Dynamical programming for off-the-grid dynamic inverse problems

Vincent Duval, Robert Tovey

ESAIM: COCV 30 (2024) 7
https://doi.org/10.1051/cocv/2023085

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

DYNAMICAL PROGRAMMING FOR OFF-THE-GRID DYNAMIC INVERSE PROBLEMS

Vincent Duval[1,*] and Robert Tovey[2]

Abstract. In this work we consider off-the-grid algorithms for the reconstruction of sparse measures from time-varying data. In particular, the reconstruction is a finite collection of Dirac measures whose locations and masses vary continuously 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. 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.

Mathematics Subject Classification. 28A33, 65K10, 65J20, 90C49.

Received February 4, 2022. Accepted November 26, 2023.

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 [1, 2]. 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 [3] 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) [1] and the unbalanced Wasserstein–Fisher–Rao (WFR) [4] 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 [5], however current methods are too slow for large-scale applications.

Keywords and phrases: Off-the-grid imaging, dynamic inverse problems, Frank–Wolfe, dynamical programming, optimal transport regularization.

[1] INRIA-Paris, MOKAPLAN, 75012 Paris, France.
[2] CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University 75016 Paris, France.

* Corresponding author: vincent.duval@inria.fr

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 $h\in\mathrm{C}([0,1])$ and trajectory $\xi\in\mathrm{C}([0,1];\overline{\Omega})$. We can think of ρ as representing the evolving physical volume that can be observed with a microscope or by eye, whereas σ more efficiently represents a collection of (trajectories of) particles, that we wish to reconstruct. The structure of the problem as seen from this second viewpoint closely resembles the Beurling-LASSO which is now well-understood [6–9]. 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 [7, 10, 11] 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 [12]) 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 [13–15].

1.1. Motivating example

To make these observations and the contribution of this work more concrete we will make reference to the motivating example described in [1] with numerical examples in [5]. We will describe this problem briefly here and postpone precise assumptions and details of function spaces to Section 6, where it is the $\delta=+\infty$ case in (6.1). For $\alpha,\beta>0$ the Benamou–Brenier penalty $\mathcal{W}$ is a map of non-negative space-time measures ρ defined by

$$\mathcal{W}(\rho)\overset{\text{def.}}{=}\inf_{v\in\mathrm{L}^2_\rho([0,1]\times\overline{\Omega};\mathbb{R}^d)}\left\{\int_{[0,1]\times\overline{\Omega}}\left[\alpha+\frac{\beta}{2}|v|^2\right]\mathrm{d}\rho\ \text{ s.t. }\ \partial_t\rho+\mathrm{div}(v\rho)=0\right\},\tag{1.1}$$

where the continuity equation is satisfied in the weak sense which will be clarified in (1.14). The function with $\alpha=0$, $\beta=1$ is referred to as the Benamou–Brenier energy whose main properties can be found in [16], Section 5.3.1. It was shown (*cf.* [1, 16]) that whenever $\mathcal{W}(\rho)<+\infty$, there exists a "disintegration" into spatial measures $\{\rho_t\}_{t\in[0,1]}$ such that the curve $t\mapsto\rho_t$ is continuous in the narrow topology and such that

$$\forall\psi\in\mathrm{L}^1([0,1]\times\overline{\Omega};\rho),\quad\int_{[0,1]\times\overline{\Omega}}\psi(t,x)\,\mathrm{d}\rho(t,x)=\int_0^1\left(\int_{\overline{\Omega}}\psi(t,x)\,\mathrm{d}\rho_t(x)\right)\mathrm{d}t.\tag{1.2}$$

Thanks to this property, given times $0=t_0<t_1<\ldots<t_T=1$, 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 narrowly continuous linear operators A_j. The more involved case of continuous-time observations is handled in [2] although we do not discuss it further in this work. To solve the corresponding inverse problem, [1], (43) proposes the minimisation of

$$\mathcal{E}(\rho)=\frac{1}{2}\sum_{j=0}^{T}\left\|A_j\rho_{t_j}-b_j\right\|_2^2+\mathcal{W}(\rho).\tag{1.3}$$

It is then shown in [1], Theorem 10 that there is a minimiser ρ^* which is a finite sum of extreme points of the Benamou–Brenier unit ball, *i.e.*

$$\text{for some } a_i \geq 0,\ \xi^i \in \mathrm{AC}^2([0,1]\,;\overline{\Omega}), \qquad \forall t \in [0,1]\,, \qquad \rho_t^* = \sum_{i=1}^{m(T+1)} a_i \delta_{\xi^i(t)}, \tag{1.4}$$

where AC^2 is the set of absolutely continuous function such that their (a.e. defined) pointwise derivative is square-integrable (see also [17], Sect. 1.1).

1.2. Outline and contributions

The main goal of the present article is to describe an algorithm for dynamic inverse problems in the space of measures which is significantly faster than the state-of-the-art method [5]. The cornerstone of our approach is a switch from Eulerian to Lagrangian point of view, in the spirit of the seminal work by Benamou and Brenier [18]: we regard a dynamic measure as a superposition of moving particles, representing a measure ρ on $[0,1] \times \overline{\Omega}$ with a measure σ on the space of (weighted) paths. While this representation already appears in the motivating works of Bredies *et. al.* [1, 2, 4, 5], it is mostly used in proofs, and the default representation is ρ (dynamic measures). In contrast, our algorithm is natively designed for problems in a Lagrangian representation, and it feels natural to introduce it for problems formulated in the space of paths, so as to fully appreciate its generality.

Therefore, we have organised the article so as to address the following questions.

- *How to switch from dynamic measures ρ on $[0,1] \times \overline{\Omega}$ to measures on paths σ, and back again.* Since [2, 4, 17], much is already known about how and when it is possible to describe a dynamic measure with a Lagrangian point of view. Section 2 gathers those results with minor adaptations, modelling this operation with a linear map $\Theta\colon \sigma \mapsto \rho$, where ρ is a dynamic measure, and σ is a measure in the space of paths. The extensions include allowing for signed measures ρ, less smoothness for curves ξ, and identifying topologies for which Θ is a continuous map.
- *Which variational problems in the space of paths we consider.* Next, Section 3 sets our framework for variational problems in the space of measures on paths. Previous works in this context have focused only on the Benamou–Brenier and Wasserstein–Fisher–Rao (a.k.a. Hellinger–Kantorovich) examples in the Eulerian setting [1, 4]. By switching to the Lagrangian setting, we are able to investigate a broad range of examples at the same time. In particular, our first main theoretical contribution is to prove that minimisers indeed exist, and at least one of them is a sparse measure: it is a finite superposition of weighted paths. This is similar to results in [1, 4] for the specific examples therein, but much easier to generalise in the Lagrangian setting where the regularisation term becomes linear. In fact, most of the difficulty is transferred to the disintegration theorems of Section 2. The biggest remaining challenge is that the topology in the space of paths $\mathrm{AC}^2\left([0,1]\,;\overline{\Omega}\right)$ is less simple than that of $[0,1] \times \overline{\Omega}$, and handling the corresponding measure space requires more care. Next, we prove that the minimisers are supported on "geodesics" of the chosen regulariser. While the geodesic structure was mentioned in [5] specifically for the Benamou–Brenier example, we prove it for a large class of regularisers. In Section 6, we confirm that the family of variational problems of Section 3 does indeed include both the Benamou–Brenier example, as well as the Wasserstein–Fisher–Rao example investigated in [4].
- *How to solve variational problems in the space of measures on paths.* Numerical work on the Benamou–Brenier example was recently presented in [5] with a generalised conditional gradient (a.k.a. Frank–Wolfe) algorithm. The most time consuming step is to compute new extreme points, *i.e.* curves γ, to add to the reconstruction. Even though only a small subset of $\mathrm{L}^\infty([0,1]\,;\mathbb{R} \times \overline{\Omega})$ is involved, optimising over such a set is a challenging task, the computation times are hardly compatible with practical applications. We offer two major algorithmic contributions. The first is the proposal and analysis of a new stochastic variant of Frank–Wolfe in Section 4, which accounts, *e.g.*, in our specific setting, to adding curves supported on a random mesh. Whilst being very similar to that used in [5], we prove almost-sure convergence for this

variant. To address the large complexity of computing new extreme points (*i.e.* adding curves γ to the reconstruction), our second algorithmic contribution reformulates this step into a shortest-path problem on an ordered, weighted, directed, acyclic graph in Section 5, which can be solved very efficiently using dynamical programming. Such a formulation crucially relies on the Lagrangian formulation studied in the previous sections.

Finally, in Section 7, we provide numerical experiments which demonstrate the efficiency of the proposed algorithm on both models: we first compare with the algorithm proposed in [5] on the Benamou–Brenier example, showing dramatic speedup, then we present numerical results for the unbalanced Wasserstein–Fisher–Rao example, for which there is no existing algorithm, to the best of our knowledge.

1.3. 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 $]\sigma_0, \sigma_1[\stackrel{\text{def.}}{=} [\sigma_0, \sigma_1] \setminus \{\sigma_0, \sigma_1\}$. A set $D \subseteq V$ is called *convex* if $]\sigma_0, \sigma_1[\subset D$ for all $\sigma_0, \sigma_1 \in D$. We say that $\sigma \in D$ is an *extreme point* (or *atom*) of D, and write $\sigma \in \operatorname{Ext}(D)$, if there are no points $\sigma_0, \sigma_1 \in D$ such that $\sigma \in]\sigma_0, \sigma_1[$. In other words,

$$\forall \lambda \in]0,1[\,, \forall \sigma_0, \sigma_1 \in D, \quad (\sigma = \lambda\sigma_0 + (1-\lambda)\sigma_1) \quad \Longrightarrow \quad (\sigma_0 = \sigma_1 = \sigma)\,. \tag{1.5}$$

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 refer to [19] for more detail.

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

$$|\sigma|(A) \stackrel{\text{def.}}{=} \sup \left\{ \sum_{i=1}^{n} |\sigma(A_i)| \mid n \in \mathbb{N}, \{A_1, \dots, A_n\} \text{ Borel partition of } A \right\} \tag{1.6}$$

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 $(\mathrm{C_b}(\Gamma))'$, 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

$$\forall \phi \in \mathrm{C_b}(\Gamma), \quad \lim_{n\to+\infty} \int_\Gamma \phi \,\mathrm{d}\sigma^n = \int_\Gamma \phi \,\mathrm{d}\sigma^*. \tag{1.7}$$

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

$$\operatorname{supp}(\sigma) \stackrel{\text{def.}}{=} \left(\bigcup \{\, U \mid \sigma(U) = 0,\ U \text{ is open} \,\} \right)^c. \tag{1.8}$$

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

$$[0,1] \times \overline{\Omega} \quad \text{for an open, bounded, convex domain} \quad \Omega \subseteq \mathbb{R}^d,\ d \geq 1 \tag{1.9}$$

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 $\psi \in \mathrm{C}([0,1] \times \overline{\Omega})$, and similarly on the space of curves $\sigma \in \mathcal{M}(\Gamma)$ and $\phi \in \mathrm{C}_{\mathrm{b}}(\Gamma)$.

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 $\rho \in \mathrm{C}_{\mathrm{w}}([0,1]\,; \mathcal{M}(\overline{\Omega})) \subset \mathcal{M}([0,1] \times \overline{\Omega})$ 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

$$\forall \psi \in \mathrm{C}(\overline{\Omega}), \qquad \left[t \mapsto \int_{\overline{\Omega}} \psi(x) \,\mathrm{d}\rho_t(x) \right] \in \mathrm{C}([0,1]) \tag{1.10}$$

and

$$\forall \psi \in \mathrm{L}^1_\rho([0,1] \times \overline{\Omega}), \qquad \int_{[0,1]\times\overline{\Omega}} \psi(t,x) \,\mathrm{d}\rho(t,x) = \int_0^1 \left(\int_{\overline{\Omega}} \psi(t,x) \,\mathrm{d}\rho_t(x) \right) \mathrm{d}t. \tag{1.11}$$

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

$$\forall \psi \in \mathrm{C}(\overline{\Omega}), \qquad \int_{\overline{\Omega}} \psi(x) \,\mathrm{d}[(e_t)_\sharp \sigma](x) \overset{\text{def.}}{=} \int_\Gamma h(t)\psi(\xi(t)) \,\mathrm{d}\sigma(h,\xi). \tag{1.12}$$

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 $\Theta(\sigma) \in \mathrm{C}_{\mathrm{w}}([0,1]\,; \mathcal{M}(\overline{\Omega}))$.

The continuity equation. In the rest of the paper, we use the following distributional definition of the continuity equation, formally

$$\partial_t \rho + \operatorname{div}(\rho v) = g\rho, \tag{1.13}$$

which expresses mass variation (or mass conservation if $g = 0$).

Definition 1.1. Let $\rho \in \mathcal{M}([0,1] \times \overline{\Omega})$ be a measure. We say that ρ satisfies the *continuity equation* if there exists $v \in \mathrm{L}^1_{|\rho|}([0,1] \times \overline{\Omega}; \mathbb{R}^d)$, $g \in \mathrm{L}^1_{|\rho|}([0,1] \times \overline{\Omega})$ such that

$$\forall \psi \in \mathrm{C}^1_c(]0,1[\times \overline{\Omega}), \qquad \int [\partial_t \psi + \nabla\psi \cdot v + \psi g] \,\mathrm{d}\rho = 0. \tag{1.14}$$

2. Preliminary results

As previously mentioned, the main function space of this work is the space of measures on paths $\mathcal{M}(\Gamma)$, where Γ is a set of continuous weighted paths in $\overline{\Omega}$, modelling particles with (varying) mass $h(t)$ at location $\xi(t)$, that is

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

For technical reasons we permit curves ξ which may not be continuous at points t where $h(t) = 0$. Intuitively, the location $\xi(t)$ is not necessarily meaningful if the particle has no mass and cannot be observed.

In this section, we review the necessary assumptions for $\mathcal{M}(\Gamma)$ to be a sufficiently well-behaved space. Firstly we require Γ to be a complete separable metric space. We follow the suggestion of [4], Proposition 3.6 where the

flat metric is used on the space of measures $\left\{ h\delta_\xi \in \mathcal{M}^+([0,1]\times\overline{\Omega}) \mid (h,\xi)\in\Gamma,\ h\geq 0 \right\}$ for a particular $\Gamma\subset\Gamma_0$. We use the isometric space $(\Gamma_0,\mathrm{d}_\Gamma)$, the properties of which are given by the following lemma.

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

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

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

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

$$(h_1,\xi_1)\sim(h_2,\xi_2) \iff h_1=h_2 \quad \textit{and} \quad \forall t\in\{h_1\neq 0\},\quad \xi_1(t)=\xi_2(t). \tag{2.4}$$

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

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

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

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

The proof is elementary but given in Appendix A for completeness as no specific reference could be found.

A key analytical tool in related prior works (*cf.* [1, 2, 4, 5]) 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 those previous results. The first theorem collects results from [4, 17] and provides minor extensions for the scope of the current work. In particular, we remove the necessity for $h\geq 0$ or elements (h,ξ) to satisfy further smoothness conditions.

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*

$$\forall\psi\in\mathrm{C}([0,1]\times\overline{\Omega}), \qquad \int_{[0,1]\times\overline{\Omega}}\psi(t,x)\,\mathrm{d}\Theta(\sigma)(t,x)=\int_{\Gamma_0}\left(\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t\right)\mathrm{d}\sigma(h,\xi). \tag{2.7}$$

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* (2.7) *holds for all* $\psi\in\mathrm{L}^1_{|\Theta(\sigma)|}([0,1]\times\overline{\Omega})$*.*
3. *If* $\int_{\Gamma_0}\|h\|_\infty\,\mathrm{d}|\sigma|<+\infty$*, then* $\Theta(\sigma)\in\mathrm{C}_\mathrm{w}([0,1];\mathcal{M}(\overline{\Omega}))$*.*
4. *Suppose* $h,\xi\in\mathrm{AC}^2([0,1])$ *for* σ*-a.e.* $(h,\xi)\in\Gamma_0$*. 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*

$$h'(t)=g(t,\xi(t))h(t) \text{ for } \sigma\text{-a.e. } (h,\xi) \text{ and a.e. } t\in\,]0,1[, \tag{2.8}$$

$$\xi'(t)=v(t,\xi(t)) \text{ for } \sigma\text{-a.e. } (h,\xi) \text{ and a.e. } t \text{ such that } h(t)\neq 0, \tag{2.9}$$

$$\textit{and}\ \int_{\Gamma_0}\int_0^1(1+|v(t,\xi(t))|+|g(t,\xi(t))|)\,|h(t)|\,\mathrm{d}t\,\mathrm{d}|\sigma|(h,\xi)<+\infty, \tag{2.10}$$

then $\int_{\Gamma_0} \|h\|_\infty \,\mathrm{d}|\sigma| < +\infty$ *and* $\Theta(\sigma)$ *satisfies the continuity equation* (1.14).
Conversely, given $\rho \in \mathcal{M}([0,1]\times\overline{\Omega})$*, if* $\rho \geq 0$ *satisfies the continuity equation* (1.14) *and*

$$\int_{[0,1]\times\overline{\Omega}} (1 + |v(t,x)|^2 + |g(t,x)|^2)\,\mathrm{d}\rho(t,x) < +\infty, \tag{2.11}$$

then $\rho = \Theta(\sigma)$ *for some* $\sigma \in \mathcal{M}^+(\Gamma_0)$ *such that* (2.8)–(2.10) *hold and* $\int_{\Gamma_0} \|h\|_\infty \,\mathrm{d}\sigma < +\infty$.

These results are mainly proved in [4, 17], we extend them to signed measures σ in the appendix (Thm. A.2). The results achieve two key relations: characterising when the mapping from $\mathcal{M}([0,1]\times\overline{\Omega})$ to $\mathcal{M}(\Gamma_0)$ is well-defined, and when $\Theta(\sigma) \in \mathrm{C_w}([0,1]\,;\mathcal{M}(\overline{\Omega}))$. Weak continuity is of practical importance in applications. Without it, for example if the data were a video, we could not consider one frame to correspond to a single instance in time. Unfortunately we have seen that not all $\sigma \in \mathcal{M}(\Gamma_0)$ satisfy this smoothness requirement. However, in the next lemma we will confirm that, if $\Gamma \subset \Gamma_0$ is sufficiently "small", then $\Theta(\sigma) \in \mathrm{C_w}([0,1]\,;\mathcal{M}(\overline{\Omega}))$ for all $\sigma \in \mathcal{M}(\Gamma)$ due to the implicit assumption that $\|\sigma\| < +\infty$. A related property is the continuity of the operator Θ which we also confirm.

Lemma 2.3. *For each* $p \in [1,+\infty]$ *define the set*

$$\Gamma_p \overset{\text{def.}}{=} \left\{ \gamma = (h,\xi) \in \Gamma_0 \mid \|h\|_p \leq 1 \right\}, \tag{2.12}$$

then

$$\left\{ \Theta(\sigma) \mid \sigma \in \mathcal{M}(\Gamma_0),\ \int_{\Gamma_0} \|h\|_p \,\mathrm{d}|\sigma| < +\infty \right\} = \{\, \Theta(\widehat{\sigma}) \mid \widehat{\sigma} \in \mathcal{M}(\Gamma_p) \,\} \tag{2.13}$$

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

$$\text{for all } p \in [1,+\infty]\,, \qquad \Theta(\sigma^n) \overset{*}{\rightharpoonup} \Theta(\sigma) \text{ narrowly in } \mathcal{M}([0,1]\times\overline{\Omega}), \tag{2.14}$$

$$\text{if } p = +\infty,\ \forall t \in [0,1]\,, \qquad (e_t)_\sharp\sigma^n \overset{*}{\rightharpoonup} (e_t)_\sharp\sigma \text{ narrowly in } \mathcal{M}(\overline{\Omega}). \tag{2.15}$$

The proof of this lemma is found in Lemma A.4, relying heavily on the definition of continuity given by [31].

Remark 2.4. 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 is a closed subset of ${}^{\Gamma_\infty}\!/\!\sim$ in order for it to be a complete separable metric space with $\Theta\colon \mathcal{M}(\Gamma) \to \mathrm{C_w}([0,1]\,;\mathcal{M}(\overline{\Omega}))$.

Summarising the results of this section, in the remainder of this work we want to use a domain $D \subset \mathcal{M}^+(\Gamma_0)$ such that $\Theta|_D$ has nice properties with respect to the space $\mathrm{C_w}([0,1]\,;\mathcal{M}(\overline{\Omega}))$. The combination of Theorem 2.2 and Lemma 2.3 show that it is sufficient to consider either $D \subset \left\{ \sigma \in \mathcal{M}^+(\Gamma_0) \mid \int_{\Gamma_0} \|h\|_\infty \,\mathrm{d}\sigma < \infty \right\}$ or simply $D \subset \mathcal{M}^+(\Gamma_\infty)$. Analytically we will always consider $D \subset \mathcal{M}^+(\Gamma_\infty)$ as it is more concise, although numerically either convention is equivalent. The generality of allowing any closed subset $\Gamma \subset \Gamma_\infty$ allows us to treat different applications with the same analysis, for example:

$\boldsymbol{\Gamma \subset \{\,(\mathbf{h},\xi) \in \boldsymbol{\Gamma}_\infty \mid \mathbf{h} \equiv \mathbf{1}\,\}$: This enforces balanced transport (*e.g.* the Benamou–Brenier example [1]). 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).

$\boldsymbol{\Gamma} \subset \{ (\mathbf{h}, \xi) \in \boldsymbol{\Gamma}_\infty \mid \mathbf{h} \in \mathrm{C}([\mathbf{0}, \mathbf{1}] ; [\mathbf{0}, \mathbf{1}]) \}$: This allows unbalanced transport of non-negative mass (*e.g.* the Wasserstein–Fisher–Rao example [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.

$\boldsymbol{\Gamma} \subset \boldsymbol{\Gamma}_\infty$: In the general case $\Theta(\sigma)$ is a general signed measure, even when $\sigma \geq 0$. In words, σ can give positive weight to curves with negative mass. The only constraint is that $\Theta(\sigma) \in \mathrm{C}_\mathrm{w}([0,1] ; \mathcal{M}(\overline{\Omega}))$ is continuous in time.

3. Core variational problem

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

$$\forall \rho \in \mathcal{M}(\overline{\Omega}),\ i = 1, \dots, m,\ j = 0, \dots, T, \qquad (A_j \rho)_i \overset{\text{def.}}{=} \int_{\overline{\Omega}} a_i^j(x) \,\mathrm{d}\rho(x). \tag{3.1}$$

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

$$\Gamma \subset \Gamma_\infty \overset{\text{def.}}{=} \left\{ \gamma = (h, \xi) \mid h \in \mathrm{C}([0,1] ; [-1,1]),\ \xi : [0,1] \to \overline{\Omega},\ \xi|_{\{h \neq 0\}} \text{ is continuous} \right\}, \tag{3.2}$$

so that $\forall t \in [0,1]$, the map $(e_t)_\sharp : \mathcal{M}(\Gamma) \to \mathcal{M}(\overline{\Omega})$ is narrowly continuous. We therefore choose a data fidelity $\mathrm{F} : \mathcal{M}(\Gamma) \to [0, +\infty[$ of the form

$$\text{for some convex } \mathrm{F}_j \in \mathrm{C}^2(\mathbb{R}^m ; [0, +\infty[), \qquad \mathrm{F}(\sigma) \overset{\text{def.}}{=} \sum_{j=0}^{T} \mathrm{F}_j(A_j \left[(e_{t_j})_\sharp \sigma \right]). \tag{3.3}$$

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

$$\forall \sigma \in D, \quad \mathrm{W}(\sigma) \overset{\text{def.}}{=} \int_\Gamma w(\gamma) \,\mathrm{d}\sigma(\gamma) \quad \text{where} \quad D \overset{\text{def.}}{=} \left\{ \sigma \in \mathcal{M}(\Gamma) \quad | \quad \sigma \geq 0, \quad \int_\Gamma \varphi(\gamma) \,\mathrm{d}\sigma \leq 1 \right\}. \tag{3.4}$$

We consider minimising the energy $\mathrm{E} : D \to]-\infty, +\infty]$ defined by

$$\forall \sigma \in D, \qquad \mathrm{E}(\sigma) \overset{\text{def.}}{=} \mathrm{F}(\sigma) + \mathrm{W}(\sigma). \tag{3.5}$$

The motivation behind this in an Inverse Problems setting is that F represents a smooth data fidelity with linear observations recorded at times t_j, and the combination of W and D represent a regularisation of the problem. The choice of φ (hence of D) is often made to ensure the well-posedness of the model, but we are mostly interested in cases where the constraint $\int_\Gamma \varphi(\gamma) \,\mathrm{d}\sigma \leq 1$ is not active, so that the choice of φ has no impact on the set of minimisers.

3.1. Existence of sparse minimisers

Any choice of energy in this framework leads to a sparse reconstruction in the space of curves.

Theorem 3.1. *If A_j are given by (3.1) and $\varphi, w : \Gamma \to [0, +\infty]$ are lower semi-continuous, then* F *and* E *are lower semi-continuous. Furthermore, recall* F *is bounded below. If 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

$$\text{for some } a_i \geq 0,\ \gamma^i \in \Gamma, \qquad \sigma^* = \sum_{i=1}^{s} a_i \delta_{\gamma^i} \tag{3.6}$$

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 B. Theorem 3.1 is reminiscent of the main result of [1] in the particular case of the Benamou–Brenier energy, but our setting is easier as we start from the formulation on paths. All the difficulty is in the disintegration results of Theorem 2.2.

The smaller value of s will often be valid in practice, for example any choice $\varphi(\gamma) \leq \frac{w(\gamma)}{\mathrm{E}(0)+1}$ is always sufficient.

Remark 3.2. The Benamou–Brenier example from Section 1.1 can be formulated in this setting with

$$\Gamma = \left\{ (h,\xi) \in \Gamma_0 \mid h \equiv 1, \xi \in \mathrm{AC}^2([0,1]\,;\overline{\Omega}) \right\}, \ \mathrm{F}_j(A_j\rho) = \frac{1}{2}\|A_j\rho - b_j\|_2^2, \quad \text{and} \quad w(h,\xi) = \int_0^1 \alpha + \frac{\beta}{2}|\xi'(t)|^2 \,\mathrm{d}t. \tag{3.7}$$

More details are given in Section 6 where the Benamou–Brenier example is the limiting case $\delta \to +\infty$ in (6.1). The choice of φ in [5] was equivalent to $\varphi(\gamma) = \frac{w(\gamma)}{\mathrm{E}(0)}$, whereas we suggest the default of $\varphi(\gamma) = \frac{\alpha}{\mathrm{E}(0)}$ which is easier to analyse. Both functions are strictly positive but sufficiently small to ensure $\int_\Gamma \varphi \,\mathrm{d}\sigma^* < 1$.

3.2. Discrete-time formulation

Recall that $\Gamma \subset L^\infty([0,1]\,;[-1,1]\times\overline{\Omega})$ is a space of continuous-time curves, denote the discrete-time space

$$\widetilde{\Gamma} \overset{\text{def.}}{=} \left\{ (\gamma(t_0), \dots, \gamma(t_T)) \mid \gamma \in \Gamma \right\} \subset ([-1,1]\times\overline{\Omega})^{T+1}. \tag{3.8}$$

Until now we have formulated E as a function of measures $\sigma \in \mathcal{M}(\Gamma)$, but in this subsection we show that it can be thought of equivalently as a function $\widetilde{\mathrm{E}}$ of $\widetilde{\sigma} \in \mathcal{M}(\widetilde{\Gamma})$. For the remainder of this section, without loss of generality, we assume there is a minimiser

$$\sigma^* \in \operatorname{argmin} \left\{ \mathrm{F}(\sigma) + \int_\Gamma w \,\mathrm{d}\sigma \mid \sigma \in \mathcal{M}^+(\Gamma) \right\} \tag{3.9}$$

which is also a minimiser of the energy considered in (3.5), *i.e.* $\int \varphi \,\mathrm{d}\sigma^* < 1$. If that is not the case, one can use a Lagrange multiplier to form a modified energy with $w \leftarrow w + \lambda\varphi$ for some $\lambda \geq 0$. Our key observation is that σ^* must be supported on "geodesics" of w (or $w+\lambda\varphi$) which interpolate the discrete-time curves.

Lemma 3.3. *Suppose* $w\colon \Gamma \to [0,+\infty]$ *is lower semi-continuous with compact sub-levelsets. Then, for all minimisers* σ^* *in* (3.9), $\gamma \in G(\gamma(t_0),\dots,\gamma(t_T))$ *for a.e.* $\gamma \in \operatorname{supp}(\sigma^*)$ *where* $G\colon \widetilde{\Gamma} \rightrightarrows \Gamma$ *is given by*

$$G(\widetilde{\gamma}) \overset{\text{def.}}{=} \operatorname*{argmin}_{\gamma\in\Gamma} \left\{ w(\gamma) \mid \forall j = 0,\dots,T,\ \gamma(t_j) = \widetilde{\gamma}_j,\ w(\gamma) < +\infty \right\}. \tag{3.10}$$

The first part of the proof uses a measurable choice theorem from [20].

Lemma 3.4. *There exists a Borel function* $g\colon \Gamma \to \Gamma$ *such that* $g(\gamma) \in G(\gamma(t_0),\dots,\gamma(t_T))$ *for all* $\gamma \in \Gamma$.

Proof of Lemma 3.4. We apply [20], Theorem 1 with $U = ([-1,1]\times\overline{\Omega})^{T+1}$, $V = \Gamma$ which is a complete separable metric space by Theorem 2.1, and a set

$$E \overset{\text{def.}}{=} \left\{ (\widetilde{\gamma}, \gamma) \in ([-1,1]\times\overline{\Omega})^{T+1} \times \Gamma \mid \gamma \in G(\widetilde{\gamma}) \right\}$$

which is Borel. For all $\widetilde{\gamma} \in ([-1,1] \times \overline{\Omega})^{T+1}$, the section $E_{\widetilde{\gamma}} = \{\, \gamma \in \Gamma \mid (\widetilde{\gamma}, \gamma) \in E \,\} = G(\gamma)$ is compact (as the intersection of a closed set and a sub-levelset of w) hence σ-compact. As a result, [20], Theorem 1 ensures that there exists a Borel selection of E, which is the desired function g. □

We can now return to the main result of this subsection.

Proof of Lemma 3.3. Let g be any function given by Lemma 3.4. The proof will show the contradiction that $\mathrm{E}(g_\sharp \sigma^*) < \mathrm{E}(\sigma^*)$ if σ^* is not strictly supported on the image of G. First we confirm that $\mathrm{F}(g_\sharp \sigma^*) = \mathrm{F}(\sigma^*)$. By (3.1), for all $i = 1, \dots, m$, $j = 0, \dots, T$

$$(A_j(e_{t_j})_\sharp(g_\sharp)\sigma^*) = \int_{\overline{\Omega}} a_i^j(x)\, \mathrm{d}(e_{t_j})_\sharp(g_\sharp)\sigma^* = \int_\Gamma h(t_j) a_i^j(\xi(t_j))\, \mathrm{d}\sigma^*(\gamma) \quad \text{where} \quad (h,\xi) = g(\gamma) \tag{3.11}$$

$$= \int_\Gamma h(t_j) a_i^j(\xi(t_j))\, \mathrm{d}\sigma^*(\gamma) \quad \text{where} \quad (h,\xi) = \gamma \tag{3.12}$$

$$= (A_j(e_{t_j})_\sharp \sigma^*). \tag{3.13}$$

Therefore each $\mathrm{F}_j(A_j(e_{t_j})_\sharp(g_\sharp)\sigma^*) = \mathrm{F}_j(A_j(e_{t_j})_\sharp \sigma^*)$ as required. On the other hand, note $w(g(\gamma)) \leq w(\gamma)$ always, suppose there exist $\varepsilon, \delta > 0$ such that $\sigma^*(\{\, \gamma \mid w(g(\gamma)) \leq w(\gamma) - \delta \,\}) = \varepsilon$. In which case

$$\mathrm{W}(g_\sharp \sigma^*) = \int_{w(g(\gamma)) \leq w(\gamma) - \delta} w(g(\gamma))\, \mathrm{d}\sigma^* + \int_{w(g(\gamma)) > w(\gamma) - \delta} w(g(\gamma))\, \mathrm{d}\sigma^* \leq \mathrm{W}(\sigma^*) - \delta\varepsilon. \tag{3.14}$$

Combining these equations, that would imply that $\mathrm{E}(g_\sharp \sigma^*) < \mathrm{E}(\sigma^*)$, contradicting the optimality of σ^*. We conclude that $\sigma^*(\{\, \gamma \mid w(g(\gamma)) < w(\gamma) \,\}) = 0$ as required. □

This shows we can perform computations in the discrete-time space $\widetilde{\Gamma}$ and later lift curves back to Γ using the geodesics G. This holds both pointwise between $\widetilde{\Gamma} \leftrightarrow \Gamma$ and with measures $\mathcal{M}(\widetilde{\Gamma}) \leftrightarrow \mathcal{M}(\Gamma)$.

Remark 3.5. For the Benamou–Brenier example, w is given in Remark 3.2. If Ω is convex, then $G(\widetilde{\gamma}) = \{\gamma\}$ where γ is the unique piecewise linear interpolant of the points $(t_j, \widetilde{\gamma}_j)$, as commented in [5], Remark 4.10. Also,

$$\forall \widetilde{\gamma} = (h_0, \xi_0, \dots, h_T, \xi_T) \in \widetilde{\Gamma}, \qquad w(G(\widetilde{\gamma})) = \alpha + \frac{\beta}{2} \sum_{j=1}^{T} \frac{|\xi_j - \xi_{j-1}|^2}{t_j - t_{j-1}}.$$

We therefore know that all minimisers σ^* are supported on the set

$$\Gamma = \left\{\, (h,\xi) \in \mathrm{C}([0,1]\,; \{1\} \times \overline{\Omega}) \mid \xi|_{]t_{j-1}, t_j[} \text{ is linear for each } j = 1, \dots, T \,\right\}. \tag{3.15}$$

Restricting to this domain of curves is much more computationally convenient without losing analytical accuracy.

4. Frank–Wolfe convergence

While Theorem 3.1 guarantees an analytical structure of minimisers, we must now choose a reconstruction algorithm which is capable of taking advantage of this structure. As previously stated, we will use a variant of the Frank–Wolfe algorithm to take advantage of the sparse structure of reconstructions.

4.1. The Frank–Wolfe algorithm

In order to derive a variant of the Frank–Wolfe algorithm with inexact or stochastic steps, we first need to highlight a few properties which, to the best of our knowledge, have not been stated in the literature. We refer the reader to [12] for a thorough introduction to the Frank–Wolfe algorithm. For the sake of generality

and future reference, we work in an abstract setting, assuming that we want to minimize a convex function $f\colon D \to \mathbb{R}$, where D is a nonempty convex set of a Hausdorff locally convex vector space $\mathcal{X}$.

Standard results in convex analysis [21], Chapter 7 ensure that for all $x \in D$ and all $h \in (D - x)$, the directional derivative

$$f'(x;h) \overset{\text{def.}}{=} \lim_{t \to 0^+} \frac{f(x+th) - f(x)}{t}$$

exists in $\mathbb{R} \cup \{-\infty\}$ and is a convex function of h. The main steps of the Frank–Wolfe algorithm are given in Algorithm 1. The standard choice [22, 23] uses the *Linear Minimisation Oracle* defined as

$$s^{n+1} = \mathrm{LMO}(x^n) \in \operatorname*{argmin}_{s \in D} \left(f(x^n) + f'(x^n; s - x^n)\right) = \operatorname*{argmin}_{s \in D} f'(x^n; s - x^n). \tag{4.1}$$

In the large majority of cases the existence of the LMO is implied by f being Gâteaux-differentiable and D compact. The analysis in [12] stresses that Algorithm 1 with the linear minimisation oracle (4.1) yields a *minimising sequence*, *i.e.* $\lim_{n\to+\infty} f(x^n) = \inf_D f$, provided the curvature of f is finite,

$$C_f \overset{\text{def.}}{=} \sup_{\substack{x \in D, \widetilde{x} \in D \\ \lambda \in]0,1[}} \frac{f(x + \lambda(\widetilde{x} - x)) - f(x) - \lambda f'(x; \widetilde{x} - x)}{\lambda^2} < +\infty. \tag{4.2}$$

In situations where the LMO is not easy to compute, we can instead choose any $s^{n+1} \in D$ and measure its suitability using the primal-dual gap

$$\forall x, s \in D, \qquad \mathrm{gap}(\mathrm{x};\mathrm{s}) \overset{\text{def.}}{=} \mathrm{f}'(\mathrm{x};\mathrm{s}-\mathrm{x}) - \inf_{\widetilde{\mathrm{s}} \in \mathrm{D}} \mathrm{f}'(\mathrm{x};\widetilde{\mathrm{s}}-\mathrm{x}). \tag{4.3}$$

In general $\mathrm{gap} \geq 0$ and $\mathrm{gap}(\mathrm{x};\mathrm{LMO}(\mathrm{x})) = 0$, suppose $\mathrm{gap}(\mathrm{x}^{\mathrm{n}};\mathrm{s}^{\mathrm{n}+1}) \leq \varepsilon_{\mathrm{n}}$ for some controlled error $\varepsilon_n \geq 0$. It has previously been shown in [12], Theorem 1 that x^n is a minimising sequence whenever $\varepsilon_n = O(1/n)$.

In our context, the guarantee $\varepsilon_n = O(1/n)$ would still be prohibitive for large n. The linear minimisation oracle consists of finding a curve $\gamma \in \Gamma$ which minimises a certain energy, see Section 4.3 below. Instead, we propose to use random discretisations of the domain so that implicitly $\liminf_{n\to+\infty} \varepsilon_n = 0$, without a guaranteed rate. In comparison to the result of [12], the relaxed assumption on ε_n is accounted for by the linesearch in Line 4 which implicitly selects large/small steps when ε_n is correspondingly large/small, *i.e.* when we have a lucky/unlucky draw. Our only assumptions on f are that it is Gâteaux differentiable with bounded curvature ($C_f < +\infty$) and $\mathrm{gap}(\mathrm{x};\mathrm{s}) < +\infty$ for all $x, s \in D$. Then, in the case that $\liminf_{n\to+\infty} \varepsilon_n = 0$ (possibly almost surely), we show that x^n is a minimising sequence (almost surely).

One final aspect of our algorithm is the freedom which is granted in Line 5: one may choose any point which has lower energy than the one provided by the linesearch. This is now a standard addition to the Frank–Wolfe algorithm to enable much faster practical convergence, see for instance [7, 10, 11]. We exploit this in Section 7.

Algorithm 1 Abstract Frank–Wolfe algorithm

1: Choose $x^0 \in D$, $n \leftarrow 0$,
2: **repeat**
3: Choose $s^{n+1} \in D$ ▷ oracle
4: $\lambda_n \leftarrow \operatorname{argmin}_{\lambda \in [0,1]} f((1-\lambda)x^n + \lambda s^{n+1})$ ▷ exact linesearch
5: Choose x^{n+1} such that $f(x^{n+1}) \leq f((1-\lambda_n)x^n + \lambda_n s^{n+1})$ ▷ improvement over the linesearch
6: $n \leftarrow n + 1$
7: **until** converged

To ease the analysis of the randomised algorithm, we first analyse the deterministic version.

Proposition 4.1. *Let f, D be as above, and assume that $C_f < +\infty$ (see (4.2)). If*

$$\liminf_{n\to+\infty} \mathrm{gap}(\mathrm{x^n}; \mathrm{s^{n+1}}) = 0, \text{i.e.} \liminf_{\mathrm{n}\to+\infty} \left(\mathrm{f'}(\mathrm{x^n}; \mathrm{s^{n+1}} - \mathrm{x^n}) - \min_{\mathrm{D}} \mathrm{f'}(\mathrm{x^n}; \cdot - \mathrm{x^n}) \right) = 0, \tag{4.4}$$

then Algorithm 1 yields a minimizing sequence.

Proof. First observe from Lines 4 and 5, and by definition of C_f that

$$\forall \lambda \in [0,1], \quad f(x^{n+1}) \leq f((1-\lambda)x^n + \lambda s^{n+1}) \leq f(x^n) + \lambda f'(x^n; s^{n+1} - x^n) + C_f \lambda^2 \tag{4.5}$$

$$\text{hence} \quad f(x^{n+1}) - f(x^n) - \lambda \min_D (f'(x^n; \cdot - x^n)) \leq \lambda \mathrm{gap}(\mathrm{x^n}; \mathrm{s^{n+1}}) + \mathrm{C_f}\lambda^2. \tag{4.6}$$

Note from the $\lambda = 0$ case we have $f(x^n) \geq f(x^{n+1})$ for each n, therefore it is clear that $(f(x^n))_{n\in\mathbb{N}}$ converges to some limit $\ell \geq \inf_{x\in D} f(x) \geq -\infty$. We are required to show that $\ell = \inf_{x\in D} f(x)$.

The case $\ell = -\infty$ is trivial, otherwise we also have $\lim_{n\to+\infty} f(x^{n+1}) - f(x^n) = 0$. Now, taking the lim inf on both sides of (4.6) gives

$$\liminf_{n\to+\infty} -\lambda \min_D (f'(x^n; \cdot - x^n)) \leq C_f \lambda^2. \tag{4.7}$$

Dividing by $\lambda \to 0^+$, we obtain $\limsup_{n\to+\infty} \min_D (f'(x^n; \cdot - x^n)) \geq 0$. On the other hand, by convexity,

$$\forall x \in D, \quad f(x) \geq \limsup_{n\to+\infty} [f(x^n) + f'(x^n; x - x^n)] \geq \ell + \limsup_{n\to+\infty} \min_D (f'(x^n; \cdot - x^n)) \geq \ell. \tag{4.8}$$

Since $x \in D$ is arbitrary, we deduce that $\ell = \lim_{n\to\infty} f(x^n) = \inf_D f$, as required. □

4.2. Stochastic variant

Now, we may study the behaviour of the algorithm in a stochastic framework. We build a random process $(X^n, S^n)_{n\in\mathbb{N}^*}$ in $D \times D$, by considering a random initialisation X^0 in D, and applying Algorithm 1 by picking a random point S^{n+1} in D on Line 3. Note that, at each step, S^{n+1} may depend on $\{X^k\}_{k=0}^n$ and $\{S^k\}_{k=1}^n$.

Typically, as in Section 4.3, we consider a setting where solving (4.1) exactly is too costly, and where one draws a random grid on which to perform the optimisation. The variable S^{n+1} is then a minimizer of $f'(X^n; \cdot - X^n)$ among a finite, small, subset of D.

Let $(\mathcal{S}^n)_{n\in\mathbb{N}}$ be the filtration generated by that random process, *i.e.* $\mathcal{S}^n$ is the σ-algebra generated by $\{X^0\} \cup \{X^k\}_{1\leq k\leq n} \cup \{S^k\}_{1\leq k\leq n}$.

Proposition 4.2. *Let $(X^n)_{n\in\mathbb{N}}$ and $(S^n)_{n\in\mathbb{N}^*}$ as described above, and let $(\mathcal{S}^n)_{n\in\mathbb{N}}$ be the filtration they generate. If for all $\varepsilon > 0$, all $n \in \mathbb{N}$, there is some deterministic $p_n(\varepsilon) > 0$ such that*

$$\mathbb{P}(\mathrm{gap}(\mathrm{X^n}; \mathrm{S^{n+1}}) < \varepsilon | \mathcal{S}^n) \geq \mathrm{p_n}(\varepsilon) \quad \textit{almost surely,}$$

and $\sum_{n=1}^\infty p_n(\varepsilon) = +\infty$, then $(f(X^n))_{n\in\mathbb{N}}$ is a minimizing sequence almost surely.

The above proposition is a consequence of the following lemma, setting $G^{n+1} = \mathrm{gap}(\mathrm{X^n}; \mathrm{S^{n+1}})$.

Lemma 4.3. *Let $(\mathcal{S}^n)_{n\in\mathbb{N}}$ be a filtration, and $(G^n)_{n\in\mathbb{N}}$ a family of random variables such that for each $n, m \in \mathbb{N}$, $n \leq m$, G^n is $\mathcal{S}^m$-measurable. If for all $\varepsilon > 0$, $n \in \mathbb{N}$, there is some deterministic $p_n(\varepsilon) \geq 0$ such that*

$$\mathbb{P}(G^{n+1} < \varepsilon | \mathcal{S}^n) \geq p_n(\varepsilon) \quad \textit{almost surely} \tag{4.9}$$

and $\sum_{n=0}^{\infty} p_n(\varepsilon) = +\infty$*, then* $\liminf_{n\to\infty} G^n \leq 0$ *almost surely.*

Proof. Fix $\varepsilon > 0$. For all $N, M \in \mathbb{N}$ with $M \geq N$,

$$\mathbb{P}\left(\bigcap_{n\geq N}\{G^n \geq \varepsilon\}\right) \leq \mathbb{P}\left(\bigcap_{n=N}^{M+1}\{G^n \geq \varepsilon\}\right) = \mathbb{E}\left[\mathbb{1}_{\{G^{M+1}>\varepsilon\}}\left(\prod_{n=N}^{M}\mathbb{1}_{\{G^n\geq\varepsilon\}}\right)\right] \tag{4.10}$$

$$\leq \mathbb{E}\left[\mathbb{E}\left(\mathbb{1}_{\{G^{M+1}\geq\varepsilon\}}|\mathcal{S}^M\right)\prod_{n=N}^{M}\mathbb{1}_{\{G^n\geq\varepsilon\}}\right] \tag{4.11}$$

$$\leq (1-p_{M+1}(\varepsilon))\mathbb{P}\left(\bigcap_{n\geq N}\{G^n \geq \varepsilon\}\right) \tag{4.12}$$

$$\leq \prod_{n=N}^{M+1}(1-p_n(\varepsilon)) \tag{4.13}$$

$$\leq \exp\left(-\sum_{n=N}^{M+1} p_n(\varepsilon)\right). \tag{4.14}$$

Letting $M \to +\infty$, we get $\mathbb{P}\left(\bigcap_{n\geq N}\{G^n \geq \varepsilon\}\right) = 0$, that is $\mathbb{P}\left(\bigcup_{n\geq N}\{G^n < \varepsilon\}\right) = 1$. As a result,

$$\mathbb{P}\left(\liminf_{n\to\infty} G^n \leq 0\right) = \mathbb{P}\left(\bigcap_{k\in\mathbb{N}^*}\bigcap_{N\in\mathbb{N}}\bigcup_{n\geq N}\left\{G^n < \frac{1}{k}\right\}\right) = \lim_{k\to+\infty}\lim_{N\to+\infty}\mathbb{P}\left(\bigcup_{n\geq N}\left\{G^n < \frac{1}{k}\right\}\right) = 1.$$

□

To summarise, the main convergence requirement of Proposition 4.1 is to guarantee $\liminf_{n\to+\infty} G^n \leq 0$. Our solution to this is to use random discretisations so that the stochastic variant of Algorithm 1 still converges asymptotically (almost surely), but the complexity of each individual iteration remains low. A similar idea in the setting of stochastic Frank–Wolfe was pursued in [24] where their proof relies on what is called Assumption P.8, in our notation this requires the sum $\sum_{n=0}^{\infty}\lambda_n G^n < +\infty$ to be finite almost surely, where λ_n is chosen deterministically. To apply this algorithm in our setting we would therefore need to bound the magnitude of G^n relative to the *a priori* choice of λ_n. With Proposition 4.2, we overcome this limitation with the linesearch for λ, so the only remaining requirement is to guarantee a uniform probability of achieving a good search direction.

4.3. Back to the dynamic inverse problem

We consider again the setting of Section 3. The function E is convex on D, and we note that, for each $\sigma \in D$, its directional derivative is given by

$$\forall \nu \in D, \quad \mathrm{E}'(\sigma; \nu-\sigma) = \int_\Gamma \left(\mathrm{F}'(\sigma)(\gamma) + \mathrm{W}'(\sigma)(\gamma)\right)\mathrm{d}(\nu-\sigma)(\gamma),$$

with $\mathrm{W}'(\sigma) = [\gamma \mapsto w(\gamma)]$ and

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

In particular, note that we can write $\mathrm{E}'(\sigma;\mu-\sigma)=\int_\Gamma \mathrm{E}'(\sigma)\,\mathrm{d}[\mu-\sigma]$ with $\mathrm{E}'(\sigma)\colon\Gamma\to\mathbb{R}$. This structure enables us to prove that E satisfies the major requirement of Section 4, that $\inf_{\mu\in D}\mathrm{E}'(\sigma;\mu-\sigma)>-\infty$ for all $\sigma\in D$. To do this, we show that the infimum is always achieved by an extreme point of D.

Lemma 4.4. *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 $\mathrm{E}'(\sigma;\cdot-\sigma)$ is lower semi-continuous and coercive on D. Moreover it has minimiser of the form*

$$\mathrm{LMO}(\sigma)\in\{0\}\cup\left\{\varphi(\gamma^*)^{-1}\delta_{\gamma^*}\mid\gamma^*\in\operatorname*{argmin}_{\gamma\in\Gamma}\frac{\eta(\gamma)+w(\gamma)}{\varphi(\gamma)}\right\}. \tag{4.15}$$

where $\eta(h,\xi)\overset{\text{def.}}{=}\sum_{j=0}^{T}h(t_j)\eta_j(\xi(t_j))\in\mathrm{C_b}(\Gamma)$.

Proof. Recall the definition of LMO,

$$\mathrm{LMO}(\sigma)\in\operatorname*{argmin}_{\widetilde\sigma\in D}\widetilde{\mathrm{E}}(\widetilde\sigma)\quad\text{where}\quad\widetilde{\mathrm{E}}(\widetilde\sigma)\overset{\text{def.}}{=}\int_\Gamma(\eta+w)\,\mathrm{d}\widetilde\sigma. \tag{4.16}$$

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

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

$$\sigma^*\in\mathrm{Ext}\,(D)\quad\text{such that}\quad\widetilde{\mathrm{E}}(\sigma^*)=\inf_{\widetilde\sigma\in D}\widetilde{\mathrm{E}}(\widetilde\sigma). \tag{4.17}$$

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

$$U\overset{\text{def.}}{=}\left\{\widetilde\sigma\in D\quad\middle|\quad\widetilde{\mathrm{E}}(\widetilde\sigma)\le 1\right\} \tag{4.18}$$

is compact by Theorem B.7, convex, and non-empty because $0\in D$, $\widetilde{\mathrm{E}}(0)=0$. Application of Bauer's principle now gives a minimizer $\sigma^*\in\mathrm{Ext}\,(U)$. To complete the proof we will confirm $\sigma^*\in\mathrm{Ext}\,(D)$.

Suppose for contradiction that there exists $\sigma_0,\sigma_1\in D\setminus\{\sigma^*\}$ with $\sigma^*\in\,]\sigma_0,\sigma_1[$. Since the restriction of $\widetilde{\mathrm{E}}$ to $]\sigma_0,\sigma_1[\subset D$ is linear and reaches its minimum at σ^*, it is constant on $]\sigma_0,\sigma_1[$. Hence $]\sigma_0,\sigma_1[\subset U$, which contradicts $\sigma^*\in\mathrm{Ext}\,(U)$. Therefore $\sigma^*\in\mathrm{Ext}\,(D)$.

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

$$\mathrm{Ext}\,(D)=\{0\}\cup\left\{\,\varphi(\gamma)^{-1}\delta_\gamma\mid\varphi(\gamma)<+\infty\,\right\}, \tag{4.19}$$

so we deduce (4.15) as required. □

This lemma is also useful for applying the stochastic Frank–Wolfe algorithm, Lemma 4.3. For each $n\in\mathbb{N}$, let $z^n\sim Z$ be an independent random discretisation of the domain $\widetilde\Gamma=([-1,1]\times\overline\Omega)^{T+1}$ (following the discrete-time formulation in Sect. 3.2). Continuing with the notation $\widetilde{\mathrm{E}}=\mathrm{E}'(\sigma^{n-1})$ from (4.16), we will use the discrete minimiser denoted

$$\mu^n=\overline{\mathrm{LMO}}(\sigma^{n-1},z^n)\in\{0\}\cup\left\{\varphi(\gamma_d^*)^{-1}\delta_{\gamma_d^*}\mid\gamma_d^*\in\operatorname*{argmin}_{\gamma\in z^n}\widetilde{\mathrm{E}}\left(\varphi(\gamma)^{-1}\delta_\gamma\right)\right\}. \tag{4.20}$$

The remaining requirements for applying both Frank–Wolfe variants simplify greatly in the Benamou–Brenier example where $\varphi > 0$ is a constant, curves have constant mass, and $\Omega =]0,1[^d$ (*i.e.* $z^n \subset [0,1]^{d(T+1)}$). An immediate consequence of this is uniform boundedness, as $\int_\Gamma \mathrm{d}\sigma = \sum_{j=0}^T \left\|(e_{t_j})_\sharp \sigma\right\| \leq \varphi^{-1}$ for all $\sigma \in D$. Because E is quadratic and A_j bounded (see (3.1), $(A_j\rho)_i = \int_{\overline{\Omega}} a_i^j \,\mathrm{d}\rho$), the curvature bound follows immediately:

$$C_{\mathrm{E}} = \sup_{\sigma,\sigma' \in D} \sum_{j=0}^T \left\|A_j(e_{t_j})_\sharp[\sigma - \sigma']\right\|_2^2 \leq \left(\sup_{\sigma,\sigma' \in D} \sum_{j=0}^T \|A_j\| \left\|(e_{t_j})_\sharp[\sigma - \sigma']\right\|\right)^2 \leq \varphi^{-2} \max_{i,j} \left\|a_i^j\right\|_\infty < +\infty. \quad (4.21)$$

Finally we must show that $\mathrm{gap}(\sigma^{\mathrm{n}-1}; \mu^{\mathrm{n}})$ is uniformly small, independently of σ^{n-1}. Ignoring the $\sigma^* = 0$ case, this quantity can be written

$$\mathrm{gap}(\sigma^{\mathrm{n}-1}; \mu^{\mathrm{n}}) = \int_\Gamma \widetilde{\mathrm{E}} \,\mathrm{d}[\mu^{\mathrm{n}} - \mathrm{LMO}(\sigma^{\mathrm{n}-1})] \leq \varphi^{-1} \min_{\gamma_{\mathrm{d}}^* \in z^{\mathrm{n}}} \max_{\gamma^* \in \Gamma} [\widetilde{\mathrm{E}}(\gamma_{\mathrm{d}}^*) - \widetilde{\mathrm{E}}(\gamma^*)] \quad (4.22)$$

where, from (3.7) and (3.15), for all $\xi \in \overline{\Omega}^{T+1}$ we have

$$\widetilde{\mathrm{E}}(1, \xi) = \left[\sum_{i=1}^m \sum_{j=0}^T \left(\int_\Omega a_i^j(x) \,\mathrm{d}(e_{t_j})_\sharp \sigma^{n-1}(x) - b_j\right)^\top a_i^j(\xi_j)\right] + \alpha + \frac{\beta}{2} \sum_{j=1}^T \frac{|\xi_j - \xi_{j-1}|^2}{t_j - t_{j-1}} \quad (4.23)$$

for some $\alpha, \beta > 0$. If moreover A_j are Lipschitz, then $\widetilde{\mathrm{E}}$ is also uniformly Lipschitz independently of σ^{n-1}. Combining this uniform Lipschitz property with, for example, uniformly sampled discrete grids z^n guarantees the uniform bound for Lemma 4.3.

4.4. Choice of constraint

In this section, we will briefly discuss the main difference between our formulation of the Benamou–Brenier example and that implemented in [5]. Although the construction in [5] looks very different, it can also be seen as Frank–Wolfe/Generalised Conditional Gradient approach to minimise the same function E (this equivalence is confirmed in Sect. 6). The parallel result to Lemma 4.15 is [5], Proposition 3.6 which shows that they are also incrementally adding new curves ξ to the support of the reconstruction σ each iteration. Both implementations actually use a stochastic variant of Frank–Wolfe, but we will discuss the classical version for simplicity.

Recall from Remark 3.2, that the un-constrained energy is

$$\forall \sigma \in \mathcal{M}^+(\Gamma), \quad \mathrm{E}(\sigma) = \frac{1}{2} \sum_{j=0}^T \left\|A_j(e_{t_j})_\sharp \sigma - b_j\right\|_2^2 + \int_\Gamma w \,\mathrm{d}\sigma \quad (4.24)$$

where $w(\xi) = \alpha + \frac{\beta}{2} \int_0^1 |\xi'(t)|^2 \,\mathrm{d}t$ for all $\xi \in \Gamma$. At iteration n, let $\eta(\xi) \overset{\text{def.}}{=} \sum_{i=1}^m \sum_{j=0}^T \left(A_j(e_{t_j})_\sharp \sigma^n - b_j\right)_i a_i^j(\xi(t_j))$ denote the linearisation of the fidelity term where operators A_j are represented by kernels $a_i^j \in \mathrm{C_b}(\overline{\Omega})$. The standard Frank–Wolfe procedure consists in iteratively minimising the function

$$\min_{\sigma \in D} \int_\Gamma [\eta(\xi) + w(\xi)] \,\mathrm{d}\sigma. \quad (4.25)$$

on a constraint set $D \subset \mathcal{M}^+(\Gamma)$ to yield descent directions. Different sets $D \supset \operatorname{argmin}_{\sigma\in\mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma)$ can change the minimisers of (4.25) without necessarily changing the original problem (4.24). A first choice is to consider

$$D_1 \overset{\text{def.}}{=} \left\{ \sigma \in \mathcal{M}^+(\Gamma) \mid \int_\Gamma w \,\mathrm{d}\sigma \leq \mathrm{E}(0) \right\}, \tag{4.26}$$

which does contain the minimisers of (4.24) since they must satisfy $\int_\Gamma w \,\mathrm{d}\sigma \leq \mathrm{E}(\sigma) \leq \mathrm{E}(0)$. By Lemma B.5, the set of extreme points of D_1 is $\{0\} \cup \left\{ \frac{\mathrm{E}(0)}{w(\xi)}\delta_\xi \mid \xi \in \Gamma \right\}$, which yields the descent direction

$$\xi^{n+1} \in \operatorname*{argmin}_{\xi\in\Gamma} \frac{\eta(\xi) + w(\xi)}{w(\xi)} = \operatorname*{argmin}_{\xi\in\Gamma} \frac{\eta(\xi)}{w(\xi)}. \tag{4.27}$$

That is precisely the descent direction used in [5], (4.30). On the other hand, we propose to use

$$D_2 \overset{\text{def.}}{=} \left\{ \sigma \in \mathcal{M}^+(\Gamma) \mid \int_\Gamma \alpha \,\mathrm{d}\sigma \leq \mathrm{E}(0) \right\}, \tag{4.28}$$

which also contains the minimisers of (4.24), 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 B.5, we see that the set of extreme points of D_2 is $\{0\} \cup \left\{ \frac{\mathrm{E}(0)}{\alpha}\delta_\xi \mid \xi \in \Gamma \right\}$, which yields

$$\xi^{n+1} \in \operatorname*{argmin}_{\xi\in\Gamma} \frac{\eta(\xi) + w(\xi)}{1} = \operatorname*{argmin}_{\xi\in\Gamma} \eta(\xi) + w(\xi). \tag{4.29}$$

We found (4.29) more convenient than (4.27) as it is amenable to dynamic programming techniques (see Sect. 5). The whole minimisation algorithm is summarised in Algorithm 2.

Remark 4.5. In the classical Frank–Wolfe algorithm D must be compact. In measure spaces, compactness is equivalent to boundedness and tightness (Prokhorov's theorem). In both cases, boundedness comes from $\alpha > 0$, and tightness comes from the fact that w has compact sub-levelsets. The set D_1 is compact by Lemma B.3. On the other hand, D_2 is only bounded, but sub-levelsets of (4.25) are still compact, as argued in Lemma 4.4.

Algorithm 2 Frank–Wolfe algorithm for the Benamou–Brenier example

1: Choose $\sigma^0 = 0$, $n \leftarrow 0$
2: **repeat**
3: Compute ξ^{n+1} according to (4.27) or (4.29) ▷ Linear oracle step
4: Choose $\mu^{n+1} \propto \delta_{\xi^{n+1}}$ such that $\mu^{n+1} \in \operatorname{Ext}(D)$ ▷ Choice of scaling
5: $\sigma^{n+1} \leftarrow (1-\lambda_n)\sigma^n + \lambda_n\mu^{n+1}$ ▷ Some stepsize $\lambda_n \in [0,1]$
6: $n \leftarrow n+1$
7: **until** converged

5. Dynamical programming method

Throughout Section 4 we have shown that we can find a minimising sequence to E from Section 3 by repeatedly evaluating a simplified linear oracle. In particular, we compute the minimiser from Lemma 4.4, but over a discretised domain. In Section 3.2 we motivated discretising in time to $\widetilde{\Gamma} = ([-1,1]\times\overline{\Omega})^{T+1}$ without loss of precision, now we also discretise in space using the domain $\Lambda \overset{\text{def.}}{=} \prod_{j=0}^{T} \Lambda_j$ for some discrete sets $\Lambda_j \subset [-1,1]\times\overline{\Omega}$ ("grids" in the mass-location space). Whilst Λ is much smaller than $\widetilde{\Gamma}$, naive computation time for γ^* still scales

exponentially in T. We propose to compute this discrete minimiser efficiently using dynamical programming, for which we require two final simplifications:

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

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

Remark 5.1. For the Benamou–Brenier example, these assumptions are satisfied by the choice

$$\varphi_0 = \frac{\alpha}{\mathrm{E}(0)}, \qquad \mathrm{step}_j(\gamma(t_{j-1}), \gamma(t_j)) = \alpha(t_j - t_{j-1}) + \frac{\beta}{2} \frac{|\xi(t_j) - \xi(t_{j-1})|^2}{t_j - t_{j-1}} \tag{5.1}$$

for all $\gamma = (h, \xi) \in \Gamma$. We will see later in Section 6.2 that these assumptions are also satisfied in the Wasserstein–Fisher–Rao example. 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.

Under assumptions (A1) and (A2), the discretised optimisation problem (4.15) can be greatly simplified to

$$\gamma^* \in \operatorname*{argmin}_{Y} \left\{ \sum_{j=0}^{T} \eta_j(Y_j) + \sum_{j=1}^{T} \mathrm{step}_j(Y_{j-1}, Y_j) \mid Y \in \prod_{j=0}^{T} \Lambda_j \right\} \tag{5.2}$$

$$\text{where} \quad \forall j = 0, \dots, T, \qquad \eta_j(h_j, \xi_j) \overset{\text{def.}}{=} h_j [A_j^* \nabla \mathrm{F}_j(A_j(e_{t_j})_\sharp \sigma)](\xi_j). \tag{5.3}$$

This minimisation problem can now be formulated as computing the minimal path on a 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. Algorithms for computing minimal paths on directed acyclic graphs are well studied, the complexity bound is given below.

Theorem 5.2 ([25], Sect. 24.2)**.** *Suppose assumptions* (A1) *and* (A2) *hold and* $|\Lambda_j| \leq N$ *for each* j*, then* γ^* *can be computed with complexity*

$$\textit{total time} = O\left(N \sum_{j=0}^{T} \mathrm{cost}(\eta_j) + N^2 \sum_{j=1}^{T} \mathrm{cost}(\mathrm{step}_j) \right) \tag{5.4}$$

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

Remark 5.3. In graph terminology, the two terms of (5.4) represent the number of vertices plus the number of edges respectively, 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 $O(NT + N^2T^{1-d}V^d)$.

Finally we outline how the minimal path can be computed efficiently in our specific example by using dynamic programming. To do so, define the truncated energies and minimal paths:

$$\widetilde{\mathrm{E}}_J(Y) \overset{\text{def.}}{=} \sum_{j=0}^{J} \eta_j(Y_j) + \sum_{j=1}^{J} \mathrm{step}_j(Y_{j-1}, Y_j), \tag{5.5}$$

$$Y^*[y,J] \in \operatorname*{argmin}_Y \left\{ \widetilde{\mathrm{E}}_J(Y) \mid Y \in \prod_{j=0}^{J} \Lambda_j,\ Y_J = y \right\} \tag{5.6}$$

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

$$\gamma^* = Y^*[y^*, T], \qquad y^* \in \operatorname*{argmin}_{y \in \Lambda_T} \widetilde{\mathrm{E}}_T(Y^*[y,T]). \tag{5.7}$$

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

$$\min_{\substack{Y \in \prod_{j=0}^{J} \Lambda_j \\ Y_J = y}} \widetilde{\mathrm{E}}_J(Y) = \min_{Y \in \prod_{j=0}^{J} \Lambda_j} \left\{ \widetilde{\mathrm{E}}_{J-1}(Y) + \eta_J(y) + \mathrm{step}_J(Y_{J-1}, y) \right\} \tag{5.8}$$

$$= \eta_J(y) + \min_{y' \in \Lambda_{J-1}} \left\{ \mathrm{step}_J(y', y) + \min_{\substack{Y \in \prod_{j=0}^{J-1} \Lambda_j \\ Y_{J-1} = y'}} \widetilde{\mathrm{E}}_{J-1}(Y) \right\}. \tag{5.9}$$

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

$$Y^*[y,J] = \begin{pmatrix} Y^*[y', J-1] & y \end{pmatrix} \quad \text{for} \quad y' \in \operatorname*{argmin}_{y' \in \Lambda_{J-1}} \left[\mathrm{step}_J(y', y) + \widetilde{\mathrm{E}}_{J-1}(Y^*[y', J-1]) \right]. \tag{5.10}$$

For each y and J each of these steps requires $O(N)$ computation, confirming the global complexity of $O(N^2 T)$.

6. The unbalanced Wasserstein–Fisher–Rao example

The two numerical examples considered in this work use the Benamou–Brenier and Wasserstein–Fisher–Rao (WFR) penalties. The former has already been discussed in previous remarks and is a limiting case of the WFR penalty so we will not discuss it in further detail here.

The penalty considered in [4] is $\mathcal{W}\colon \mathcal{M}^+([0,1] \times \overline{\Omega}) \to \mathbb{R} \cup \{+\infty\}$ such that for all $\rho \in \mathcal{M}^+([0,1] \times \overline{\Omega})$,

$$\mathcal{W}(\rho) \stackrel{\text{def.}}{=} \inf_{\substack{v \in \mathrm{L}^2_\rho([0,1] \times \overline{\Omega}; \mathbb{R}^d) \\ g \in \mathrm{L}^2_\rho([0,1] \times \overline{\Omega})}} \left\{ \int_{[0,1] \times \overline{\Omega}} \left[\alpha + \frac{\beta}{2} |v|^2 + \frac{\beta \delta^2}{2} g^2 \right] \mathrm{d}\rho \ \text{ s.t. } \ \partial_t \rho + \operatorname{div}(v\rho) = g\rho \right\} \tag{6.1}$$

for some $\alpha, \beta, \delta > 0$, and the continuity equation is satisfied in the sense of (1.14). This leads to the energy

$$\mathcal{E}(\rho) = \frac{1}{2} \sum_{j=0}^{T} \left\| A_j \rho_{t_j} - b_j \right\|_2^2 + \mathcal{W}(\rho) \tag{6.2}$$

where the properties of A_j are as stated in (3.1). Much is already known about minimisers of this energy.

Theorem 6.1 ([4], Thms. 4.2, 6.4)**.** *Let $\alpha, \beta, \delta > 0$ and*

$$\Gamma \stackrel{\text{def.}}{=} \left\{ (h, \xi)\colon\ [0,1] \to [0,1] \times \overline{\Omega} \mid \sqrt{h} \in \mathrm{AC}^2([0,1]),\ \sqrt{h}\xi \in \mathrm{AC}^2([0,1]; \mathbb{R}^d) \right\}. \tag{6.3}$$

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

$$\mathcal{W}(\rho) = \int_{[0,1]\times\overline{\Omega}} \left[\alpha + \frac{\beta}{2}|v|^2 + \frac{\beta\delta^2}{2}g^2\right] \mathrm{d}\rho, \qquad \partial_t\rho + \mathrm{div}(v\rho) = g\rho, \tag{6.4}$$

and

$$\begin{aligned} &\xi'(t) = v(t,\xi(t)) \quad \textit{for} \quad \textit{a.e. } t \in \{h>0\} \textit{ and } \sigma\textit{-a.e. } (h,\xi),\\ &h'(t) = g(t,\xi(t))h(t) \quad \textit{for} \quad \textit{a.e. } t \in [0,1] \textit{ and } \sigma\textit{-a.e. } (h,\xi). \end{aligned} \tag{6.5}$$

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

$$\textit{for some } a_i \geq 0,\ (h^i,\xi^i) \in \Gamma, \qquad \forall t \in [0,1]\,, \qquad \rho^*_t = \sum_{i=1}^{m(T+1)} a_i h^i(t)\delta_{\xi^i(t)}. \tag{6.6}$$

Proof. Fix $\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})$ with $\mathcal{W}(\rho) < +\infty$. The only aspect not directly covered by [4] is the existence of the minimal pair (v,g). To confirm this, note that the set

$$\left\{ \begin{array}{c} v \in \mathrm{L}^2_\rho([0,1]\times\overline{\Omega};\mathbb{R}^d) \\ g \in \mathrm{L}^2_\rho([0,1]\times\overline{\Omega};\mathbb{R}) \end{array} \;\middle|\; \int_{[0,1]\times\overline{\Omega}} \left[\alpha + \frac{\beta}{2}|v|^2 + \frac{\beta\delta^2}{2}g^2\right] \mathrm{d}\rho \leq \mathcal{W}(\rho)+1,\ \partial_t\rho + \mathrm{div}(v\rho) = g\rho \right\} \tag{6.7}$$

is bounded, hence compact in the weak topology of L^2_ρ. There is therefore a weakly-convergent sequence converging to a point (v,g) which achieves the desired infimum in (6.1). As the weak form of the continuity equation is preserved under weak limits in (v,g), the triplet (ρ,v,g) also satisfies the continuity equation. □

6.1. Reformulation in the space of measures on paths

The results of Theorem 6.1 highlight the close relationship between the representations ρ and σ, *i.e.* dynamical measures and measures on paths. We now reformulate the energy $\mathcal{E}$ into an equivalent energy $\mathrm{E}\colon \mathcal{M}^+([0,1]\times\overline{\Omega}) \to \mathbb{R}\cup\{+\infty\}$ of the form in Section 3. This energy can be written

$$\forall \sigma \in \mathcal{M}^+(\Gamma), \qquad \mathrm{E}(\sigma) \overset{\text{def.}}{=} \frac{1}{2}\sum_{j=0}^{T} \left\| A_j(e_{t_j})_\sharp\sigma - b_j \right\|_2^2 + \mathrm{W}(\sigma), \tag{6.8}$$

$$\mathrm{W}(\sigma) \overset{\text{def.}}{=} \int_\Gamma \int_0^1 \left[\alpha + \frac{\beta}{2}|\xi'|^2 + \frac{\beta\delta^2}{2}\left(\frac{h'}{h}\right)^2\right] h \,\mathrm{d}t\,\mathrm{d}\sigma(h,\xi). \tag{6.9}$$

Remark 6.2. 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 (6.9) makes sense when regarded as

$$w(h,\xi) \overset{\text{def.}}{=} \int_{\{h>0\}} \left[\alpha + \frac{\beta}{2}\left|\left(\frac{\sqrt{h}\xi}{\sqrt{h}}\right)'(t)\right|^2\right] h + 2\beta\delta^2\left|(\sqrt{h})'(t)\right|^2 \mathrm{d}t. \tag{6.10}$$

We will use the operator $\Theta\colon \mathcal{M}(\Gamma) \to \mathcal{M}([0,1]\times\overline{\Omega})$ introduced in Section 2 to map between σ and ρ. The formula in (6.9) comes from [4], (3.9), but it was only shown that $\mathrm{W}(\delta_\gamma) = \mathcal{W}(\Theta(\delta_\gamma))$. We cannot hope that this holds on the whole of $\mathcal{M}(\Gamma)$ because Θ is not a one-to-one mapping, $\mathcal{M}(\Gamma)$ is a much larger space than $\mathcal{M}([0,1]\times\overline{\Omega})$. The next lemma is needed to confirm the equivalence of minimisers between E and $\mathcal{E}$.

Lemma 6.3. *Choosing $\varphi(\gamma) = \frac{\alpha}{\mathrm{E}(0)}$, the function* E *is lower semi-continuous with compact sub-levelsets with sparse minimisers in $D = \left\{ \sigma \in \mathcal{M}^+(\Gamma) \mid \int_\Gamma \varphi \,\mathrm{d}\sigma \leq 1 \right\}$ (Thm. 3.1). Also, for any $\rho \in \mathcal{M}^+([0,1]\times\overline{\Omega})$,*

$$\mathcal{W}(\rho) = \min_{\sigma\in\mathcal{M}^+(\Gamma)} \left\{ \mathrm{W}(\sigma) \mid \Theta(\sigma) = \rho \right\} \tag{6.11}$$

where $\min\emptyset = +\infty$. We conclude that

$$\min\left\{ \mathrm{E}(\sigma) \mid \sigma \in \mathcal{M}^+(\Gamma) \right\} = \min\left\{ \mathcal{E}(\rho) \mid \rho \in \mathcal{M}^+([0,1]\times\overline{\Omega}) \right\}, \tag{6.12}$$

and

$$\operatorname*{argmin}_{\sigma\in\mathcal{M}^+(\Gamma)} \mathrm{E}(\sigma) = \left\{ \sigma \in \mathcal{M}^+(\Gamma) \mid \Theta(\sigma) \in \operatorname*{argmin}_{\rho\in\mathcal{M}^+([0,1]\times\overline{\Omega})} \mathcal{E}(\rho) \text{ and } \mathcal{W}(\Theta(\sigma)) = \mathrm{W}(\sigma) \right\}. \tag{6.13}$$

Proof. The only requirement for Theorem 3.1 which is not explicitly assumed is that the function w is lower semi-continuous with compact sub-levelsets. This is proved in the appendix in Lemma B.4.

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 6.1 satisfying $\rho = \Theta(\widehat{\sigma})$, then

$$\begin{aligned}
\mathcal{W}(\rho) &= \int_{[0,1]\times\overline{\Omega}} \left[\alpha + \frac{\beta}{2}|v|^2 + \frac{\beta\delta^2}{2} g^2 \right] \mathrm{d}\rho && \text{Theorem 6.1} && (6.14)\\
&= \int_\Gamma \int_0^1 \left[\alpha + \frac{\beta}{2}|v(t,\xi(t))|^2 + \frac{\beta\delta^2}{2} g(t,\xi(t))^2 \right] h(t)\,\mathrm{d}t\,\mathrm{d}\widehat{\sigma}(h,\xi) && \text{Theorem 2.2(2)} && (6.15)\\
&= \int_\Gamma \int_0^1 \left[\alpha + \frac{\beta}{2}|\xi'(t)|^2 + \frac{\beta\delta^2}{2} \left(\frac{h'(t)}{h(t)} \right)^2 \right] h(t)\,\mathrm{d}t\,\mathrm{d}\widehat{\sigma}(h,\xi) = \mathrm{W}(\widehat{\sigma}). && (6.5) && (6.16)
\end{aligned}$$

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[4], Lemma A.6, by [26], Theorem 5 there exists a collection $\{(c_i,\psi_i)\}_{i\in I} \subset \mathbb{R}\times\mathrm{C}([0,1]\times\overline{\Omega})$ such that

$$\forall \rho \in \mathcal{M}^+([0,1]\times\overline{\Omega}), \qquad \mathcal{W}(\rho) = \sup_{i\in I} \left(c_i + \int_{[0,1]\times\overline{\Omega}} \psi_i \,\mathrm{d}\rho \right). \tag{6.17}$$

As $\mathcal{W}$ is positively homogeneous, also $c_i = 0$. 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

$$\begin{aligned}
\int_{[0,1]\times\overline{\Omega}} \psi_i \,\mathrm{d}\Theta(\widehat{\sigma}) = \int_\Gamma \int_0^1 h(t)\psi_i(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\widehat{\sigma} &\leq \int_\Gamma \sup_{j\in I} \int_0^1 h(t)\psi_j(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\widehat{\sigma}\\
&= \int_\Gamma \left[\sup_{j\in I} \int_{[0,1]\times\overline{\Omega}} \psi_j \,\mathrm{d}\Theta(\delta_\gamma) \right] \mathrm{d}\widehat{\sigma} = \int_\Gamma \mathcal{W}(\Theta(\delta_\gamma))\,\mathrm{d}\widehat{\sigma}. \qquad (6.18)
\end{aligned}$$

It is shown in [4], Proposition 3.9 that $\mathcal{W}(\Theta(\delta_\gamma)) = w(\gamma)$ from (6.10). Now taking a supremum over $i \in I$ gives

$$\mathcal{W}(\Theta(\widehat{\sigma})) \leq \int_\Gamma \mathcal{W}(\Theta(\delta_\gamma))\,\mathrm{d}\widehat{\sigma} = \int_\Gamma w(\gamma)\,\mathrm{d}\widehat{\sigma} = \mathrm{W}(\widehat{\sigma}). \tag{6.19}$$

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 confirms (6.11) when combined with the "$\geq$" result.

Finally we consider the equivalence of minimums and minimisers. Notice that the equality $\rho = \Theta(\sigma)$ implies the equality of the times slices ρ_t and $(e_t)_\sharp\sigma$, for $t \in [0,1]$, and thus

$$\frac{1}{2}\sum_{j=0}^{T} \left\|A_j\rho_{t_j} - b_j\right\|_2^2 = \frac{1}{2}\sum_{j=0}^{T} \left\|A_j(e_{t_j})_\sharp\sigma - b_j\right\|_2^2. \tag{6.20}$$

As a result, (6.11) implies that

$$\mathcal{E}(\rho) = \min_{\sigma\in\mathcal{M}^+(\Gamma)} \{\,\mathrm{E}(\sigma) \mid \Theta(\sigma) = \rho\,\},$$

and thus

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

If $\sigma \in \operatorname{argmin}\mathrm{E}$, then

$$\mathcal{E}(\Theta(\sigma)) \leq \mathrm{E}(\sigma) = \min\mathrm{E} = \min\mathcal{E},$$

so that $\Theta(\sigma) \in \operatorname{argmin}_{\rho\in\mathcal{M}^+([0,1]\times\overline{\Omega})} \mathcal{E}(\rho)$ and $\mathcal{W}(\Theta(\sigma)) = \mathrm{W}(\sigma)$, which yields the first inclusion in (6.13). Conversely, if σ belongs to the r.h.s. of (6.13), then $\mathrm{E}(\sigma) = \min_\rho \mathcal{E}(\rho) = \min\mathrm{E}$, so that the converse inclusion holds. □

6.2. Discrete-time formulation

Problems of the form (6.8) also have a discrete-time structure inherited from discrete-time data. The same argument from Section 3.2 is valid for the WFR penalty, although it is very hard to find the explicit form. Analytically the WFR penalty can be expressed in the form required for Assumption A2. In particular, for

$$\mathrm{step}_j(\gamma_{j-1},\gamma_j) = \inf_{\gamma\in\Gamma} \left\{ \int_{t_{j-1}}^{t_j} \left[\alpha + \frac{\beta}{2}|\xi'(t)|^2 + \frac{\beta\delta^2}{2}\left(\frac{h'(t)}{h(t)}\right)^2\right] h(t)\,\mathrm{d}t \mid \gamma = (h,\xi),\ \begin{array}{c}\gamma(t_{j-1}) = \gamma_{j-1},\\ \gamma(t_j) = \gamma_j\end{array} \right\}, \tag{6.21}$$

geodesics of the WFR penalty satisfy

$$w(\gamma) = \int_0^1 \left[\alpha + \frac{\beta}{2}|\xi'|^2 + \frac{\beta\delta^2}{2}\left(\frac{h'}{h}\right)^2\right] h\,\mathrm{d}t = \sum_{j=1}^{T} \mathrm{step}_j(\gamma(t_{j-1}),\gamma(t_j)). \tag{6.22}$$

This formula can be verified with the Euler–Lagrange equation. The key point is that the left-hand side only involves up to first order derivatives so only the zeroth order constraints (*i.e.* $\gamma(t_j) = \gamma_j$) are needed to interpolate

on intervals $]t_{j-1}, t_j[$. The exact shape or formulae for the WFR-geodesics is not so convenient as for the Benamou–Brenier penalty ($\delta \to +\infty$), although for $\alpha = 0$ it is known from [27], Theorem 5.6

$$\mathrm{step}_j((h_{j-1},\xi_{j-1}),(h_j,\xi_j)) = \frac{4\beta\delta^2}{t_j - t_{j-1}} \left[\tfrac{h_j+h_{j-1}}{2} - \sqrt{h_j h_{j-1}} \cos\left(\min\left(\tfrac{|\xi_j - \xi_{j-1}|}{2\delta}, \pi \right) \right) \right]. \tag{6.23}$$

More details can be found in Corollary 4.1(i) of the preprint (https://arxiv.org/pdf/1506.06430v2.pdf) of [28]. In summary, the map $\gamma \mapsto \mathrm{argmin}_{\widetilde{\gamma}\in\Gamma} \{ w(\widetilde{\gamma}) \mid \widetilde{\gamma}(t_j) = \gamma(t_j) \}$ is continuous, $\widetilde{h}(t)$ is a simple quadratic on each $]t_j, t_{j+1}[$, and $\widetilde{\xi}(t)$ follows a straight line between $\xi(t_j)$ and $\xi(t_{j+1})$ with speed varying like arctan.

7. Numerical results

For numerical experiments we implement variants of Algorithm 1 using the linear-oracle strategy discussed in Section 5. 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 [5], Section 5.1.5 as a "multistart" parameter. The idea is that the approximation of optimal curves (*i.e.* $Y^*[\cdot, T]$) 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 $\overline{\Omega}$. Assuming T is fixed, Theorem 5.2 states 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 $O((N^d M)^2 k^{-1})$ per curve (see [29, 30]).

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

$$\Lambda_j \overset{\text{def.}}{=} \{ (h,x) \mid h \in H,\ x \in X \}. \tag{7.1}$$

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

$$\Lambda_j = \Lambda_j(\widetilde{N}) \overset{\text{def.}}{=} \left\{ (h,(x_1,x_2)) \mid h \in \{0, M^{-1}, \ldots, 1\},\ x_1, x_2 \in \{0, \widetilde{N}^{-1}, \ldots, 1\} \right\} \tag{7.2}$$

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 1.1, 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 Lemma 4.3. 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

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

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 $D_d \overset{\text{def.}}{=} \left\{ \sigma \in D \mid \operatorname{supp}(\sigma) \subset \prod_{j=0}^{T} \Lambda_j(N) \right\}$ be the discretisation of D. Expanding on the linesearch computation, due to the convexity and smoothness of E,

$$\forall \lambda \in [0,1]\,, \qquad \mathrm{E}((1-\lambda)\sigma^n + \lambda\mu^{n+1}) - \mathrm{E}(\sigma^n) = \lambda \int_\Gamma (\mathrm{F}'(\sigma^n) + w)\,\mathrm{d}[\mu^{n+1} - \sigma^n] + O(\lambda^2). \tag{7.3}$$

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 [12]), can therefore be stated

$$\mathrm{E}(\sigma^n) - \min_{\sigma \in D_d} \mathrm{E}(\sigma) \leq \sup_{\sigma \in D_d} \int_\Gamma (\mathrm{F}'(\sigma^n) + w)\,\mathrm{d}[\sigma^n - \sigma] = \int_\Gamma (\mathrm{F}'(\sigma^n) + w)\,\mathrm{d}[\sigma^n - \mu^{n+1}] \leq 0. \tag{7.4}$$

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 $\eta_j = A_j^* \nabla \mathrm{F}_j(A_j(e_{t_j})_\sharp \sigma^n) \in \mathrm{C}(\overline{\Omega})$ for $j = 0, \dots, T$, and
    $\widetilde{\mathrm{E}}((h,\xi)) = \sum_{j=0}^{T} \eta_j(\xi_j) h_j + \sum_{j=1}^{T} \mathrm{step}_j((h_{j-1}, \xi_{j-1}), (h_j, \xi_j))$
  Choose $\Lambda_j \subset [0,1] \times \overline{\Omega}$, $j = 0, \dots T$ ▷ the discrete mesh
  Compute $Y^*[y,T] \in \prod_{j=0}^{T} \Lambda_j$ for all $y \in \Lambda_T$, as in (5.10) ▷ discrete optimal paths
  Choose $\widetilde{\gamma}^1, \dots, \widetilde{\gamma}^k \in \mathrm{Image}(Y^*[\cdot, T])$ with least energy in $\widetilde{\mathrm{E}}$ ▷ select best k endpoints
  Find $\gamma^1, \dots \gamma^k \in [0,1] \times \overline{\Omega}$ 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 $\widetilde{\mathrm{E}}(\gamma^1) \leq \widetilde{\mathrm{E}}(\gamma^2), \dots$
  for $i = 1, \dots, k$ **do**
    $\lambda \leftarrow \operatorname{argmin}_{\lambda \in [0,1]} \mathrm{E}((1-\lambda)\sigma + \lambda\varphi(\gamma^i)^{-1}\delta_{\gamma^i})$ ▷ exact linesearch
    $\sigma \leftarrow (1-\lambda)\sigma + \lambda\varphi(\gamma^i)^{-1}\delta_{\gamma^i}$
  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

7.1. Benamou–Brenier example

First we compare directly with the numerical results presented in [5] for the model discussed in Section 1.1. In particular, the energy we seek to minimise is $\mathrm{E}\colon \mathcal{M}^+(\Gamma) \to]-\infty, +\infty]$ defined by

$$\mathrm{E}(\sigma) = \frac{1}{2} \sum_{j=0}^{T} \left\| A_j(e_{t_j})_\sharp \sigma - b_j \right\|_2^2 + \int_\Gamma \int_0^1 \left(\alpha + \frac{\beta}{2} |\xi'(t)|^2 \right) \mathrm{d}t\,\mathrm{d}\sigma(h,\xi) \tag{7.5}$$

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

$$\Gamma \overset{\text{def.}}{=} \left\{ (h,\xi) \in \{1\} \times \mathrm{AC}^2([0,1]\,; \overline{\Omega}) \mid \xi' \text{ is constant on }]t_{j-1}, t_j[,\ 1 \leq j \leq T \right\}, \tag{7.6}$$

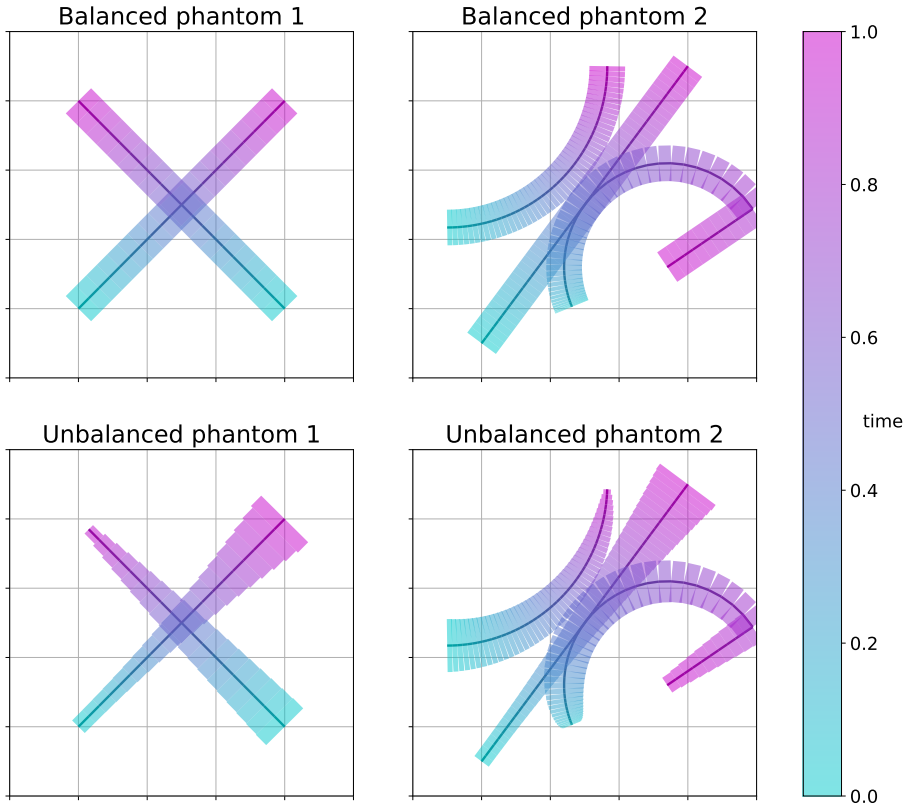


FIGURE 1. Synthetic phantoms of the form $\sigma = \sum_i \delta_{(h^i,\xi^i)}$ for balanced (top row) and unbalanced (bottom row) examples. Colour indicates the time t, the solid line indicates the positions $\xi^i(t)$, and the width of the overlayed band is proportional to $h^i(t)$.

and $A_j : \mathcal{M}([0,1]^2) \to \mathbb{R}^m$ represents a finite number of smoothed Fourier samples. The precise details are given in [5], Section 6. In view of the convexity of Ω, (7.5) reformulates to

$$\mathrm{E}(\sigma) = \frac{1}{2}\sum_{j=0}^{T} \left\| A_j(e_{t_j})_\sharp \sigma - b_j \right\|_2^2 + \int_\Gamma \left(\sum_{j=1}^{T} (t_j - t_{j-1}) \left(\alpha + \frac{\beta}{2} \frac{|\xi(t_j) - \xi(t_{j-1})|^2}{(t_j - t_{j-1})^2} \right) \right) \mathrm{d}\sigma(h,\xi). \tag{7.7}$$

The two phantoms are also from [5], Section 6. In the notation of this work, we would say that, for example, phantom 1 is represented by

$$\sigma = \delta_{(h^1,\xi^1)} + \delta_{(h^2,\xi^2)} \quad \text{where} \quad h^1 = h^2 = 1,\ \xi^1(t) = (0.2 + 0.6t,\ 0.2 + 0.6t), \tag{7.8}$$

$$\xi^2(t) = (0.8 - 0.6t,\ 0.2 + 0.6t). \tag{7.9}$$

Similarly phantom 2 is the sum of 3 Dirac masses, both phantoms are shown here in Figure 1. In the setting of Section 5, 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

$$\mathrm{step}_j((h_0,\xi_0),(h_1,\xi_1)) = \alpha(t_j - t_{j-1}) + \frac{\beta}{2}\frac{|\xi_1-\xi_0|^2}{t_j - t_{j-1}} = \frac{\alpha}{T} + \frac{\beta T}{2}|\xi_1-\xi_0|^2. \tag{7.10}$$

For phantom 1 we have $\alpha = \beta = 0.5$, $T = 21$, and $\alpha = \beta = 0.1$, $T = 51$ for phantom 2. The algorithm of [5] 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 [5] in both examples.

We also replicate the noise-scenarios for phantom 2 as tested in [5]. 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 (*cf.* [5], Tab. 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 4.3, the random variant often finds an even lower energy, despite having a much smaller value of N.

7.2. Wasserstein–Fisher–Rao example

In this section, we show numerical results for the unbalanced transport example presented in Section 6, the data fidelity is the same as in the Benamou–Brenier example. Ideally we would use the exact function $d_{\alpha,\beta,\delta}$ from (6.21) to define step_j, but it lacks a closed-form expression, hence for computational reasons we use the approximation

$$\mathrm{step}_j(\gamma_0,\gamma_1) \overset{\text{def.}}{=} \alpha\frac{h_0+h_1}{2}(t_j - t_{j-1}) + d_{0,\beta,\delta}(\gamma_0, t_{j-1}, \gamma_1, t_j). \tag{7.11}$$

where $d_{0,\beta,\delta}$ is given in (6.23). As in Section 7.1 we use $\alpha = \beta = 0.5$ or 0.1 for the first and second phantom respectively, also $\varphi = \delta = 0.1$ throughout. This leads to the explicit form of step_j

$$\mathrm{step}_j(\gamma_0,\gamma_1) = \frac{\alpha}{T}\frac{h_0+h_1}{2} + \frac{\beta T}{25}\left[\frac{h_0+h_1}{2} - \sqrt{h_0 h_1}\cos\left(\min\left(5|\xi_0-\xi_1|,\pi\right)\right)\right]. \tag{7.12}$$

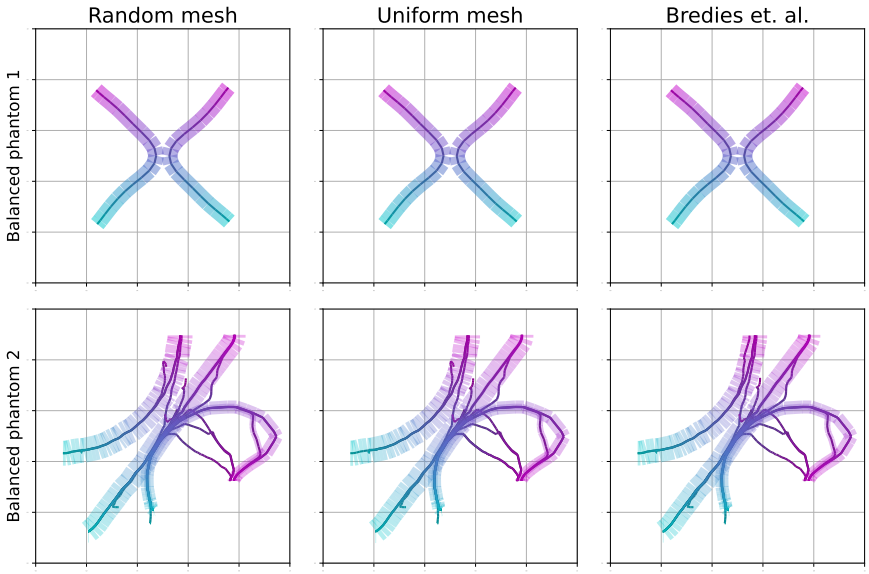
Numerically we re-parametrise the mass to $\widetilde{h}_j \overset{\text{def.}}{=} \sqrt{h_j}$ for each j so that step_j is a C^1 function of $\widetilde{h}_j$. Note that this only effects the sliding step of the Frank–Wolfe algorithm, the remaining steps are unchanged.

Our synthetic phantoms are equivalent to those in Section 7.1 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

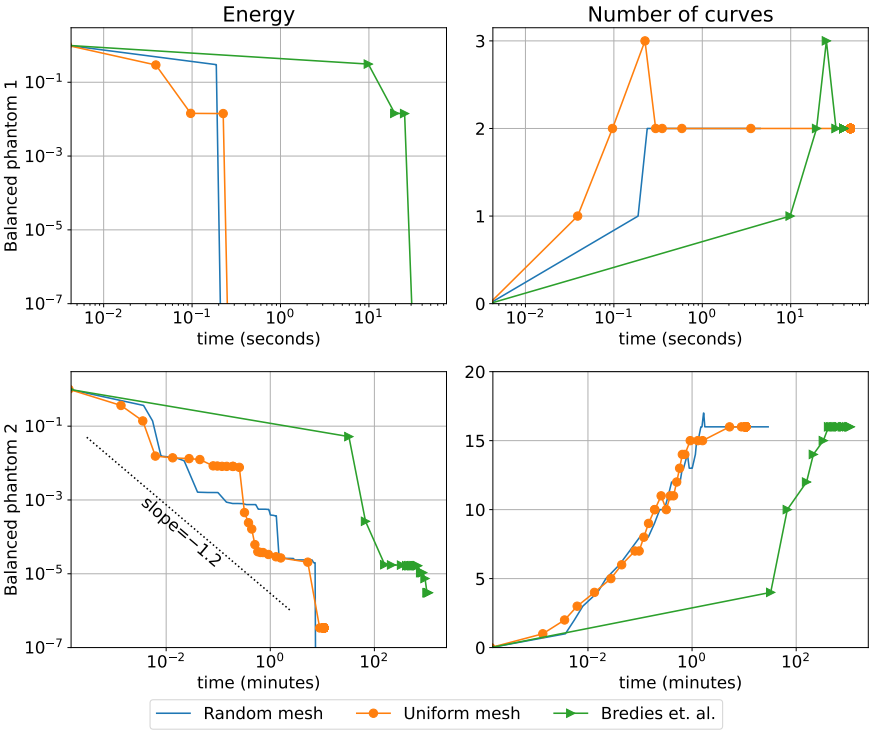
$$h^0(t) = \tfrac{1}{2}(1+3t^2), \qquad \xi^0(t) = (0.2+0.6t,\ 0.2+0.6t), \tag{7.13}$$

$$h^1(t) = \tfrac{3}{2}\sqrt{1-t}, \qquad \xi^1(t) = (0.8-0.6t,\ 0.2+0.6t). \tag{7.14}$$

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



(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 [5], 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).

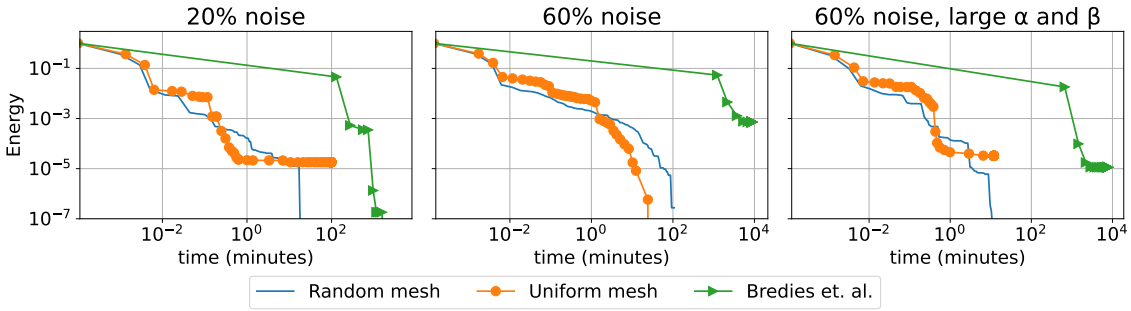


FIGURE 3. Convergence of three default algorithms (see Fig. 2) with different levels of noise and choice of $\alpha = \beta \in \{0.1, 0.3\}$.

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 $\int_0^1 h \, \mathrm{d}t = 1$, $\|h\|_\infty \leq 2$.

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.

7.3. 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.

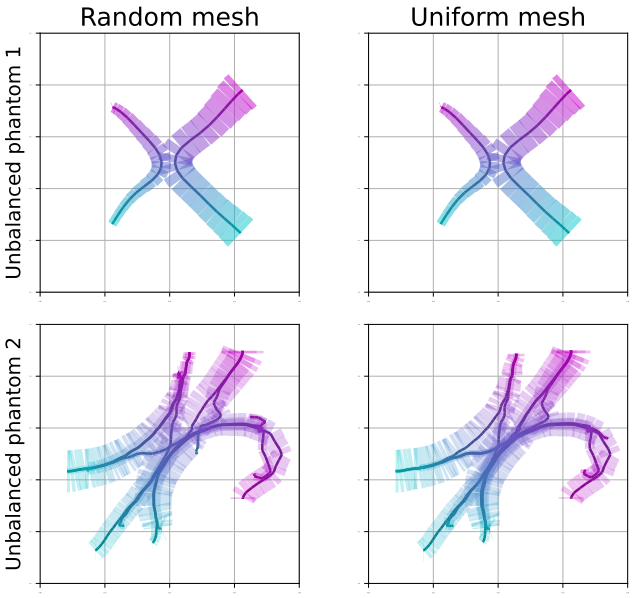
The classical choice of $k = 1$ was always a competitive choice. Large k potentially enables the algorithm to find multiple atoms in one iteration, but slows down the sliding steps. We found that $k \in [1, 5]$ was a reasonable range depending on the sparsity of the signal to be recovered.

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.

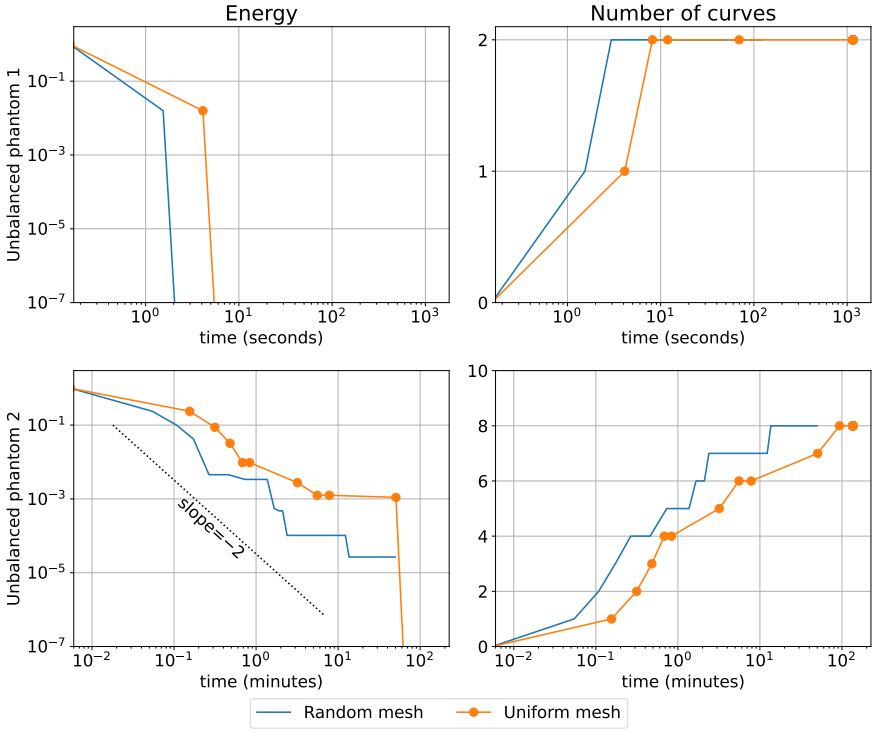
In the unbalanced experiments the value of M had little effect. Although we don't include this figure, we observed that the choice $M = 0$ was also competitive for our unbalanced phantom 2, achieving energies within $10^{-3}\%$ of the best observed energy. It's unclear whether this will generalise to more complicated examples, we chose $M = 10$ as a default. Computational complexity of the linear oracle also scales with M^2, so M 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 all but one 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



(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).

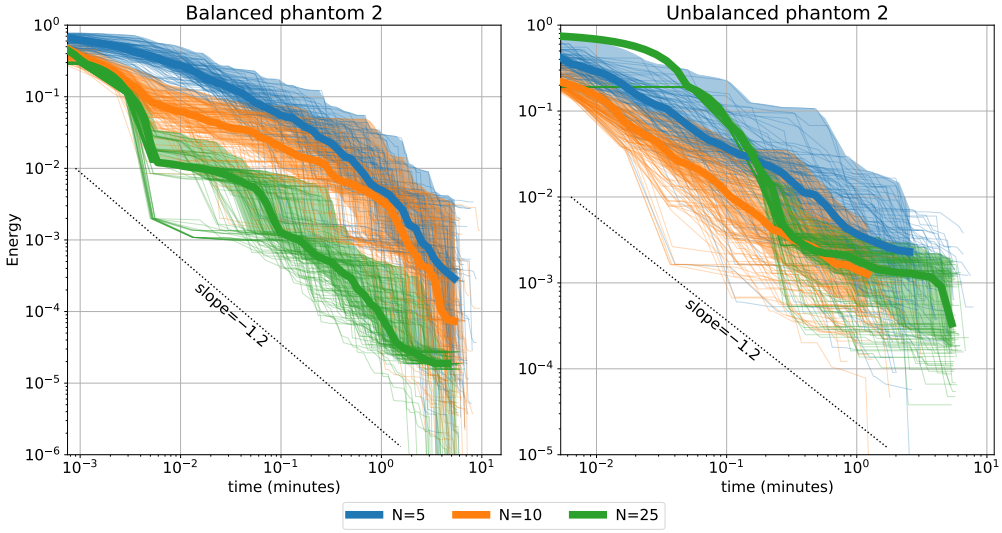


FIGURE 5. Convergence of different random realisations of the random mesh algorithms applied to the balanced and unbalanced phantom 2. The two algorithms are $(1, N, 0)$- and $(1, N, 10)$-random mesh run for 1000 and 100 iterations respectively. 100 random instances of each algorithm are run, each drawn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

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 $O(N^4M^2)$ cost is starting to dominate the computation time. It's possible that $N = 5$ is slower than $N = 10$ for the unbalanced example because the sliding step has to work much harder to find local minima.

8. 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 [5] 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. As an example, 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

$$\widetilde{y} \mapsto \eta_j(\widetilde{y}) + \mathrm{step}_{j-1}(y, \widetilde{y}) + \mathrm{step}_j(\widetilde{y}, y). \tag{8.1}$$

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. 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.

Appendix A. Preliminary results

A.1 Properties of the flat metric

Lemma A.1 (Lem. 2.1)**.** *Define* $\mathrm{d}_\Gamma : \Gamma_0 \times \Gamma_0 \to [0, +\infty[$ *by*

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

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

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

$$(h_1,\xi_1) \sim (h_2,\xi_2) \iff h_1 = h_2 \quad \textit{and} \quad \forall t \in \{h_1 \neq 0\}, \quad \xi_1(t) = \xi_2(t). \tag{A.3}$$

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

$$\left[\gamma_n \overset{\mathrm{d}_\Gamma}{\to} (h,\xi)\right] \iff \left[h_n \to h \text{ in } \mathrm{C}([0,1]) \text{ and for all } \varepsilon>0,\ \xi_n \to \xi \text{ in } \mathrm{C}(\{|h|\ge\varepsilon\})\right] \tag{A.4}$$

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

$$\forall t \in [0,1]\,,\ (h,\xi)\in\Gamma_0, \qquad \Psi(t,h,\xi) \overset{\text{def.}}{=} h(t)\psi(t,\xi(t)). \tag{A.5}$$

Proof. Complete metric space. In [4], Proposition 3.6 it was shown that $\left\{\, t\mapsto h(t)\delta_{\xi(t)} \mid (h,\xi)\in\Gamma_0,\ h\ge 0\,\right\}$ is a complete separable metric space with respect to the flat metric, which is simply $\mathrm{d}(h_1\delta_{\xi_1}, h_2\delta_{\xi_2}) = \mathrm{d}_\Gamma((h_1,\xi_1),(h_2,\xi_2))$. Reducing to $\Gamma_0/\!\sim$ reduces to a single representative all couples (h,ξ) which map to the same element $h\delta_\xi$. After removing this redundancy from Γ_0, it is clear that $(\{\,(h,\xi)\in\Gamma_0/\!\sim \mid h\ge 0\,\}, \mathrm{d}_\Gamma)$ is a complete separable metric space isometrically equivalent to that in [4], Proposition 3.6. In the signed case, consider $h_i^\pm = \max(0,\pm h_i)$, our choice of d_Γ is such that for all $(h_1,\xi_1),(h_2,\xi_2)\in\Gamma_0$

$$\mathrm{d}_\Gamma((h_1,\xi_1),(h_2,\xi_2)) = \mathrm{d}_\Gamma((h_1^+,\xi_1),(h_2^+,\xi_2)) + \mathrm{d}_\Gamma((h_1^-,\xi_1),(h_2^-,\xi_2)). \tag{A.6}$$

It follows that $(\Gamma_0/\!\sim, \mathrm{d}_\Gamma)$ is also a complete separable metric space.

Convergence. First, note from the definition that for all $r_i\in\mathbb{R}$, $x_i\in\overline{\Omega}$,

$$|r_1-r_2| \le \mathrm{d}_F((r_1,x_1),(r_2,x_2)) \le |r_1|+|r_2|. \tag{A.7}$$

Now, fix a sequence $\gamma_n = (h_n,\xi_n)\in\Gamma_0$ and point $\gamma=(h,\xi)\in\Gamma_0$. Suppose $\mathrm{d}_\Gamma(\gamma_n,\gamma)\to 0$, then from (A.7) we have $\|h_n-h\|_\infty \to 0$. Also, for any $\varepsilon>0$ choose $N_\varepsilon\in\mathbb{N}$ such that

$$\forall n \ge N_\varepsilon, \qquad \|h_n-h\|_\infty \le \frac{\varepsilon}{2} \quad \text{and} \quad \mathrm{d}_\Gamma(\gamma_n,\gamma) \le \frac{\varepsilon}{2}. \tag{A.8}$$

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

$$\sup_{|h(t)|\geq\varepsilon} |\xi_n(t) - \xi(t)| \leq \sup_{|h(t)|\geq\varepsilon} \frac{2}{\varepsilon} \min(|h_n(t)|, |h(t)|)|\xi_n(t) - \xi(t)| \leq \frac{2}{\varepsilon} \mathrm{d}_\Gamma(\gamma_n, \gamma) \overset{n\to+\infty}{\longrightarrow} 0. \tag{A.9}$$

This concludes the " $\Longrightarrow$ " direction of (A.4). Conversely, suppose $\|h_n - h\|_\infty \to 0$ and for any $\varepsilon > 0$, $\|\xi_n - \xi\|_{\mathrm{L}^\infty(\{|h|\geq\varepsilon\})} \to 0$. Fix $\varepsilon \in]0, 2[$ and $N_\varepsilon \in \mathbb{N}$ such that for all

$$\forall n \geq N_\varepsilon, \qquad \|h_n - h\|_\infty \leq \frac{\varepsilon}{2} \quad \text{and} \quad \|\xi_n - \xi\|_{\mathrm{L}^\infty(\{|h|\geq\varepsilon\})} \leq \varepsilon. \tag{A.10}$$

Then, from (A.7), for all $n \geq N_\varepsilon$ we have

$$\mathrm{d}_\Gamma(\gamma_n, \gamma) \leq \sup_{t\in[0,1]} \begin{cases} |h_n(t)| + |h(t)| & \text{if } |h(t)| < \varepsilon, \\ |h_n(t) - h(t)| + \min(|h_n(t)|, |h(t)|)|\xi_n(t) - \xi(t)| & \text{else} \end{cases} \tag{A.11}$$

$$\leq \sup_{t\in[0,1]} \begin{cases} 5\varepsilon/2 & \text{if } |h(t)| < \varepsilon, \\ \varepsilon/2 + \|h\|_\infty \varepsilon & \text{else.} \end{cases} \tag{A.12}$$

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

Continuity. Fix $\psi \in \mathrm{C}([0,1] \times \overline{\Omega})$ and define Ψ as in (A.5). 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 $\tau_n \to t$, $\gamma_n \overset{\mathrm{d}_\Gamma}{\to} \gamma$. Observe for each n we have

$$|\Psi(\tau_n, \gamma_n) - \Psi(t, \gamma)| = |h_n(\tau_n)\psi(\tau_n, \xi_n(\tau_n)) - h(t)\psi(t, \xi(t))| \tag{A.13}$$

$$= |(h_n(\tau_n) - h(t) + h(t))\psi(\tau_n, \xi_n(\tau_n)) - h(t)\psi(t, \xi(t))| \tag{A.14}$$

$$\leq |h_n(\tau_n) - h(t)|\|\psi\|_\infty + |h(t)||\psi(\tau_n, \xi_n(\tau_n)) - \psi(t, \xi(t))|. \tag{A.15}$$

$$\leq [\|h_n - h\|_\infty + |h(\tau_n) - h(t)|]\|\psi\|_\infty + |h(t)||\psi(\tau_n, \xi_n(\tau_n)) - \psi(t, \xi(t))|. \tag{A.16}$$

Now consider the limit of $n \to +\infty$. As $h \in \mathrm{C}([0,1])$, $h(\tau_n) \to h(t)$. The characterisation of convergence in d_Γ in (A.4) also confirms $\|h_n - h\|_\infty \to 0$ and either $h(t) = 0$, or

$$|\xi_n(\tau_n) - \xi(t)| \leq |\xi_n(\tau_n) - \xi(\tau_n)| + |\xi(\tau_n) - \xi(t)| \to 0. \tag{A.17}$$

In either case, we see that

$$\limsup_{n\to+\infty} |\Psi(\tau_n, \gamma_n) - \Psi(t, \gamma)| \leq \limsup_{n\to+\infty} |h(t)||\psi(\tau_n, \xi_n(\tau_n)) - \psi(t, \xi(t))| = 0, \tag{A.18}$$

therefore Ψ is continuous at (t, γ).

□

A.2 Projection properties

Theorem A.2 (Thm. 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*

$$\forall \psi \in \mathrm{C}([0,1] \times \overline{\Omega}), \qquad \int_{[0,1]\times\overline{\Omega}} \psi(t, x) \,\mathrm{d}\Theta(\sigma)(t, x) = \int_{\Gamma_0} \left(\int_0^1 h(t)\psi(t, \xi(t)) \,\mathrm{d}t \right) \mathrm{d}\sigma(h, \xi). \tag{A.19}$$

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* (A.19) *holds for all* $\psi \in \mathrm{L}^1_{|\Theta(\sigma)|}([0,1]\times\overline{\Omega})$.
3. *If* $\int_{\Gamma_0} \|h\|_\infty \,\mathrm{d}|\sigma| < +\infty$, *then* $\Theta(\sigma) \in \mathrm{C_w}([0,1]\,;\mathcal{M}(\overline{\Omega}))$.
4. *Suppose* $h,\xi \in \mathrm{AC}^2([0,1])$ *for* σ*-a.e.* $(h,\xi)\in\Gamma_0$. *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*

$$h'(t) = g(t,\xi(t))h(t) \text{ for } \sigma\text{-a.e. } (h,\xi) \text{ and a.e. } t\in\,]0,1[, \tag{A.20}$$
$$\xi'(t) = v(t,\xi(t)) \text{ for } \sigma\text{-a.e. } (h,\xi) \text{ and a.e. } t \text{ such that } h(t)\neq 0, \tag{A.21}$$
$$\text{and } \int_{\Gamma_0}\int_0^1 (1+|v(t,\xi(t))|+|g(t,\xi(t))|)\,|h(t)|\,\mathrm{d}t\,\mathrm{d}|\sigma|(h,\xi) < +\infty, \tag{A.22}$$

then $\int_{\Gamma_0}\|h\|_\infty\,\mathrm{d}|\sigma| < +\infty$ *and* $\Theta(\sigma)$ *satisfies the continuity equation* (1.14).
Conversely, given $\rho\in\mathcal{M}([0,1]\times\overline{\Omega})$, *if* $\rho\geq 0$ *satisfies the continuity equation* (1.14) *and*

$$\int_{[0,1]\times\overline{\Omega}} (1+|v(t,x)|^2+|g(t,x)|^2)\,\mathrm{d}\rho(t,x) < +\infty, \tag{A.23}$$

then $\rho=\Theta(\sigma)$ *for some* $\sigma\in\mathcal{M}^+(\Gamma_0)$ *such that* (A.20)–(A.22) *hold and* $\int_{\Gamma_0}\|h\|_\infty\,\mathrm{d}\sigma<+\infty$.

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

$$(h,\xi)\mapsto \int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t = \int_0^1 \Psi(t,h,\xi)\,\mathrm{d}t \tag{A.24}$$

is continuous (hence Borel) in Γ_0 and dominated by $\|h\|_1\|\psi\|_\infty$, therefore the right-hand side of (A.19) is well-defined. This induces a linear form on $\mathrm{C}([0,1]\times\overline{\Omega})$ which is moreover bounded, since

$$\left|\int_{\Gamma_0}\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\sigma(h,\xi)\right| \leq \|\psi\|_\infty \int_{\Gamma_0}\|h\|_1\,\mathrm{d}|\sigma|. \tag{A.25}$$

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 [4] 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 A.3. *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*

$$\sigma^\pm\left(\left\{(h,\xi)\in\Gamma_0 \mid \inf_{t\in[0,1]} h(t) < 0\right\}\right) = 0 \tag{A.26}$$

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

$$\forall (h,\xi)\in\Gamma_0,\qquad T^\pm(h,\xi)\overset{\text{def.}}{=}(\max(0,\pm h),\xi). \tag{A.27}$$

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, (A.26) holds, and by construction of the image measure

$$\int_{\Gamma_0} \phi(h,\xi)\,\mathrm{d}\sigma^\pm = \int_{\Gamma_0} \phi(\max(0,\pm h),\xi)\,\mathrm{d}\sigma \tag{A.28}$$

for all $\phi \in \mathrm{C}_\mathrm{b}(\Gamma_0)$, and by monotone convergence

$$\int_{\Gamma_0} \|h\|_1\,\mathrm{d}\sigma^\pm = \int_{\Gamma_0} \|\max(0,\pm h)\|_1\,\mathrm{d}\sigma \leq \int_{\Gamma_0} \|h\|_1\,\mathrm{d}\sigma. \tag{A.29}$$

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

$$\begin{aligned}
\int_{[0,1]\times\overline{\Omega}} \psi(t,x)\,\mathrm{d}\Theta(\sigma) &= \int_{\Gamma_0}\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\sigma && \text{(A.30)}\\
&= \int_{\Gamma_0}\int_0^1 [\max(0,h(t)) - \max(0,-h(t))]\psi(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}\sigma && \text{(A.31)}\\
&= \int_{\Gamma_0}\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t\,\mathrm{d}(\sigma^+ - \sigma^-) && \text{(A.32)}\\
&= \int_{[0,1]\times\overline{\Omega}} \psi(t,x)\,\mathrm{d}(\Theta(\sigma^+) - \Theta(\sigma^-)) && \text{(A.33)}
\end{aligned}$$

as required. □

With this decomposition, we can apply the results of [4] to each $\sigma^\pm$. Point (2) becomes a direct consequence of [4], 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 (1.11). In particular,

$$\forall \psi \in \mathrm{C}(\overline{\Omega}),\ t \in [0,1]\,, \qquad \int_{\overline{\Omega}} \psi(x)\,\mathrm{d}\rho_t^\pm(x) = \int_{\Gamma_0} h(t)\psi(\xi(t))\,\mathrm{d}\sigma^\pm(h,\xi). \tag{A.34}$$

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 $\psi \in \mathrm{C}(\overline{\Omega})$. By Lemma A.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

$$\left|\int_{\overline{\Omega}} \psi\,\mathrm{d}(\rho_\tau - \rho_t)\right| \leq \int_{\Gamma_0} |h(\tau)\psi(\xi(\tau)) - h(t)\psi(\xi(t))|\,\mathrm{d}|\sigma|(h,\xi) = \int_{\Gamma_0} |\Psi(\tau,\gamma) - \Psi(t,\gamma)|\,\mathrm{d}|\sigma|(\gamma). \tag{A.35}$$

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 [4], Theorem 4.2. For the forward direction, we again consider the modified Hahn–Jordan decomposition. The assumptions (A.20)–(A.22) 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 [4], Theorem 4.2 to each component and sum for the result. The converse is exactly the statement of [4], Theorem 4.2 because we assume $\rho \geq 0$. □

Lemma A.4 (Lem. 2.3)**.** *For each $p \in [1, +\infty]$ define the set*

$$\Gamma_p \overset{\text{def.}}{=} \left\{ \gamma = (h, \xi) \in \Gamma_0 \mid \|h\|_p \leq 1 \right\}, \tag{A.36}$$

then

$$\left\{ \Theta(\sigma) \mid \sigma \in \mathcal{M}(\Gamma_0),\ \int_{\Gamma_0} \|h\|_p \,\mathrm{d}|\sigma| < +\infty \right\} = \{\, \Theta(\widehat{\sigma}) \mid \widehat{\sigma} \in \mathcal{M}(\Gamma_p) \,\} \tag{A.37}$$

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

$$\text{for all } p \in [1, +\infty], \qquad \Theta(\sigma^n) \overset{*}{\rightharpoonup} \Theta(\sigma) \text{ narrowly in } \mathcal{M}([0,1] \times \overline{\Omega}), \tag{A.38}$$

$$\text{if } p = +\infty,\ \forall t \in [0,1], \qquad (e_t)_\sharp \sigma^n \overset{*}{\rightharpoonup} (e_t)_\sharp \sigma \text{ narrowly in } \mathcal{M}(\overline{\Omega}). \tag{A.39}$$

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

$$\int_{\Gamma_0} \|h\|_p \,\mathrm{d}|\widehat{\sigma}| \leq \int_{\Gamma_p} 1 \,\mathrm{d}|\widehat{\sigma}| = \|\widehat{\sigma}\| < +\infty \tag{A.40}$$

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 $\mathrm{C}([0,1])$ 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 $T(h, \xi) \overset{\text{def.}}{=} \left(\frac{h}{\max(1, \|h\|_p)}, \xi \right)$. T is Borel measurable, therefore $\widehat{\sigma}$ is a Borel measure which satisfies

$$\forall \phi \in \mathrm{C_b}(\Gamma_p), \qquad \int_{\Gamma_p} \phi \,\mathrm{d}\widehat{\sigma} = \int_{\Gamma_0} \max(1, \|h\|_p) \phi \left(\frac{h}{\max(1, \|h\|_p)}, \xi \right) \mathrm{d}\sigma(h, \xi). \tag{A.41}$$

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

$$\left| \int_{\Gamma_p} \phi \,\mathrm{d}\widehat{\sigma} \right| \leq \|\phi\|_\infty \int_{\Gamma_0} (1 + \|h\|_p) \,\mathrm{d}|\sigma|(h, \xi), \tag{A.42}$$

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

$$\int_{[0,1] \times \overline{\Omega}} \psi(t, x) \,\mathrm{d}\Theta(\widehat{\sigma})(t, x) = \int_{\Gamma_p} \left(\int_0^1 h(t) \psi(t, \xi(t)) \,\mathrm{d}t \right) \mathrm{d}\widehat{\sigma}(h, \xi) \tag{A.43}$$

$$= \int_{\Gamma_0} \max(1, \|h\|_p) \left(\int_0^1 \frac{h(t)}{\max(1, \|h\|_p)} \psi(t, \xi(t)) \,\mathrm{d}t \right) \mathrm{d}\sigma(h, \xi) \tag{A.44}$$

$$= \int_{[0,1] \times \overline{\Omega}} \psi(t, x) \,\mathrm{d}\Theta(\sigma)(t, x). \tag{A.45}$$

This shows the "$\subset$" direction, therefore we conclude the equality in (A.37). For continuity of Θ, [31], Theorem 1, Section 1.6 states that Θ is continuous from $(\mathrm{C_b}(\Gamma_p))'$ to $(\mathrm{C}([0,1]\times\overline{\Omega}))'$ if and only if for every $\psi\in\mathrm{C}([0,1]\times\overline{\Omega})$ there exists a finite collection $\Phi\subset\mathrm{C_b}(\Gamma_p)$ such that

$$\forall\sigma\in\mathcal{M}(\Gamma_p),\qquad \left|\int_{[0,1]\times\overline{\Omega}}\psi(t,x)\,\mathrm{d}\Theta(\sigma)\right|\leq\max_{\phi\in\Phi}\left|\int_{\Gamma_p}\phi(h,\xi)\,\mathrm{d}\sigma(h,\xi)\right|. \tag{A.46}$$

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 $(h,\xi)\in\Gamma_p$, $p\geq1$. This confirms the continuity of Θ because

$$\int_{[0,1]\times\overline{\Omega}}\psi(t,x)\,\mathrm{d}\Theta(\sigma)=\int_{\Gamma_p}\left(\int_0^1 h(t)\psi(t,\xi(t))\,\mathrm{d}t\right)\mathrm{d}\sigma(h,\xi)=\int_{\Gamma_p}\phi(h,\xi)\,\mathrm{d}\sigma(h,\xi) \tag{A.47}$$

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

$$\int_{\overline{\Omega}}\psi(t,x)\,\mathrm{d}(e_t)_\sharp\sigma=\int_{\Gamma_\infty}(h(t)\psi(t,\xi(t)))\,\mathrm{d}\sigma(h,\xi)=\int_{\Gamma_\infty}\phi(h,\xi)\,\mathrm{d}\sigma(h,\xi), \tag{A.48}$$

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

Appendix B. Results for the structure of E

Here we prove Theorem 3.1 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

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

B.1 Properties of W

The following lower semi-continuity result is well known (see for instance [16], Prop. 7.1, or [17], Prop. 5.1.7).

Lemma B.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 B.2 (Prokhorov's theorem on measure spaces, *e.g.* [32], Thms. 7.1.7, 8.6.7–8)**.**
If Γ is a complete separable metric space and $\mathcal{A}\subset\mathcal{M}(\Gamma)$ is a family of Borel measures, then

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

Because of this lemma, we can give sufficient conditions for the compactness of sub-levelsets of W.

Lemma B.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 B.1 shows that W is lower semi-continuous, therefore U_t is closed. As U_t is bounded, Lemma B.2 equates compactness with tightness. Finally, [4], Proposition A.2 (adapted from [17], 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 1.1 and 6.

Lemma B.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*

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

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

Proof. It has already been shown that the map $h\delta_\xi \mapsto w(h,\xi)$ is lower semi-continuous with compact sub-levelsets in $\left\{ t \mapsto h(t)\delta_{\xi(t)} \mid w(h,\xi) < +\infty \right\}$ with the flat metric [4], Proposition 3.10. In the proof of Lemma A.1, we showed that this metric is isometrically equivalent to d_Γ, so w is also lower semi-continuous and coercive with respect to d_Γ. □

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

$$(h,\xi) \mapsto w(h,\xi) + \begin{cases} 0 & h = 1 \\ +\infty & \text{else.} \end{cases} \tag{B.4}$$

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

B.2 Properties of D

We consider D of the form in (3.4), that is for some lower semi-continuous $\varphi \colon \Gamma \to]0, +\infty]$,

$$D \stackrel{\text{def.}}{=} \left\{ \sigma \in \mathcal{M}^+(\Gamma) \quad \text{s.t.} \quad \int_\Gamma \varphi \,\mathrm{d}\sigma \leq 1 \right\}. \tag{B.5}$$

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

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

$$\mathrm{Ext}\,(D) = \{0\} \cup \left\{ \varphi(\gamma)^{-1}\delta_\gamma \mid \varphi(\gamma) < +\infty \right\}. \tag{B.6}$$

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 B.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 B.3 confirms that D is compact.

It remains to prove (B.6), we begin with the "$\supset$" inclusion. By non-negativity, note that for all $\sigma, \sigma_0, \sigma_1 \in D$ and $\lambda \in]0,1[$,

$$\sigma = \lambda\sigma_0 + (1-\lambda)\sigma_1 \qquad \Longrightarrow \qquad \sigma_0, \sigma_1 \ll \sigma, \ \textit{i.e.}\ \operatorname{supp}(\sigma_i) \subset \operatorname{supp}(\sigma). \tag{B.7}$$

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

$$\sigma_0 = \frac{\alpha}{\varphi(\gamma)}\delta_\gamma, \ \sigma_1 = \frac{\beta}{\varphi(\gamma)}\delta_\gamma, \ \text{for some } \alpha, \beta \in [0,1]. \tag{B.8}$$

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 B.6. *If $\sigma \in D$ and there exists $\gamma_0, \gamma_1 \in \operatorname{supp}(\sigma)$ distinct, then $\sigma \in]\sigma_0, \sigma_1[$ for some $\sigma_0, \sigma_1 \in D$ such that $\gamma_0 \in \operatorname{supp}(\sigma_0) \setminus \operatorname{supp}(\sigma_1)$ and $\gamma_1 \in \operatorname{supp}(\sigma_1) \setminus \operatorname{supp}(\sigma_0)$.*

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

$$\sigma = \lambda\frac{\sigma_0}{\lambda} + (1-\lambda)\frac{\sigma_1}{1-\lambda}, \tag{B.9}$$

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

$$\operatorname{Ext}(D) \subset \{0\} \cup \left\{ \lambda\delta_\gamma \mid \gamma \in \Gamma, \ \varphi(\gamma) < +\infty, \ 0 < \lambda \leq \varphi(\gamma)^{-1} \right\}. \tag{B.10}$$

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.

□

B.3 Lower semi-continuity of E

Recall that $\mathrm{E}\colon \mathcal{M}^+(\Gamma) \to \mathbb{R}$ is defined by $\mathrm{E}(\sigma) = \mathrm{F}(\sigma) + \mathrm{W}(\sigma)$. In Theorem 3.1 we assume $\varphi, w\colon \Gamma \to [0, +\infty]$ are lower semi-continuous, therefore Lemmas B.1 and B.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.

B.4 Compactness of sub-levelsets of E

We now 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 (Lem. B.5), so the following theorem is not required.

Theorem B.7. *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 B.3 shows that $\mathrm{W}\,|_D$ has compact sub-levelsets. Therefore for any $t \in \mathbb{R}$

$$\{\,\sigma \in D \mid \mathrm{E}(\sigma) \leq t\,\} = \{\,\sigma \in D \mid \mathrm{F}(\sigma) + \mathrm{W}(\sigma) \leq t\,\} \tag{B.11}$$

$$\subset \left\{\,\sigma \in D \mid \mathrm{W}(\sigma) \leq t - \inf_{\widetilde{\sigma} \in D} \mathrm{F}(\widetilde{\sigma})\,\right\}. \tag{B.12}$$

The function E is lower semi-continuous, so the left-hand side is closed and the right-hand side is compact by assumption. We conclude that the left-hand side is compact for each t. □

B.5 E has sparse minimisers

Since the function F is of the form

$$\mathrm{F}(\sigma) = \sum_{j=0}^{T} \mathrm{F}_j(A_j(e_{t_j})_\sharp \sigma) \quad \text{for some} \quad A_j\colon \mathcal{M}(\Gamma) \to \mathbb{R}^m, \text{ convex } \mathrm{F}_j, \tag{B.13}$$

with A_j as in (3.1), F is convex lower semi-continuous. We can therefore use a representer theorem (*e.g.* [13], Cor. 3.8) to demonstrate the sparsity of minimisers.

Lemma B.8. *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 (B.5). Then there exists a minimiser $\sigma^* \in D$ such that*

$$\sigma^* = \sum_{i=1}^{s} a_i \delta_{\gamma_i} \quad \textit{for some} \quad a_i \geq 0,\ \gamma_i \in \Gamma \textit{ with } \varphi(\gamma_i) < +\infty, \tag{B.14}$$

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

$$\min_{\sigma \in V} H(\widetilde{A}\sigma) + R(\sigma) \tag{B.15}$$

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

$$R(\sigma) \overset{\text{def.}}{=} W(\sigma) + \chi_D(\sigma) \quad \text{where} \quad \chi_D(\sigma) \overset{\text{def.}}{=} \begin{cases} 0 & \sigma \in D \\ +\infty & \text{else,} \end{cases} \tag{B.16}$$

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

$$\sigma^* = \sum_{i=1}^{k+1} \theta_i \mu^i, \quad \text{with} \quad \sum_{i=1}^{k+1} \theta_i = 1, \text{ and } \forall i,\ \theta_i \geq 0,\ \mu^i \in \operatorname{Ext}(D). \tag{B.17}$$

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

$$\widetilde{D} \overset{\text{def.}}{=} D \cap \{\, \sigma \in V \mid \langle w, \sigma \rangle = t \,\} \subset \{R \leq t\}. \tag{B.18}$$

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

Case $t = \inf_V R$: In the case $t = \inf_V R = 0$, [13], Corollary 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 CarathÃ©odory's theorem, we can express σ^* in the form in (B.17) but with $\mu^i \in \operatorname{Ext}\left(\widetilde{D}\right)$.

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

$$\widetilde{D} = \left\{ \sigma \in \mathcal{M}^+(\{w = 0\}) \mid \int_\Gamma \varphi \,\mathrm{d}\sigma \leq 1 \right\} \subset D \tag{B.19}$$

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

$$\operatorname{Ext}\left(\widetilde{D}\right) = \{0\} \cup \left\{\, \varphi(\gamma)^{-1} \delta_\gamma \mid \varphi(\gamma) < +\infty,\ w(\gamma) = 0 \,\right\} \subset \operatorname{Ext}(D), \tag{B.20}$$

so (B.17) is satisfied.

Case $t > \inf_V R$: In this case the minimiser σ^* from [13], Corollary 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 [19]) that σ^* belongs to a face of D with dimension at most k. The formulation of (B.17) is again given by CarathÃ©odory's theorem.

We may now deduce (B.14) from (B.17). From Lemma B.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 (B.14) 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 $s \leq k$.

□

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

References

[1] K. Bredies, M. Carioni, S. Fanzon and F. Romero, On the extremal points of the ball of the Benamou-â€“Brenier energy. *Bull. Lond. Math. Soc.* **53** (2021) 1436–1452.

[2] K. Bredies and S. Fanzon, An optimal transport approach for solving dynamic inverse problems in spaces of measures. *ESAIM: M2AN* **54** (2020) 2351–2382.

[3] G.S. Alberti, H. Ammari, F. Romero and T. Wintz, Dynamic spike superresolution and applications to ultrafast ultrasound imaging. *SIAM J. Imaging Sci.* **12** (2019) 1501–1527.

[4] 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, 2020.

[5] 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, 2020.

[6] J.-M. Azaïs, Y. de Castro and F. Gamboa, Spike detection from inaccurate samplings. *Appl. Computat. Harmonic Anal.* **38** (2015) 177–195.
[7] K. Bredies and H.K. Pikkarainen, Inverse problems in spaces of measures. *ESAIM: Control Optim. Calc. Var.* **19** (2013) 190–218.
[8] V. Duval and G. Peyré, Exact support recovery for sparse spikes deconvolution. *Found. Computat. Math.* **15** (2015) 1315–1355, 2015.
[9] C. Poon, N. Keriven and G. Peyré, The geometry of off-the-grid compressed sensing. *Found. Computat. Math.* (2021).
[10] N. Boyd, G. Schiebinger and B. Recht, The alternating descent conditional gradient method for sparse inverse problems. *SIAM J. Optim.* **27** (2017) 616–639.
[11] Q. Denoyelle, V. Duval, G. Peyre and E. Soubies, The Sliding Frank-Wolfe Algorithm and its application to super-resolution microscopy. *Inverse Probl.* (2019).
[12] M. Jaggi, Revisiting Frank–Wolfe: projection-free sparse convex Optimization, in *International Conference on Machine Learning*. PMLR (2013) 427–435.
[13] C. Boyer, A. Chambolle, Y.D. Castro, V. Duval, F. De Gournay and P. Weiss, On representer theorems and convex regularization. *SIAM J. Optim.* **29** (2019) 1260–1281.
[14] K. Bredies and M. Carioni, Sparsity of solutions for variational inverse problems with finite-dimensional data. *Calc. Var. Part. Diff. Eq.* **59** (2019) 14.
[15] M. Unser, J. Fageot and J.P. Ward, Splines are Universal Solutions of linear inverse problems with generalized-tv regularization. *SIAM Rev.* **59** (2017) 769–793.
[16] 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).
[17] L. Ambrosio, N. Gigli and G. Savaré, Gradient Flows: In Metric Spaces and in the Space of Probability Measures, 2nd edn. Birkhäuser Basel (2008).
[18] J.-D. Benamou and Y. Brenier, A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numer. Math.* **84** (2000) 375–393.
[19] L.E. Dubins, On extreme points of convex sets. *J. Math. Anal. Applic.* **5** (1962) 237–244.
[20] L.D. Brown and R. Purves, Measurable selections of extrema. *Ann. Statist.* (1973) 902–912.
[21] C.D. Aliprantis and K.C. Border, Infinite Dimensional Analysis: A Hitchhiker's Guide, 3rd [rev. and enl.] edn. Springer, Berlin; New York (2006). OCLC: ocm69983226.
[22] V.F. Demyanov and A.M. Rubinov, Approximate Methods in Optimization Problems, Vol. 32. Elsevier Publishing Company (1970).
[23] M. Frank and P. Wolfe, An algorithm for quadratic programming. *Naval Res. Logist. Quart.* **3** (1956) 95–110.
[24] 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.
[25] T.H. Cormen, C.E. Leiserson, R.L. Rivest and C. Stein, Introduction to Algorithms, 3rd edn. MIT Press (2009).
[26] R.T. Rockafellar, Conjugate Duality and Optimization. SIAM (1974).
[27] L. Chizat, G. Peyré, B. Schmitzer and F.-X. Vialard, Unbalanced optimal transport: Dynamic and Kantorovich formulations. *J. Funct. Anal.* **274** (2018) 3090–3123.
[28] L. Chizat, G. Peyré, B. Schmitzer and F.-X. Vialard, An interpolating distance between optimal transport and Fisher–Rao metrics. *Found. Computat. Math.* **18** (2018) 1–44.
[29] L. Ding, J. Fan and M. Udell, kFW: a Frank–Wolfe style algorithm with stronger subproblem oracles. arXiv preprint arXiv:2006.16142, 2020.
[30] A. Flinth, F. de Gournay and P. Weiss, On the linear convergence rates of exchange and continuous methods for total variation minimization. *Math. Program.* **190** (2021) 221–257.
[31] K. Yosida, Functional Analysis, 6th edn. Springer Berlin Heidelberg (1980).
[32] V.I. Bogachev, Measure Theory. Springer-Verlag Berlin Heidelberg (2007).