AC}^2([0,1]\,;\overline{\Omega})$. Since ξ is absolutely continuous, for all $r \in \mathscr{C}_c^\infty(]0,1[\,;\mathbb{R}^d)$, we have (see [Federer, 1996, 2.5.18 and 2.9.21])

$$\int_0^1 \xi'(t) \cdot r(t) \,\mathrm{d}t = -\int_0^1 \xi(t) \cdot r'(t) \,\mathrm{d}t.$$

As a result, ξ has a weak derivative which is given by the square integrable function $t \mapsto \xi'(t)$ (where $\xi'(t)$ is the pointwise derivative, defined for a.e. t). Moreover $\xi \in \mathscr{C}([0,1]\,;\overline{\Omega}) \subseteq L^2([0,1]\,;\mathbb{R}^d)$, hence $\xi \in H^1([0,1]\,;\mathbb{R}^d)$.

Conversely, if $\xi \in H^1(0,1;\mathbb{R}^d)$, then [Brezis, 2011, Thm. 8.2],

$$\forall t, \tau \in [0,1]\,, \quad \xi(t) - \xi(\tau) = \int_\tau^t D\xi(t) \,\mathrm{d}t.$$

where $D\xi \in L^2([0,1]\,;\mathbb{R}^d)$ is the weak derivative of ξ. As a result, $\xi \in \mathrm{AC}^2([0,1]\,;\mathbb{R}^d)$.

Lower semi-continuity and coercivity. Define the sub-levelsets of w by

$$\forall \lambda \in \mathbb{R}, \qquad U_\lambda \stackrel{\text{def.}}{=} \{\, \gamma \in \Gamma \mid w(\gamma) \leq \lambda \,\}\,. \tag{A.4}$$

The function w is lower semi-continuous and coercive if and only if each U_λ is closed and compact. As Γ is a metric space, it is sufficient to show that for every $\lambda \in \mathbb{R}$ and sequence $\gamma_n = (h_n, \xi_n) \in U_\lambda$, there exists a sub-sequence (still referred to as γ_n for convenience) and limit point $\gamma = (h,\xi) \in U_\lambda$ such that $\gamma_n \stackrel{\mathrm{d}_\Gamma}{\to} \gamma$.

Existence of h. First note that, if $\sqrt{h_n} \in \mathrm{AC}^2([0,1])$, then

$$\int_0^1 (\sqrt{h_n(t)})^2 + ((\sqrt{h_n(t)})')^2 \,\mathrm{d}t = \int_{h>0} h_n(t)^2 + \frac{h_n'(t)^2}{4h_n(t)} \,\mathrm{d}t \leq \left[\frac{1}{\alpha} + \frac{1}{2\beta\delta^2}\right] w(\gamma_n). \tag{A.5}$$

We conclude that $\{\sqrt{h_n}\}_{n\in\mathbb{N}}$ are contained in a bounded set in $H^1([0,1])$, therefore there is a sub-sequence which converges to a weak limit $f \in H^1([0,1])$. Passing to this sub-sequence, by the Rellich-Kondrachov theorem we have

$$\sqrt{h_n} \to f \text{ in } \mathscr{C}([0,1]) \quad \text{and} \quad (\sqrt{h_n})' \rightharpoonup f' \text{ weakly in } L^2([0,1]). \tag{A.6}$$

Without loss of generality assume $\left\|\sqrt{h_n}\right\|_\infty \leq 2\|f\|_\infty$ for all $n \in \mathbb{N}$. Now, we let $h \stackrel{\text{def.}}{=} f^2$, and we observe that

$$\sup_{t\in[0,1]} |h_n(t) - f(t)^2| = \sup_{t\in[0,1]} |\sqrt{h_n(t)} - f(t)||\sqrt{h_n(t)} + f(t)| \tag{A.7}$$

$$\leq 3\left\|\sqrt{h}\right\|_\infty \left\|\sqrt{h_n} - f\right\|_\infty. \tag{A.8}$$

We conclude that $h_n \to h$ in $\mathscr{C}([0,1])$, $\sqrt{h} = f \in \mathrm{AC}^2([0,1])$. Moreover, by the lower semi-continuity of the L^2 norm w.r.t. weak convergence,

$$\int_0^1 \left[\alpha(\sqrt{h(t)})^2 + 2\beta\delta^2((\sqrt{h})'(t))^2\right] \mathrm{d}t \leq \liminf_{n\to+\infty} \int_0^1 \left[\alpha(\sqrt{h_n(t)})^2 + 2\beta\delta^2((\sqrt{h_n})'(t))^2\right] \mathrm{d}t. \tag{A.9}$$

Existence of ξ. Similarly to the case with h, we will show that the H^1 norm of $\sqrt{h_n}\xi_n$ is uniformly bounded. Recall that $h_n \to h$ in $\mathscr{C}([0,1])$ and Ω is bounded, say $|x| \leq R$ for all $x \in \Omega$. Therefore both $\left\|\sqrt{h_n}\right\|_\infty$ and $\|\xi_n\|_\infty$ are uniformly bounded, so $\left\|\sqrt{h_n}\xi_n\right\|_2$ is also uniformly bounded. For the derivative, observe

$$\int_0^1 |(\sqrt{h_n}\xi_n)'(t)|^2 = \int_{h_n>0} |\sqrt{h_n(t)}\xi_n'(t) - (\sqrt{h_n})'(t)\xi_n(t)|^2 \,\mathrm{d}t \tag{A.10}$$

$$\leq 2\int_{h_n>0} |\xi_n'(t)|^2 h_n(t) + |(\sqrt{h_n})'(t)\xi_n(t)|^2 \,\mathrm{d}t \tag{A.11}$$

$$\leq \left[\frac{4}{\beta} + \frac{R^2}{\beta\delta^2}\right] w(\gamma_n), \tag{A.12}$$

which is uniformly bounded as $w(\gamma_n) \leq \lambda$. We conclude that there is a limit point $g \in \mathrm{AC}^2([0,1]\,;\mathbb{R}^d)$ such that, after passing to a sub-sequence,

$$\sqrt{h_n}\xi_n \to g \text{ in } \mathscr{C}([0,1]\,;\mathbb{R}^d) \quad \text{and} \quad (\sqrt{h_n}\xi_n)' \rightharpoonup g' \text{ in } L^2([0,1]\,;\mathbb{R}^d). \tag{A.13}$$

To conclude, we need to show that $\xi \overset{\text{def.}}{=} \frac{g}{\sqrt{h}}$ is well-defined. Set

$$\forall h(t) > 0, \qquad \xi(t) \overset{\text{def.}}{=} \frac{g(t)}{\sqrt{h(t)}}. \tag{A.14}$$

Then by a limit argument, $\xi(t) \in \overline{\Omega}$. Moreover, being the ratio of a.e. differentiable functions, ξ is differentiable for a.e. t such that $h(t) > 0$, with

$$\xi'(t) = \frac{g'(t) - (\sqrt{h})'(t)\xi(t)}{\sqrt{h(t)}}. \tag{A.15}$$

We may extend ξ by an arbitrary constant $x_0 \in \overline{\Omega}$ to obtain a function $\xi\colon\ [0,1] \to \overline{\Omega}$ continuous on $\{h > 0\}$.

Now, let $\varepsilon > 0$, and let us prove that ξ_n converges uniformly to ξ on $\{h \geq \varepsilon\}$. There exists $N_\varepsilon \in \mathbb{N}$ such that $\|h_n - h\|_\infty \leq \frac{\varepsilon}{2}$, so we have

$$\forall n \geq N_\varepsilon, \qquad \sup_{h(t)\geq\varepsilon} |\xi_n(t) - \xi(t)| = \sup_{h(t)\geq\varepsilon} \left|\frac{\sqrt{h_n(t)}\xi_n(t)}{\sqrt{h_n(t)}} - \frac{g(t)}{\sqrt{h(t)}}\right| \tag{A.16}$$

$$\leq \sup_{h(t)\geq\varepsilon} \frac{|\sqrt{h_n(t)}\xi_n(t) - g(t)|}{\sqrt{h_n(t)}} + \frac{|g(t)||\sqrt{h_n(t)} - \sqrt{h(t)}|}{\sqrt{h_n(t)h(t)}} \tag{A.17}$$

$$\leq \sqrt{\frac{2}{\varepsilon}}\left\|\sqrt{h_n}\xi_n - g\right\|_\infty + \frac{\sqrt{2}\|g\|_\infty}{\varepsilon}\left\|\sqrt{h_n} - f\right\|_\infty. \tag{A.18}$$

As a result, $\xi_n \to \xi$ uniformly in $\{h \geq \varepsilon\}$. Since this is true for all $\varepsilon > 0$ and we already know that $h_n \to h$ in $\mathscr{C}([0,1])$, we deduce from Lemma 2.1 that $\gamma_n \xrightarrow{\mathrm{d}_\Gamma} \gamma$, where $\gamma \overset{\text{def.}}{=} (h, \xi)$.

Conclusion. It remains to prove that $\gamma \in U_\lambda$, which ensures that U_λ is compact (hence that w is both lower semi-continuous and coercive). Since, $\xi_n \to \xi$ in $L^\infty(\{h \geq \varepsilon\}; \overline{\Omega})$, we have $(\sqrt{h_n})'\xi_n \rightharpoonup (\sqrt{h})'\xi$ in $L^2(\{h \geq \varepsilon\}; \mathbb{R}^d)$. Hence, by the monotone convergence theorem and

the weak lower semi-continuity of the L^2 norm,

$$\int_{h>0} |\xi'|^2 h \,\mathrm{d}t = \lim_{\varepsilon \searrow 0} \int_{h \geq \varepsilon} |g'(t) - (\sqrt{h})'(t)\xi(t)|^2 \,\mathrm{d}t \tag{A.19}$$

$$\leq \limsup_{\varepsilon \searrow 0} \liminf_{n \to +\infty} \int_{h \geq \varepsilon} |(\sqrt{h_n}\xi_n)'(t) - (\sqrt{h_n})'(t)\xi_n(t)|^2 \,\mathrm{d}t \tag{A.20}$$

$$\leq \limsup_{\varepsilon \searrow 0} \liminf_{n \to +\infty} \int_{h_n > 0} |\xi_n'|^2 h_n \,\mathrm{d}t \tag{A.21}$$

$$= \liminf_{n \to +\infty} \int_{h_n > 0} |\xi_n'|^2 h_n \,\mathrm{d}t. \tag{A.22}$$

To conclude, we know that $\sqrt{h}, \sqrt{h}\xi \in \mathrm{AC}^2$, and from (A.9) and (A.22), we obtain

$$w(\gamma) = \int_0^1 \left[\alpha(\sqrt{h(t)})^2 + 2\beta\delta^2((\sqrt{h})'(t))^2\right] \mathrm{d}t + \int_{h>0} \frac{\beta}{2} |\sqrt{h(t)}\xi'(t)|^2 \,\mathrm{d}t \leq \liminf_{n \to +\infty} w(\gamma_n) \leq \lambda.$$

Hence $\gamma \in U_\lambda$ and w is both coercive and lower semi-continuous. □

Note that the coercivity and lower semi-continuity of the Benamou-Brenier penalty follows immediately if we consider it to be the function

$$(h, \xi) \mapsto w(h, \xi) + \begin{cases} 0 & h = 1 \\ +\infty & \text{else.} \end{cases} \tag{A.23}$$

The constraint-set is closed, therefore the function is still coercive and lower semi-continuous.

A.2 Properties of D

We consider D of the form in (5.6), that is for some lower semi-continuous $\varphi \colon \Gamma \to]0, +\infty]$,

$$D \overset{\text{def.}}{=} \left\{ \sigma \in \mathcal{M}^+(\Gamma) \quad \text{s.t.} \quad \int_\Gamma \varphi \,\mathrm{d}\sigma \leq 1 \right\}. \tag{A.24}$$

The only difference between w and φ is that $\varphi > 0$, so we can re-use many of the results for W.

Lemma A.5. *If $\varphi \colon \Gamma \to]0, +\infty]$ is lower semi-continuous, then D is closed and*

$$\mathrm{Ext}\,(D) = \{0\} \cup \left\{ \varphi(\gamma)^{-1}\delta_\gamma \mid \varphi(\gamma) < +\infty \right\}. \tag{A.25}$$

Furthermore, if $\inf_{\gamma \in \Gamma} \varphi(\gamma) \geq \varepsilon > 0$, then D is bounded. Finally, if φ has compact sub-levelsets, then D is also compact.

Proof. Note by Lemma A.1 that D is a sub-levelset of the lower semi-continuous function $\sigma \mapsto \int_\Gamma \varphi \,\mathrm{d}\sigma$, therefore it is closed. Also for any $\sigma \in \mathcal{M}^+(\Gamma)$ with $\|\sigma\| > \varepsilon^{-1}$, we have $\int_\Gamma \varphi \,\mathrm{d}\sigma > 1$. In particular, $\sigma \notin D$, so D is bounded. If φ has compact sub-levelsets, taking $w = \varphi$ in Lemma A.3 confirms D is compact.

It remains to prove (A.25), and begin with the "$\supset$" inclusion. By non-negativity, note that for all $\sigma, \sigma_0, \sigma_1 \in D$ and $\lambda \in]0, 1[$,

$$\sigma = \lambda\sigma_0 + (1-\lambda)\sigma_1 \qquad \Longrightarrow \qquad \sigma_0, \sigma_1 \ll \sigma. \tag{A.26}$$

In particular, taking $\sigma = 0$, we obtain $\sigma_0 = \sigma_1 = 0$, hence $0 \in \mathrm{Ext}\,(D)$. Now, setting $\sigma = \varphi(\gamma)^{-1}\delta_\gamma$ for some $\gamma \in \Gamma$ with $\varphi(\gamma) < +\infty$, we deduce that

$$\sigma_0 = \frac{\alpha}{\varphi(\gamma)}\delta_\gamma, \ \sigma_1 = \frac{\beta}{\varphi(\gamma)}\delta_\gamma, \ \text{for some } \alpha, \beta \in [0, 1]. \tag{A.27}$$

Since $\int_\Gamma \varphi \,\mathrm{d}\sigma = 1$, we must have $\lambda\alpha + (1-\lambda)\beta = 1$ hence $\alpha = \beta = 1$, so that $\sigma_0 = \sigma_1 = \sigma$. As a result, $\varphi(\gamma)^{-1}\delta_\gamma$ is an extreme point of D.
To show that this inclusion is sharp, we make the following claim.

> Claim: If $\sigma \in D$ and there exists $\gamma_0, \gamma_1 \in \mathrm{supp}(\sigma)$ distinct, then $\sigma \in]\sigma_0, \sigma_1[$ for some $\sigma_0, \sigma_1 \in D$ such that $\gamma_0 \in \mathrm{supp}(\sigma_0) \setminus \mathrm{supp}(\sigma_1)$ and $\gamma_1 \in \mathrm{supp}(\sigma_1) \setminus \mathrm{supp}(\sigma_0)$.

Proof of claim: Let $r = \frac{1}{2}\,\mathrm{d}_\Gamma(\gamma_0, \gamma_1)$ and set $\Gamma_1 = \{\mathrm{d}_\Gamma(\gamma, \gamma_1) < r\}$. Define $\sigma_1 = \mathbb{1}_{\Gamma_1}\sigma$ and $\sigma_0 = \sigma - \sigma_1$. By the definition of support, for both $i = 0, 1$ we have $\sigma_i \neq 0$, $\gamma_i \in \operatorname{supp}(\sigma_i) \setminus \operatorname{supp}(\sigma_{1-i})$. Also, $\alpha\sigma_0 + \beta\sigma_1 \in \mathcal{M}^+(\Gamma)$ for all $\alpha, \beta \geq 0$, the only challenge is the constraint with φ.

Case $\int_\Gamma \varphi\,\mathrm{d}\sigma_i = 0$: In this case, by non-negativity, $\sigma_i = 0$ so $\operatorname{supp}(\sigma_i) = \emptyset$ contradicts the assumption. We conclude that $\int_\Gamma \varphi\,\mathrm{d}\sigma_i \in]0, 1[$ for both $i = 0, 1$.

Else: Set $\lambda \overset{\text{def.}}{=} \int_\Gamma \varphi\,\mathrm{d}\sigma_0 \in]0, 1[$ and $\int_\Gamma \varphi\,\mathrm{d}\sigma_1 = \int_\Gamma \varphi(\mathrm{d}\sigma - \mathrm{d}\sigma_0) \leq 1 - \lambda \in]0, 1[$, therefore

$$\sigma = \lambda\frac{\sigma_0}{\lambda} + (1 - \lambda)\frac{\sigma_1}{1 - \lambda}, \tag{A.28}$$

which confirms $\sigma \in]\sigma_0, \sigma_1[$ as required.

This claim immediately confirms that all extreme points must have at most one point in their support. Combined with the constraint $\int_\Gamma \varphi\,\mathrm{d}\sigma \leq 1$, we must have

$$\operatorname{Ext}(D) \subset \{0\} \cup \left\{\, \lambda\delta_\gamma \mid \gamma \in \Gamma,\ \varphi(\gamma) < +\infty,\ 0 < \lambda \leq \varphi(\gamma)^{-1} \,\right\}. \tag{A.29}$$

Finally, if $0 < \lambda < \varphi(\gamma)^{-1}$, then $\lambda\delta_\gamma \in \left]0, \varphi(\gamma)^{-1}\delta_\gamma\right[$, so $\lambda\delta_\gamma \notin \operatorname{Ext}(D)$. This confirms that the only extreme points are those found in the first half of the proof.

□

A.3 Lower semi-continuity of E

In Theorem 5.1 we assume $\varphi, w \colon \Gamma \to [0, +\infty]$ are lower semi-continuous, therefore Lemmas A.1 and A.5 confirm that W is lower semi-continuous and D is closed. It remains to show that F is lower semi-continuous. Lemma 2.3 shows that $(e_t)_\sharp$ is narrowly continuous, therefore F inherits lower semi-continuity from each F_j. From now on we consider any closed subset $\Gamma \subset \Gamma_\infty$ with the topology induced by d_Γ.

A.4 Compactness of sub-levelsets of E

Recall that $\mathrm{E} \colon \mathcal{M}^+(\Gamma) \to \mathbb{R}$ is defined by $\mathrm{E}(\sigma) = \mathrm{F}(\sigma) + \mathrm{W}(\sigma)$. We prove that minimisers of E exist by showing that E has compact sub-levelsets. Note in the case φ has compact sub-levelsets, then D is already compact (Lemma A.5), so the following theorem is not required.

Theorem A.6. *Suppose $D \subset \mathcal{M}^+(\Gamma)$ is narrowly closed, bounded, and* F *is convex lower semi-continuous. If* F *is bounded from below, $w \colon \Gamma \to [0, +\infty]$ is lower semi-continuous and has compact sub-levelsets, then $\mathrm{E}\,|_D$ also has compact sub-levelsets.*

Proof. Lemma A.3 shows that $\mathrm{W}\,|_D$ has compact sub-levelsets. Therefore for any $t \in \mathbb{R}$

$$\{\, \sigma \in D \mid \mathrm{E}(\sigma) \leq t \,\} = \{\, \sigma \in D \mid \mathrm{F}(\sigma) + \mathrm{W}(\sigma) \leq t \,\} \tag{A.30}$$

$$\subset \left\{\, \sigma \in D \mid \mathrm{W}(\sigma) \leq t - \inf_{\widetilde{\sigma} \in D} \mathrm{F}(\widetilde{\sigma}) \,\right\}. \tag{A.31}$$

By Lemma A.1, W is lower semi-continuous, therefore E is lower semi-continuous. We conclude that the left-hand side is closed and the right-hand side is compact by assumption. □

A.5 E has sparse minimisers

Since the function F is of the form

$$\mathrm{F}(\sigma) = \sum_{j=0}^{T} \mathrm{F}_j(A_j(e_{t_j})_\sharp\sigma) \quad \text{for some} \quad A_j \colon \mathcal{M}(\Gamma) \to \mathbb{R}^m,\ \text{convex}\ \mathrm{F}_j, \tag{A.32}$$

with A_j as in (5.1), F is convex lower semi-continuous. We can therefore use a representer theorem (e.g. [Boyer et al., 2019, Cor. 3.8]) to demonstrate the sparsity of minimisers.

Lemma A.7. *Suppose* $\operatorname{argmin}_{\sigma\in D} \mathrm{E}(\sigma) \neq \emptyset$ *for some* $D \subset \mathcal{M}^+(\Gamma)$ *closed and bounded of the form in* (A.24)*. Then there exists a minimiser* $\sigma^* \in D$ *such that*

$$\sigma^* = \sum_{i=1}^{s} a_i \delta_{\gamma_i} \quad \textit{for some} \quad a_i \geq 0,\ \gamma_i \in \Gamma \textit{ with } \varphi(\gamma_i) < +\infty, \tag{A.33}$$

where $s \leq m(T+1)+1$*. If in addition* $\int \varphi \,\mathrm{d}\sigma^* < 1$*, then* $s \leq m(T+1)$*.*

Proof. We reformulate the problem $\min_{\sigma\in D} E(\sigma)$ as

$$\min_{\sigma\in V} H(\widetilde{A}\sigma) + R(\sigma) \tag{A.34}$$

where V is the vector space of all $\sigma \in \mathcal{M}(\Gamma)$ such that $w \in L^1_{|\sigma|}(\Gamma)$, R is a convex "regulariser"

$$R(\sigma) \overset{\text{def.}}{=} W(\sigma) + \chi_D(\sigma) \quad \text{where} \quad \chi_D(\sigma) \overset{\text{def.}}{=} \begin{cases} 0 & \sigma \in D \\ +\infty & \text{else,} \end{cases} \tag{A.35}$$

with linearly closed level sets. The observation operator $\widetilde{A}\colon V \to (\mathbb{R}^m)^{(T+1)}$, is linear, defined as $\sigma \mapsto (A_j(e_{t_j})_\sharp \sigma)_{0\leq j\leq T}$, and $H\colon (a_0,\dots,a_T) \mapsto \sum_{j=0}^T F_j(a_j)$ is a convex "fidelity term".

Set $k \overset{\text{def.}}{=} m(T+1)$. We claim that there exists a minimiser σ^* such that

$$\sigma^* = \sum_{i=1}^{k+1} \theta_i \mu^i, \quad \text{with} \quad \sum_{i=1}^{k+1} \theta_i = 1, \text{ and } \forall i,\ \theta_i \geq 0,\ \mu^i \in \operatorname{Ext}(D)\,. \tag{A.36}$$

To prove this, fix $t \overset{\text{def.}}{=} R(\sigma)$ for some arbitrary $\sigma \in \operatorname{argmin}_{\sigma\in D} \mathrm{E}(\sigma)$ and define

$$\widetilde{D} \overset{\text{def.}}{=} D \cap \{\, \sigma \in V \mid \langle w, \sigma\rangle = t \,\} \subset \{R \leq t\}. \tag{A.37}$$

We now split the analysis into two cases depending on the value of t.

Case $t = \inf_V R$**:** In the case $t = \inf_V R = 0$, [Boyer et al., 2019, Cor. 3.8] tells us that there exists a minimiser σ^* which belongs to an elementary face of $\{R \leq t\}$ with dimension at most k, therefore also a face of $\widetilde{D}$ with dimension at most k. In particular, by Carathéodory's theorem, we can express σ^* in the form in (A.36) but with $\mu^i \in \operatorname{Ext}\left(\widetilde{D}\right)$.

Observe that for $t = 0$, we can equivalently write $\widetilde{D}$ as

$$\widetilde{D} = \left\{ \sigma \in \mathcal{M}^+(\{w = 0\}) \mid \int_\Gamma \varphi \,\mathrm{d}\sigma \leq 1 \right\} \subset D \tag{A.38}$$

therefore the extreme points of $\widetilde{D}$ can be computed explicitly. By Lemma A.5,

$$\operatorname{Ext}\left(\widetilde{D}\right) = \{0\} \cup \left\{\, \varphi(\gamma)^{-1}\delta_\gamma \mid \varphi(\gamma) < +\infty,\ w(\gamma) = 0 \,\right\} \subset \operatorname{Ext}(D)\,, \tag{A.39}$$

so (A.36) is satisfied.

Case $t > \inf_V R$**:** In this case the minimiser σ^* from [Boyer et al., 2019, Cor. 3.8] belongs to an elementary face of $\{R \leq t\}$ with dimension at most $(k-1)$ (hence also for $\widetilde{D}$).

Since $\{\, \sigma \in V \mid \langle w, \sigma\rangle = t \,\}$ is a hyperplane, it is possible to prove (see for instance Dubins [1962]) that σ^* belongs to a face of D with dimension at most k.

The formulation of (A.36) is again given by Carathéodory's theorem.

We may now deduce (A.33) from (A.36). From Lemma A.5, we know that the extreme points of D are either 0 or of the form $\varphi(\gamma)^{-1}\delta_\gamma$ where $\varphi(\gamma) < +\infty$, hence the general form of (A.33) where $s \leq k+1$.

In the special case of $\int \varphi \,\mathrm{d}\sigma^* < 1$, one of the atoms μ^i must be 0. As a result, we may remove it from the sum, so that $s \leq k$.

□

B Support of optimal measures

In this section we discuss the continuous-time interpolation of discrete measures. In particular, suppose

$$\sigma^* \in \operatorname*{argmin}_{\sigma \in \mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma), \qquad \mathrm{E}(\sigma) = \sum_{j=0}^{T} \mathrm{F}_j(A_j(e_{t_j})_\sharp \sigma) + \int_\Gamma w(\gamma)\,\mathrm{d}\sigma(\gamma) \tag{B.1}$$

for lower semi-continuous $w\colon \Gamma \to [0, +\infty]$. Note that we have removed the constraint set D in these computations. If there are two terms which are sensitive to the continuous-time interpolation of γ (rather than just $\gamma(t_j)$), then the interpolation must balance both terms. Removing the constraint of D, the statement is as follows.

Lemma B.1. *Assume that there is a Borel-measurable mapping $G\colon \Gamma \to \Gamma$ such that for all $\gamma \in \Gamma$,*

$$G(\gamma) \in \operatorname*{argmin}_{\widetilde{\gamma} \in \Gamma} \{\, w(\widetilde{\gamma}) \mid \widetilde{\gamma}(t_j) = \gamma(t_j) \text{ for all } j = 0, \dots, T \,\}. \tag{B.2}$$

Then,

$$\text{for a.e. } \gamma \in \mathrm{supp}(\sigma^*), \qquad w(\gamma) = \min_{\widetilde{\gamma} \in \Gamma} \{\, w(\widetilde{\gamma}) \mid \widetilde{\gamma}(t_j) = \gamma(t_j) \,\}. \tag{B.3}$$

Proof. Since G is Borel-measurable, we may define the image measure $\sigma^{**} \overset{\text{def.}}{=} G_\# \sigma^*$ by $\sigma^{**}(B) = \sigma^*(G^{-1}(B))$ for all Borel subset B of Γ. Then, since $w \geq 0$ is lower semi-continuous,

$$\int_\Gamma w\,\mathrm{d}\sigma^{**} = \int_\Gamma w(G(\gamma))\,\mathrm{d}\sigma^*(\gamma) \tag{B.4}$$

(indeed, the result holds for all continuous bounded function, see [Ambrosio et al., 2008, Sec. 5.2]. One may extend it to w by approximation with an increasing sequence of such functions and the monotone convergence theorem).

Since σ^* is optimal,

$$0 \geq \mathrm{E}(\sigma^*) - \mathrm{E}(\sigma^{**}) = \int_\Gamma [w(\gamma) - w(G(\gamma))]\,\mathrm{d}\sigma^*(\gamma) \tag{B.5}$$

and since $w(\gamma) - w(G(\gamma)) \geq 0$, we deduce that $w(\gamma) = w(G(\gamma))$ for σ^*-a.e. $\gamma \in \Gamma$. □

For the example in Section 3, because Ω is convex, the map $G(\gamma)$ returns the piece-wise linear interpolation of points $\gamma(t_0), \dots, \gamma(t_T)$. This is continuous, therefore measurable. The Wasserstein-Fisher-Rao example of Section 4 is more involved but details for the $\alpha = 0$ can be found in Corollary 4.1(i) of the preprint (https://arxiv.org/pdf/1506.06430.pdf) of Chizat et al. [2018b]. In summary, $G(\gamma) = (\widetilde{h}, \widetilde{\xi})$ is continuous, $\widetilde{h}(t)$ is a simple quadratic between each $[h(t_j), h(t_{j+1})]$, and $\widetilde{\xi}(t)$ follows a straight line between $\xi(t_j)$ and $\xi(t_{j+1})$ with speed varying like arctan.

C Conditional expectation lemma

Lemma C.1 (Lemma 6.2)**.** *Suppose X is an $\mathcal{S}$-measurable random variable and $Z \perp X$ is an independent random variable. Then, for all integrable Borel measurable functions g*

$$\mathbb{E}(g(X, Z)|\mathcal{S}) = \mathbb{E}\left[\mathbb{E}_Z(g(X, Z))|\mathcal{S}\right]. \tag{C.1}$$

Proof. Let Y be an arbitrary $\mathcal{S}$-measurable random variable. We are required to show that

$$\mathbb{E}_{\mathcal{S}}\left[Y \mathbb{E}_Z(g(X, Z))\right] = \mathbb{E}_{\mathcal{S},Z}\left[Y g(X, Z)\right]. \tag{C.2}$$

Recall that if $X, Y \perp Z$, then the joint law $\mathbb{P}(X = x, Y = y, Z = z)$ factorises into $\mathbb{P}(X = x, Y = y)\mathbb{P}(Z = z)$. We can therefore compute

$$\begin{aligned}
\mathbb{E}_{\mathcal{S},Z}\left[Yg(X,Z)\right] &= \int yg(x,z)\,\mathrm{d}\mathbb{P}(X = x, Y = y, Z = z) && \text{(C.3)}\\
&= \iint yg(x,z)\,\mathrm{d}\mathbb{P}(X = x, Y = y)\,\mathrm{d}\mathbb{P}(Z = z) && (X, Y \perp Z)\ \text{(C.4)}\\
&= \int y\left[\int g(x,z)\,\mathrm{d}\mathbb{P}(Z = z)\right]\mathrm{d}\mathbb{P}(X = x, Y = y) && \text{(Fubini's theorem)}\ \text{(C.5)}\\
&= \mathbb{E}_{\mathcal{S}}\left[Y\mathbb{E}_Z(g(X,Z))\right] && \text{(C.6)}
\end{aligned}$$

as required. □