Journal of Nonlinear Science (2022) 32:27
https://doi.org/10.1007/s00332-022-09780-2

Journal of Nonlinear Science

Semi-discrete Optimization Through Semi-discrete Optimal Transport: A Framework for Neural Architecture Search

Nicolás García Trillos[1] · **Javier Morales**[2]

Received: 5 December 2020 / Accepted: 7 January 2022 / Published online: 14 March 2022

Abstract

In this paper, we introduce a theoretical framework for semi-discrete optimization using ideas from optimal transport. Our primary motivation is in the field of deep learning, and specifically in the task of neural architecture search. With this aim in mind, we discuss the geometric and theoretical motivation for new techniques for neural architecture search [in the companion work (García-Trillos et al. in Traditional and accelerated gradient descent for neural architecture search, 2021); we show that algorithms inspired by our framework are competitive with contemporaneous methods]. We introduce a Riemannian like metric on the space of probability measures over a semi-discrete space $\mathbb{R}^d \times \mathcal{G}$ where $\mathcal{G}$ is a finite weighted graph. With such Riemannian structure in hand, we derive formal expressions for the gradient flow of a relative entropy functional, as well as second-order dynamics for the optimization of said energy. Then, with the aim of providing a rigorous motivation for the gradient flow equations derived formally we also consider an iterative procedure known as minimizing movement scheme (i.e., Implicit Euler scheme, or JKO scheme) and apply it to the relative entropy with respect to a suitable cost function. For some specific choices of metric and cost, we rigorously show that the minimizing movement scheme of the relative entropy functional converges to the gradient flow process provided by the formal Riemannian structure. This flow coincides with a system of reaction–diffusion equations on $\mathbb{R}^d$.

Communicated by Mary Pugh.

✉ Javier Morales
javierm1@cscamm.umd.edu

Nicolás García Trillos
garciatrillo@wisc.edu

[1] Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA

[2] Center for Scientific Computation and Mathematical Modeling (CSCAMM), University of Maryland, College Park, MD 20742, USA

Keywords Neural architecture search · Semi-discrete optimization · Optimal transport · Gradient flows

Mathematics Subject Classification 35K57 · 35A24 · 58J60

Contents

1 Introduction . . . 2
1.1 Motivation from Euclidean Space: Otto Calculus in $\mathcal{P}(\mathbb{R}^d)$ . . . 3
1.2 Outline . . . 7
2 Semi-discrete Optimal Transport and Gradient Flows . . . 9
2.1 Some Differential Operators on Graphs . . . 9
2.2 A Riemannian Structure for Semi-discrete OT . . . 11
2.2.1 A Dynamic Optimal Transport Problem in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ . . . 11
2.2.2 A Formal Riemannian Structure for $(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G}), W_2)$ . . . 14
2.3 Computation of Gradient Flows Using the Riemannian Formalism . . . 16
2.4 Hamiltonian Dynamics: Formal Computation of Geodesic Equations and Accelerated Methods for Optimization . . . 19
2.4.1 Geodesics . . . 19
2.4.2 Second-Order Dynamics . . . 19
2.5 Main Theoretical Result . . . 20
3 Metric and Geometric Properties of W_2 . . . 24
3.1 Proof of Theorem 2.7 . . . 24
3.2 Tangent Plane Characterization . . . 28
3.3 A Formal Computation of the Acceleration of a Curve in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$: Geodesic Equations and Accelerated Methods for Optimization . . . 33
4 Properties of Minimizing Pairs of the Static Semi-discrete Optimal Transport Problem . . . 36
5 Properties of JKO Minimizers and Maximum Principle . . . 48
6 Convergence of the JKO Scheme: Proof of Theorem 2.14 . . . 55
7 Summary and Discussion on Applications . . . 63
7.1 From Semi-discrete Optimal Transport to Neural Architecture Search . . . 64
References . . . 65

1 Introduction

Let $(\mathcal{G}, K)$ be a weighted graph over the finite set $\mathcal{G}$ and consider the semi-discrete space $\mathbb{R}^d \times \mathcal{G}$; the function $K : \mathcal{G} \times \mathcal{G} \to [0, \infty)$ is assumed to be symmetric. In this paper, we study, from geometric and variational perspectives, the system of reaction diffusion PDEs:

$$\begin{aligned}\partial_t f_t(x,g) &= \Delta_x f_t(x,g) + \operatorname{div}_x(f_t(x,g)\nabla_x V(x,g)) \\ &\quad + \sum_{g' \in \mathcal{G}} \big[\log f_t(x,g) + V(x,g) - (\log f_t(x,g') + V(x,g'))\big] K(g,g')\theta_{x,g,g'} \\ &\quad (f_t(x,g), f_t(x,g')),\end{aligned} \tag{1.1}$$

for $g \in \mathcal{G}$. In the above, $V : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ is a potential function defined on the semi-discrete space $\mathbb{R}^d \times \mathcal{G}$. The function f_t, i.e., the solution to the system of PDEs, is a function from $\mathbb{R}^d \times \mathcal{G}$ into $\mathbb{R}$ (alternatively, f_t can be thought of as a collection of real valued functions on $\mathbb{R}^d$ indexed by $\mathcal{G}$), and can be interpreted as the density of a probability distribution on $\mathbb{R}^d \times \mathcal{G}$. Finally, the *mobility* function $\theta_{x,g,g'} : [0,\infty)\times[0,\infty) \to [0,\infty)$ serves as "interpolator" for the masses at the points (x, g) and (x, g') and in general dictates the rate at which mass can be exchanged between nodes in $\mathcal{G}$.

In the first part of the paper, we provide a geometric interpretation of system (1.1) by casting it as a formal gradient flow of a relative entropy functional defined on the space $\mathcal{P}(\mathbb{R}^d \times \mathcal{G})$ of probability measures on $\mathbb{R}^d \times \mathcal{G}$ with respect to an appropriate semi-discrete optimal transport metric; this optimal transport metric is reminiscent to the Wasserstein metric in Euclidean space in its dynamic form. While the geometric interpretation that we study here is largely formal, the framework that we introduce is quite rich and allows us to give formal definitions of geodesic equations and second-order dynamics in the space $\mathcal{P}(\mathbb{R}^d \times \mathcal{G})$.

The second perspective that we take has a variational flavor. We introduce a static optimal transport problem that serves as cost function in a minimizing movement scheme (a.k.a. JKO scheme) for the relative entropy functional $\mathcal{E}$. Then, we rigorously show that for a mobility that is independent of the masses to be interpolated (i.e., $\theta_{x,g,g'}$ does not depend on $f_t(x, g)$ and $f_t(x, g')$), system (1.1) can be recovered as the limit of the minimizing movement scheme as the time discretization converges to zero; see Theorem 2.14 for a precise statement.

Regardless of the perspective taken, the main conceptual insight stemming from our work is that the system of equations (1.1) can be interpreted as a gradient flow of relative entropy in the space of probability measures $\mathcal{P}(\mathbb{R}^d \times \mathcal{G})$. What interests us from this interpretation is that it allows us to motivate new schemes for the optimization of an objective function of the form $V : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$, with applications in machine learning such as *neural architecture search* in mind (see the discussion in Sect. 7). The discussion in the next section in the familiar Euclidean setting will help us motivate the prospects of using semi-discrete optimal transport for semi-discrete optimization; we also motivate the theoretical results that we seek in this paper by providing a brief historical background on gradient flows in the space of probability measures. Our companion paper (Trillos et al. 2021) discusses more concretely how part of the theoretical framework presented in this work can be used to define scalable neural architecture search algorithms.

1.1 Motivation from Euclidean Space: Otto Calculus in $\mathcal{P}(\mathbb{R}^d)$

Consider an optimization problem on $\mathbb{R}^d$ of the form

$$\min_{x\in\mathbb{R}^d} V(x),$$

where for the sake of exposition V is assumed to be a nice enough differentiable function. Let us consider the following dynamics on the state space $\mathbb{R}^d$:

$$\begin{cases} \mathrm{d}x(t) = -\nabla_x V(x(t))\mathrm{d}t, \quad t > 0 \\ x(0) = x_0, \end{cases} \tag{1.2}$$

$$\begin{cases} \mathrm{d}x(t) = -\nabla_x V(x(t))\mathrm{d}t + \frac{\sqrt{\eta}}{2}\mathrm{d}B_t, \quad t > 0 \\ x(0) = x_0, \end{cases} \tag{1.3}$$

The usual calculation in normal coordinates at x yields:

$$\begin{aligned} &\int_{\mathcal{M}} \eta\left(\frac{d_{\mathcal{M}}(x,y)}{\epsilon}\right)(\rho(x)-\rho(y))\,\mathrm{dVol}_{\mathcal{M}}(y) \\ &\quad = \epsilon^m \int_{B(0,1)\subset T_x\mathcal{M}} \eta(|w|)\big(\epsilon\langle\nabla\rho(x), w\rangle + \mathcal{O}(\epsilon^2)\big)(1+\mathcal{O}(\epsilon^2))\,\mathrm{d}w \\ &\quad = \mathcal{O}(\epsilon^{m+2}), \end{aligned}$$

for $\epsilon \ll 1$ when the integration variable y is close to $x \in \mathcal{M}$.

$$\begin{cases} \mathrm{d}x^j(t) = -C_t\nabla_x V(x^j(t))\mathrm{d}t + \sqrt{2C_t}dB_t^j, \quad t > 0 \quad j = 1,\dots,J \\ C_t := \frac{1}{J}\sum_{j=1}^J (x^j(t)-\overline{x}(t)) \otimes (x^j(t)-\overline{x}(t)). \end{cases} \tag{1.4}$$

All of the above dynamics can be interpreted as gradient-based continuous time algorithms for the optimization of the function V. (1.2) is gradient descent. (1.3) is gradient descent with Brownian noise; in principle useful to help gradient descent scape local minima. (1.4) is a preconditioned gradient descent with noise. In (1.4) multiple interacting particles are used to define the preconditioning matrix C_t (in this case the running covariance matrix associated to the particles). Besides being used for the optimization of the objective V defined on $\mathbb{R}^d$, Eqs. (1.2), (1.3) and (1.4) share a common underlying structure: they can be associated to certain gradient flows in the space of probability measures $\mathcal{P}(\mathbb{R}^d)$ when endowed with an appropriate optimal transport cost. In what follows we revisit this connection for (1.3) (notice that while degenerate, (1.2) can be seen as a special case of (1.3)) and refer the interested reader to Garbuno-Inigo et al. (2019) for details on how to interpret (1.4).

It is well known that the law of the process $x(t)$ in (1.3) denoted μ_t solves a Fokker–Planck equation of the form:

$$\dot{\mu}_t - \mathrm{div}_x(\mu_t\nabla_x V) - \eta\Delta_x(\mu_t) = 0, \quad t > 0, \tag{1.5}$$

with initial datum μ_0, where in the above div_x is the divergence operator in $\mathbb{R}^d$, ∇_x the gradient operator, and Δ_x the *Laplacian* operator $\Delta_x := \mathrm{div}_x \circ \nabla_x$. In general, Eq. (1.5) must be interpreted in weak form.

Mathematicians and physicists have studied Fokker–Planck equations for decades, and more recently, the seminal work of Jordan et al. (1998) has provided a gradient

flow interpretation for these equations. This interpretation uses the setting of gradient flows in the space of probability measures endowed with the Wasserstein distance. To be more precise let us first recall the definition of the Wasserstein distance with quadratic cost for a pair of probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ (i.e., probability measures with finite second moments):

$$W_2(\mu, \nu)^2 := \min_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \mathrm{d}\pi(x, y), \tag{1.6}$$

where $\Gamma(\mu, \nu)$ is the set of couplings between μ and ν. The above definition can be thought of as describing a *static* optimal transport problem, where one seeks for an optimal assignment of sources and targets of mass without specifying how said transport is actually realized dynamically in time. An alternative *dynamic* reformulation due to Benamou and Brenier (2000) states that

$$W_2(\mu, \nu)^2 = \inf_{t \in [0,1] \mapsto (\mu_t, \nabla_x \varphi_t)} \int_0^1 \int_{\mathbb{R}^d} |\nabla_x \varphi_t|^2 \mathrm{d}\mu_t \mathrm{d}t,$$

where the minimum is taken over all solutions $(\mu_t, \nabla_x \varphi_t)$ to the continuity equation

$$\dot{\mu}_t + \mathrm{div}(\mu_t \nabla_x \varphi_t) = 0, \tag{1.7}$$

with $\mu_0 = \mu$ and $\mu_1 = \nu$. The Benamou–Brenier reformulation highlights the otherwise unclear dynamic nature of the optimal transport problem (1.6) and it reveals a deeper geometric structure that we now discuss. First, solutions to the continuity equation $t \in [0, 1] \mapsto (\mu_t, \nabla_x \varphi_t)$ which represent the different ways in which one can dynamically transport mass from μ_0 to μ_1 can be mathematically interpreted as curves in the space of probability measures. Here, μ_t specifies the location of a particle at time t while the potential $\varphi_t : \mathbb{R}^d \to \mathbb{R}$ is interpreted as "tangent vector" characterizing an allowed infinitesimal change to the location μ_t. Second, the objective function in the Benamou–Brenier problem can be interpreted as the "length" of a given curve (in this case a kinetic energy). A formal Riemannian metric tensor $\langle \cdot, \cdot \rangle_\mu$ can be defined according to:

$$\langle \varphi, \varphi' \rangle_\mu := \int_{\mathbb{R}^d} \nabla_x \varphi \cdot \nabla_x \varphi' \mathrm{d}\mu$$

for any two potentials $\varphi, \varphi' : \mathbb{R}^d \to \mathbb{R}$ (i.e., any two tangent vectors at μ). From the above discussion one can now see that the Wasserstein distance corresponds to the *geodesic distance* associated to the above formal metric tensor, and reveals that the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ can be treated (at least formally) as a Riemannian manifold.

Now, seeing $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ as a formal Riemannian manifold allows one to give a heuristic definition for the gradient flow of a functional $\mathcal{E}$ defined on $\mathcal{P}_2(\mathbb{R}^d)$:

$$\begin{cases} \dot{\mu}(t) = -\nabla_{W_2}\mathcal{E}(\mu(t)) \\ \mu(0) = \mu_0. \end{cases} \tag{1.8}$$

With the Fokker–Planck equation in (1.5) in mind, let us consider the functional

$$\mathcal{E}(\mu) = \int_{\mathbb{R}^d} V \mathrm{d}\mu + \eta H(\mu),$$

where H is the negative Shannon entropy

$$H(\mu) = \begin{cases} \int_{\mathbb{R}^d} f \log f \mathrm{d}x & \text{if } \mathrm{d}\mu = f(x)\mathrm{d}x, \\ +\infty & \text{otherwise .} \end{cases}$$

In the Riemannian formalism, $\nabla_{W_2}\mathcal{E}(\mu)$ must be interpreted as a tangent vector to μ (i.e., a potential) which serves as Riesz representer to the map of directional derivatives of the energy $\mathcal{E}$. Namely, for an arbitrary curve $t \mapsto \mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ which at time $t = 0$ passes through μ with tangent vector φ one must have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\mu_t)|_{t=0} = \langle \nabla_{W_2}\mathcal{E}(\mu), \varphi \rangle_\mu.$$

The set of heuristic computations used to determine the gradient $\nabla_{W_2} E(\mu)$ from the above formula is nowadays widely known as Otto Calculus (see chapter 15 in Villani 2009), and in the case of the relative entropy it gives the formula:

$$-\nabla_{W_2}\mathcal{E}(\mu) = -V - \eta \log f,$$

for every $\mathrm{d}\mu = f(x)\mathrm{d}x$; a similar computation will be presented in more detail in Sect. 2.3 for the semi-discrete setting explored here. Plugging the above potential back in the continuity equation, we recover the Fokker–Planck equation (1.5). In other words, through heuristic arguments from Riemannian geometry that rely on the geometric structure of the optimal transport distance W_2, the dynamics (1.3) used for optimization of V can be lifted to the space $\mathcal{P}_2(\mathbb{R}^d)$ where one can give a gradient flow interpretation.

There is a second way of motivating an interpretation of (1.8) which coincides with the one coming from the Riemannian formalism. To discuss this alternative let us first consider a more general setting and let us assume that $\mathcal{M}$ is an arbitrary topological space, $E : \mathcal{M} \to (-\infty, \infty]$ is an objective function to optimize, $C : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ is a driving cost function, and $\tau > 0$ is a time step. One can then consider the *minimizing movement scheme* (also known as JKO scheme)

$$\mu_{k+1} \in \arg\min_{\mu \in \mathcal{M}} E(\mu) + \frac{1}{2\tau} C(\mu_k, \mu)^2, \tag{1.9}$$

as a discrete time scheme for optimization.

Under suitable conditions, in the limit $\tau \to 0$ iterates (1.9) define a function in time describing what one can refer to as a "gradient flow of E" with respect to the cost function C. Notice that when $\mathcal{M} = \mathbb{R}^d$ and C is the Euclidean metric, the above scheme is essentially the variational formulation of implicit Euler iterates (i.e., the computation of a proximal operator for the function E).

When $\mathcal{M} = \mathcal{P}_2(\mathbb{R}^d)$, C is the Wasserstein distance W_2, and $E = \mathcal{E}$ is the relative entropy, the iterates $\mu_0, \mu_1, \ldots, \mu_k, \ldots$ (where μ_0 is assumed to satisfy $\mathcal{E}(\mu_0) < \infty$) defined recursively by the JKO scheme, i.e.,

$$\mu_{k+1} \in \arg\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{E}(\mu) + \frac{1}{2\tau} W_2^2(\mu_k, \mu), \tag{1.10}$$

can be shown to converge as $\tau \to 0$, to a solution of the Fokker–Planck equation (1.5) (see Jordan et al. 1998). Historically, the JKO scheme (1.10) was the first approach used to give a "gradient flow" interpretation to the Fokker–Planck equation (1.5). In more generality, evolution equations of the form

$$\dot{\mu}_t = \mathrm{div}_x\Big(\nabla_x \mu_t + \mu_t \nabla_x V + \mu_t(\nabla_x U * \mu_t)\Big),$$

are limits of the JKO scheme (1.9) for appropriate functionals defined on $\mathcal{P}_2(\mathbb{R}^d)$ using the Wasserstein distance as cost function. The gradient flow interpretation via the minimizing movement scheme allows one to prove entropy estimates and functional inequalities (see Villani 2009 for more details on this area, which is still very active and in constant evolution).

The minimization problem can be stated in a Lagrangian form as the problem of finding

$$\mu_{k+1} \in \arg\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{E}(\mu) + \mathcal{A}^\tau(\mu_k, \mu), \tag{1.11}$$

where $\mathcal{A}^\tau(\mu_k, \mu)$ denotes the action of the curve in the tangent bundle of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ with minimal kinetic energy connecting μ_k and μ in τ units of time.

In summary, the gradient-based dynamics (1.3) used for optimization of an objective V defined on the state space $\mathbb{R}^d$ are closely linked to a gradient flow on the space of probability measures $\mathcal{P}(\mathbb{R}^d)$. This gradient flow can be motivated using either the formal Riemannian structure that the dynamic formulation of optimal transport has, or the minimizing movement scheme with driving cost taken to be the Wasserstein distance (given that the two interpretations coincide).

1.2 Outline

We organize the rest of the paper as follows. In Sect. 2, we introduce the main objects studied in the paper and state our main results precisely. We start in Sect. 2.1 introducing the basic analytical objects on graphs used throughout the paper. In Sect. 2.2, we

introduce a family of distances on the space of probability measures over $\mathbb{R}^d \times \mathcal{G}$ based on a dynamic formulation of optimal transport. We highlight the formal Riemannian structure of the metric introduced and explore the connections between our definition and the literature on discrete optimal transport. In Sect. 2.3, we use the Riemannian formalism from Sect. 2.2 in order to motivate a definition for the gradient flow of a relative entropy energy closely related to the objective function in the semi-discrete optimization problem of interest. In Sect. 2.4, we use the Riemannian formalism once again and motivate a method for optimization of the relative entropy. In Sect. 2.5, we provide concrete theoretical support for the formal definitions and computations presented in the earlier sections. In particular, we state our main theoretical result, which establishes a connection between the formal definitions from Sect. 2.3 and the minimizing movement scheme discussed in the introduction. To realize the JKO scheme, we introduce a new cost that can be interpreted as a *static* semi-discrete optimal transport cost.

Section 3 explores metric and geometric properties of the transport distances introduced in Sect. 2.2 (i.e., the dynamic semi-discrete transport problems). More specifically, in Sect. 3.1 we prove that these "distances" are indeed metrics. Section 3.2 aims at providing concrete and rigorous support for the heuristic discussion in Sect. 2.2. The discussion in this section motivates more concretely (and rigorously) the characterization of tangent planes of the space of probability measures over $\mathbb{R}^d \times \mathcal{G}$. Section 3.3 presents some heuristic computations justifying the definition of the accelerated method for optimization presented in Sect. 2.4.

Section 4 studies the static semi-discrete transport problem introduced in Sect. 2.5. This section is used later on in the paper, but is also of independent interest. We establish a characterization for solutions to the static semi-discrete optimal transportation problem that is analogous to the celebrated result by Brenier characterizing solutions to the quadratic (Euclidean) optimal transport problem.

Section 5 studies properties of the variational problem used to define the JKO scheme relative to the static semi-discrete cost. We provide a full characterization of solutions to this variational problem. We also establish a maximum principle that is characteristic of Fokker–Plank equations.

In Sect. 6, we put together the results proved in Sects. 4 and 5 and prove our main theoretical result Theorem 2.14, i.e., we show the convergence of the JKO scheme proposed in Sect. 2.5.

We wrap up the paper in Sect. 7 where we provide some conclusions, perspective on future research directions, and discussion on some of the applications in machine learning that have motivated this work.

Note Throughout the paper, some computations will be carried out at a formal level. One of our aims is to stress the importance of the intuition emanating from the formal Riemannian structure that the dynamic formulation of optimal transport has. After all, it is this Riemannian formalism that motivates the algorithms that are implemented in our companion paper (Trillos et al. 2021) for the purposes of neural architecture search (including accelerated methods). The formal computations (or heuristic arguments) that we present here are, for the most part, accompanied by rigorous counterparts.

2 Semi-discrete Optimal Transport and Gradient Flows

2.1 Some Differential Operators on Graphs

In this section, we introduce the discrete differential operators that will later be used to introduce a semi-discrete optimal transport problem on $\mathbb{R}^d \times \mathcal{G}$.

Throughout the paper, we assume that $(\mathcal{G}, K)$ is connected, meaning that for every $g, g' \in \mathcal{G}$, there exists a path $g_0, \dots, g_m \in \mathcal{G}$ with $g_0 = g, g_m = g'$ and $K(g_l, g_{l+1}) > 0$ for every $l = 0, \dots, m-1$.

Given a function $\phi : \mathcal{G} \to \mathbb{R}$, we define its *discrete gradient* as the function $\nabla_g \phi : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$

$$\nabla_g \phi(g, g') := \phi(g') - \phi(g).$$

We use the subscript g in ∇_g to distinguish the discrete gradient from the gradient of a function defined on $\mathbb{R}^d$ (where we use the notation ∇_x). This distinction will become important later on when we consider functions $\phi : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ for which we can compute its gradient ∇_x as well as its discrete gradient ∇_g.

Given a function $h : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ (i.e., a discrete vector field), we define its *discrete divergence* as the function $\mathrm{div}_g h : \mathcal{G} \to \mathbb{R}$ defined by

$$\mathrm{div}_g \quad h(g) := \sum_{g'} (h(g, g') - h(g', g)) K(g, g').$$

Discrete gradients and discrete divergences are related to each other via a discrete integration by parts formula. Namely, a straightforward computation shows that for every $h : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ and $\phi : \mathcal{G} \to \mathbb{R}$ it holds

$$\sum_g \mathrm{div}_g(h)(g)\phi(g) = -\sum_{g,g'} h(g, g') \nabla_g \phi(g, g') K(g, g'). \tag{2.1}$$

In particular if h is of the form $h = \nabla_g \psi \cdot S$ (where $\cdot$ is interpreted as a coordinate wise product) for some $S : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$, then

$$\sum_g \mathrm{div}_g(\nabla_g \psi \cdot S)(g)\phi(g) = -\sum_{g,g'} \nabla_g \phi \cdot \nabla_g \psi S(g, g') K(g, g'). \tag{2.2}$$

In the remainder, we use the following result establishing existence and uniqueness of solutions to elliptic graph PDEs.

Proposition 2.1 *Suppose that the graph $(\mathcal{G}, K)$ is connected. Let $\phi : \mathcal{G} \to \mathbb{R}$ be such that*

$$\sum_g \phi(g) = 0,$$

and let $S : \mathcal{G} \times \mathcal{G} \to [0, \infty)$ *be a symmetric function which is strictly positive whenever* $K(g, g') > 0$. *Then, there exists a unique solution* $\eta : \mathcal{G} \to \mathbb{R}$ *to the graph PDE*

$$\mathrm{div}_g(\nabla_g \eta \cdot S) = \phi \tag{2.3}$$

satisfying

$$\sum_g \eta(g) = 0.$$

Moreover,

$$\sum_{g,g'} |\nabla_g \eta(g, g')|^2 S(g, g') K(g, g') \leq \frac{1}{\lambda_S} \sum_g |\phi(g)|^2,$$

where λ_S *represents the first nonzero eigenvalue of the graph Laplacian matrix* L_S *with entries:*

$$L_S(g, g') := \mathbb{1}_{g=g'} \sum_{g''} 2S(g, g'')K(g, g'') - 2S(g, g')K(g, g').$$

Proof The graph PDE can be written in matrix form as

$$L_S \eta = -\phi,$$

where ϕ and η are interpreted as vectors whose coordinates are indexed by the elements in $\mathcal{G}$, and where the matrix L_S is the (unnormalized) graph Laplacian for a weighted graph $(\mathcal{G}, \omega)$ with weights $\omega_{g,g'} := 2S(g, g')K(g, g')$—see Chung (1996) for the definition of graph Laplacians. The assumptions on S guarantee that the graph $(\mathcal{G}, \omega)$ is connected and thus its graph Laplacian L_S is a positive semi-definite matrix with zero eigenvalue of multiplicity one. The assumption on ϕ guarantees that it belongs to the orthogonal complement of the null space of L_S, and thus is an element of the range of L_S. We conclude that the graph PDE indeed has a unique solution η with average zero.

Finally, according to (2.2),

$$\begin{aligned}\sum_{g,g'} |\nabla_g \eta(g, g')|^2 S(g, g') K(g, g') &= \sum_g -\mathrm{div}_g(\nabla_g \eta S)\eta(g) = -\sum_g \phi(g)\eta(g) \\ &= \sum_g L_S \eta(g)\eta(g),\end{aligned}$$

and thus from Cauchy–Schwarz inequality it follows that

$$\sum_{g,g'} |\nabla_g \eta(g, g')|^2 S(g, g') K(g, g') \leq \left(\sum_g |\phi(g)|^2\right)^{1/2} \left(\sum_g |\eta(g)|^2\right)^{1/2}.$$

From the fact that the graph $(\mathcal{G}, \omega)$ is connected it follows that

$$\sum_g |\eta(g)|^2 \leq \frac{1}{\lambda_S} \sum_g L_S \eta(g)\eta(g),$$

where λ_S is the first nonzero eigenvalue of L_S. Combining the above two inequalities, we obtain the desired result.

□

2.2 A Riemannian Structure for Semi-discrete OT

Let us denote by $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ the space of Borel probability measures on $\mathbb{R}^d \times \mathcal{G}$ with finite second moments. In this section we introduce a metric W_2 on $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ which can be formally interpreted as the geodesic distance associated to a formal Riemannian structure on $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$. Viewing $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ as a Riemannian manifold, in Sect. 2.3 we will be able to give a concrete heuristic interpretation for the gradient descent equation:

$$\begin{cases} \dot{\mu}(t) = -\nabla_{W_2} \mathcal{E}(\mu(t)) \\ \mu(0) = \mu_0, \end{cases} \tag{2.4}$$

for a conveniently chosen function $\mathcal{E}$ on $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ that depends on the objective function V in (7.1). Here, $t \mapsto \mu_t$ describes a path in the space $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$.

2.2.1 A Dynamic Optimal Transport Problem in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$

Motivated by the (Euclidean) Otto Calculus discussed in Sect. 1.1, in order to define an optimal transport problem in the semi-discrete setting, we first introduce an appropriate notion of continuity equation. As in the Euclidean case, semi-discrete continuity equations are used to describe paths in the space $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$.

The definition of a semi-discrete continuity equation depends on the choice of a *mobility function* θ which in full generality is a function of the form

$$\theta : \mathbb{R}^d \times \mathcal{G} \times \mathcal{G} \times \mathbb{R}_+ \times \mathbb{R}_+ \longrightarrow \mathbb{R}_+.$$

In the remainder, we will often write $\theta_{x,g,g'}(s,t)$ and drop the subscripts when no confusion may arise from doing so. The mobility function is used to quantify how easy it is to move mass from a point (x, g) to a point (x, g') when the amount of mass at each of these points is s and t respectively. Mobilities as described above are motivated by the literature on discrete optimal transport. See Chow et al. (2012), Maas (2011), Mielke (2011) and Mielke (2013)) where discrete optimal transport was first introduced and Erbar and Maas (2012), Erbar et al. (2016) and Esposito et al. (2019) for other references where the topic has been developed further. A rigorous passage to the limit from discrete OT to OT in $\mathbb{R}^d$, at least for certain classes of geometric

graphs, has been explored in Gigli and Maas (2013), García Trillos (2021), Gladbach et al. (2020) and Gladbach et al. (2020).

Throughout the paper, we will make the following assumptions on θ. These assumptions are closely related to those in Erbar and Maas (2012) and Maas (2011) for discrete OT.

Assumption 2.2 The mobility function θ satisfies either:

(A0) θ is nonzero, does not depend on s, t and satisfies the symmetry condition: $\theta_{x,g,g'}$ is equal to $\theta_{x,g',g}$ for all $x \in \mathbb{R}^d$, $g, g' \in \mathcal{G}$. In addition, $\theta_{x,g,g}$ is uniformly bounded away from zero on compact sets of $\mathbb{R}^d \times \mathcal{G} \times \mathcal{G}$.

or all of the following

(A1) Symmetry: $\theta_{x,g,g'}(s,t) = \theta_{x,g,g'}(t,s)$ for all s, t.
(A2) Differentiability: The function $\theta_{x,g,g'}(\cdot,\cdot)$ is differentiable.
(A3) Monotonicity: $\theta_{x,g,g'}(r,t) \leq \theta_{x,g,g'}(s,t)$ for all $r \leq s$ and all t.
(A4) Positive homogeneity: $\theta_{x,g,g'}(\lambda s, \lambda t) = \lambda \theta_{x,g,g'}(s,t)$ for all $\lambda \geq 0$ and all s, t.
(A5) The quantity

$$C_{x,g,g'} := \int_0^1 \frac{1}{\sqrt{\theta_{x,g,g'}(1-t,t)}} \mathrm{d}t,$$

is uniformly bounded above on compact subsets of $\mathbb{R}^d \times \mathcal{G} \times \mathcal{G}$, and the quantity $\theta_{x,g,g'}(1,1)$ is uniformly bounded away from zero on compact subsets of $\mathbb{R}^d \times \mathcal{G} \times \mathcal{G}$.

Definition 2.3 In what follows, we consider $v_t : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}^d$, $h_t : \mathbb{R}^d \times \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ and $\mu_t \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$. We say that $t \in [0,T] \mapsto (\mu_t, v_t, h_t)$ satisfies the semi-discrete continuity equation and write

$$\dot{\mu}_t + \mathrm{div}_x(v_t \mu_t) + \mathrm{div}_g(h_t \mu_t) = 0, \tag{2.5}$$

if for all smooth test functions $\zeta \in C_c^\infty(\mathbb{R}^d \times \mathcal{G})$ (i.e., $\zeta(\cdot, g)$ is $C_c^\infty(\mathbb{R}^d)$ for all $g \in \mathcal{G}$) we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d} \sum_g \zeta(x,g) \mathrm{d}\mu_t &= \int_{\mathbb{R}^d} \sum_g \nabla_x \zeta(x,g) \cdot v_t(x,g) \mathrm{d}\mu_t \\ &\quad + \int_{\mathbb{R}^d} \sum_{g,g'} \nabla_g \zeta(x,g,g') h_t(x,g,g') \mathrm{d}\hat{\mu}_t(x,g,g'). \end{aligned} \tag{2.6}$$

In the above expression, for a given $\mu \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$, we use $\hat{\mu}$ to denote the measure on $\mathbb{R}^d \times \mathcal{G} \times \mathcal{G}$ given by

$$\mathrm{d}\hat{\mu}(x,g,g') = \theta_{x,g,g'} \mathrm{d}x \mathrm{d}g \mathrm{d}g'$$

when θ satisfies $(A0)$ in Assumption 2.2 and

$$\mathrm{d}\hat{\mu}(x, g, g') = \theta(\mu_{g|x}(g), \mu_{g|x}(g'))\mathrm{d}\mu_x(x)\mathrm{d}g\mathrm{d}g'.$$

when θ satisfies $(A1)$–$(A5)$ instead. Here, $\mu_{g|x}$ denotes the conditional distribution of g given x. Also, here and in the remainder $\mathrm{d}g$ represents the measure on $\mathcal{G}$ that gives mass one to every element of $\mathcal{G}$.

Remark 2.4 We notice that when μ has a density with respect to $\mathrm{d}x\mathrm{d}g$, i.e.,

$$\mathrm{d}\mu(x, g) = f(x, g)\mathrm{d}x\mathrm{d}g,$$

then

$$\mathrm{d}\hat{\mu}(x, g, g') = \theta(f(x, g), f(x, g'))\mathrm{d}x\mathrm{d}g\mathrm{d}g'.$$

Indeed, this is immediate if θ satisfies $(A0)$ and otherwise follows from the homogeneity of the mobility θ, i.e., condition $(A4)$.

Remark 2.5 Let $t \in [0, T] \mapsto (\mu_t, v_t, h_t)$ be a solution to the semi-discrete continuity equation and suppose that for every t, μ_t is absolutely continuous with respect to $\mathrm{d}x\mathrm{d}g$ and has density $f_t(x, g)$. Additionally, suppose that the mappings $(t, x, g) \mapsto f(t, x, g)$, $(t, x, g) \mapsto v_t(x, g, g')$ and $(t, x, g) \mapsto h_t(x, g, g')$ are all smooth. In that case we can see that for every test function $\zeta \in C_c^\infty(\mathbb{R}^d \times \mathcal{G})$ we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} \sum_g \zeta(x, g) \frac{\partial}{\partial t} f_t(x, g)\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d} \sum_g \zeta(x, g)\mathrm{d}\mu_t(x, g) \\
&= \int_{\mathbb{R}^d} \sum_g \nabla_x \zeta(x, g) \cdot v_t(x, g)\mathrm{d}\mu_t \\
&\quad + \int_{\mathbb{R}^d} \sum_{g,g'} \nabla_g \zeta(x, g, g') h_t(x, g, g') K(g, g') \theta(f_t(x, g), f_t(x, g'))\mathrm{d}x \\
&= -\int_{\mathbb{R}^d} \sum_g \zeta \,\mathrm{div}_x(v_t f_t)\mathrm{d}x - \int_{\mathbb{R}^d} \sum_g \zeta \,\mathrm{div}_g(h_t \cdot \hat{f}_t)\mathrm{d}x,
\end{aligned}$$

where $\hat{f}_t(x, g, g') := \theta(f_t(x, g), f_t(x, g'))$. The last equality follows using integration by parts in x for the first term and in g for the second term (i.e., identity (2.1)). We conclude that

$$\frac{\partial}{\partial t} f_t + \mathrm{div}_x \quad (v_t f_t) + \mathrm{div}_g \quad (h_t \hat{f})(x, g) = 0, \quad \forall t, x, g.$$

which justifies the notation (2.5) used in Definition 2.3.

With the above notion of continuity equation in hand, we are now able to introduce the following dynamic optimal transport problem.

Definition 2.6 Let μ_0 and μ_1 be two elements in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$. We define

$$W_2(\mu_0,\mu_1)^2 := \inf_{t\in[0,1]\mapsto(\mu_t,\nabla_x\phi_t,\nabla_g\psi_t)} \int_0^1 \Bigg(\int_{\mathbb{R}^d} \sum_{g\in\mathcal{G}} |\nabla_x \phi_t(x,g)|^2 \mathrm{d}\mu_t(x,g) + \int_{\mathbb{R}^d} \sum_{g,g'} (\nabla_g \psi_t(x,g,g'))^2 K(g,g') \mathrm{d}\hat{\mu}_t(x,g,g') \Bigg) \mathrm{d}t, \tag{2.7}$$

where the infimum is taken among all solutions to the semi-discrete continuity equation of the form $t \in [0,1] \mapsto (\mu_t, \nabla_x\phi_t, \nabla_g\psi_t)$, where $\phi_t : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ and $\psi_t : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$.

In words, $W(\mu_0,\mu_1)^2$ is obtained by minimizing the *total kinetic energy* associated to paths connecting μ_0 and μ_1. In Sect. 3, we rigorously show that W_2 as defined above is indeed a metric on the space $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$. The precise statement is the following.

Theorem 2.7 *Let $(\mathcal{G}, K)$ be a connected weighted graph, where K is a symmetric weight matrix with nonnegative entries. Suppose that the mobility function $\theta : \mathbb{R}^d \times \mathcal{G} \times \mathcal{G} \times \mathbb{R} \times \mathbb{R} \to [0,\infty)$ satisfies Assumptions 2.2. Then, W_2 as introduced in Definition 2.6 is a metric on the space $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$.*

Remark 2.8 In the above definition, we have introduced the semi-discrete Wasserstein distance as an optimization problem over a specific class of solutions to the continuity equation, namely, solutions whose driving vector fields are gradients of potentials. It is actually possible to show that removing the restriction to this smaller class of vector fields does not change the definition given. We have introduced W_2 in this way for convenience.

Later on we will show that the class of vector fields can actually be restricted even further (at least for regular enough measures). In particular the potentials ϕ and ψ may be taken to be the same. This observation will be useful when interpreting $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ as a formal Riemannian manifold with geodesic distance that coincides with W_2.

Remark 2.9 The definition given in (2.6) is a particular case of the formal definition given in Mielke (2011). Here, we present some heuristic computations providing a characterization of tangent planes (see the informal Theorem 2.10 and its rigorous counterpart in Sect. 3.2), and a formal computation of the acceleration of curves which in turn motivates: (1) geodesic equations, and (2) accelerated methods for optimization (see Sects. 2.4 and 3.3).

2.2.2 A Formal Riemannian Structure for $(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G}), W_2)$

In differential geometry, when working in the setting of a smooth manifold $\mathcal{M}$, a tangent vector at a given point q is interpreted as the velocity of a curve in $\mathcal{M}$ when passing through q. The collection of tangent vectors at q, i.e., q's tangent plane, is typically denoted by $\mathcal{T}_q\mathcal{M}$. When $\mathcal{M}$ is endowed with a Riemannian structure, one can compute inner products $\langle p, \tilde{p}\rangle_q$ between elements $p, \tilde{p} \in T_q\mathcal{M}$ and introduce a

notion of distance between points $q, \tilde{q} \in \mathcal{M}$ according to

$$d(q,\tilde{q})^2 := \inf_{t\in[0,1]\mapsto q(t)} \int_0^1 \langle \dot{q}(t), \dot{q}(t)\rangle_{q(t)} \mathrm{d}t,$$

where the infimum ranges over all paths connecting q to $\tilde{q}$.

We now provide some heuristics that motivate how the space $(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G}), W_2)$ can actually be interpreted in light of this Riemannian formalism. The first step is an informal statement that will justify some of the subsequent discussion. A precise (and rigorous) version will be presented in Sect. 3.2.

Theorem 2.10 Characterization of potentials (informal) *Let $t \to \mu_t$, be an arbitrary curve in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ with velocity fields generated by the potentials (ϕ_t, ψ_t). Then, we can replace the potentials with a pair of the form (φ_t, φ_t) such that it acts as a velocity field for the same curve $t \to \mu_t$, and has minimal total kinetic energy.*

The above suggests that there is some redundancy when considering different potentials ϕ, ψ and actually one may take both potentials to be the same. Indeed, such a characterization allows us to formally identify the tangent plane at a measure μ in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ as:

$$\begin{aligned} T_\mu \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G}) := \Big\{ \varphi \quad : \quad & \int_{\mathbb{R}^d \times \mathcal{G}} |\nabla_x \varphi|^2 \mathrm{d}\mu(x,g) \\ & + \int_{\mathbb{R}^d \times \mathcal{G} \times \mathcal{G}} [\varphi(x,g') - \varphi(x,g)]^2 K(g,g') \mathrm{d}\hat{\mu}(x,g,g') < \infty \Big\}. \end{aligned} \tag{2.8}$$

endowed with the inner product:

$$\begin{aligned} \langle \varphi, \widetilde{\varphi} \rangle_\mu := & \int_{\mathbb{R}^d} \sum_g \nabla_x \varphi(x,g) \cdot \nabla_x \widetilde{\varphi}(x,g) \mathrm{d}\mu(x,g) \\ & + \int_{\mathbb{R}^d} \sum_{g,g'} \nabla_g \varphi \cdot \nabla_g \tilde{\varphi} K(g,g') \mathrm{d}\hat{\mu}(x,g,g'). \end{aligned} \tag{2.9}$$

A rigorous definition of the tangent plane is out of the scope of this work. Putting aside all technicalities, we can observe formally that the semi-discrete Wasserstein distance W_2 from Definition (2.6) can be rewritten as

$$W_2^2(\mu_0, \mu_1) = \inf \int_0^1 \langle \varphi_t, \varphi_t \rangle_{\mu_t} \mathrm{d}t,$$

where the inf ranges over solutions to the continuity equation $t \in [0,1] \mapsto (\mu_t, \nabla_x \varphi_t, \nabla_g \varphi_t)$ connecting μ_0 with μ_1 (i.e., over paths in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$, according to the informal Theorem 2.10): this formula and its interpretation reveal the Riemannian structure of the metric W_2. In the next subsection, we use this Riemannian formalism to motivate a concrete interpretation for (2.4).

2.3 Computation of Gradient Flows Using the Riemannian Formalism

In this section, we use the Riemannian formalism for $(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G}), W_2)$ discussed in the previous section to motivate a definition for the gradient of a given energy function $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G}) \to \mathbb{R} \cup \{\infty\}$, and ultimately give a concrete meaning to the gradient flow ODE (2.4). Looking forward to our applications, here we will focus on energies of the form

$$\mathcal{E}(\mu) := \begin{cases} \int_{\mathbb{R}^d} \sum_g \vartheta(f(x,g),x,g) \quad \mathrm{d}x \quad \text{if} \quad \mathrm{d}\mu(x,g) = f(x,g)\,\mathrm{d}x\mathrm{d}g \\ +\infty \quad \text{otherwise,} \end{cases} \tag{2.10}$$

where $\vartheta : [0,\infty) \times \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ is given by

$$\vartheta(r,x,g) := r\log r + V(x,g)r.$$

We think of the function $V : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ as the objective of the semi-discrete optimization problem (7.1). Here, we assume for simplicity that V is differentiable in the x coordinate. Notice that $\mathcal{E}$ is a relative entropy and can be written as the sum of the two terms

$$\mathcal{E}(\mu) = H(\mu) + \int_{\mathbb{R}^d \times \mathcal{G}} V(x,g)\mathrm{d}\mu(x,g),$$

where H denotes the (negative) entropy of μ when the base measure on $\mathbb{R}^d \times \mathcal{G}$ is the product measure $\mathrm{d}x\mathrm{d}g$. The entropy term H may be multiplied by a positive factor for generality without that entailing any meaningful changes in the computations below. This choice of energy is motivated by the discussion presented in Sect. 1.1.

Let us recall that in Riemannian geometry, the gradient of a differentiable function $E : \mathcal{M} \to \mathbb{R}$ at a point q is defined as a tangent vector $\nabla_{\mathcal{M}} E(q)$ at q characterized by: for every smooth curve $t \in (-\varepsilon,\varepsilon) \mapsto q(t) \in \mathcal{M}$ with $q(0) = q$,

$$\langle \nabla_{\mathcal{M}} E(q), \dot{q}(0)\rangle_q = \frac{\mathrm{d}}{\mathrm{d}t} E(q(t))\Big|_{t=0}.$$

In words, the above means that the gradient of a given function E at a given point q on the Riemannian manifold $\mathcal{M}$ serves as Riesz representer (with respect to the inner product at that point) for the map of directional derivatives of the function E at the point q.

Using the above discussion as motivation, we notice that for arbitrary $\mu \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ such that $\mathcal{E}(\mu) < \infty$, the gradient of $\mathcal{E}$ (with respect to W_2) at the point μ must be interpreted as a potential φ_μ. Our goal is to identify φ_μ. In order to achieve this, we consider $t \in (-\varepsilon,\varepsilon) \mapsto (\mu_t, \nabla_x\psi_t, \nabla_g\psi_t)$ an arbitrary curve in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ which at time $t = 0$ passes through the point μ (i.e., $\mu_0 = \mu$). We assume $d\mu_t = f_t \mathrm{d}x\mathrm{d}g$

and write $f = f_0$. We want φ_μ to satisfy

$$\langle \varphi_\mu, \psi_0 \rangle_\mu = \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{E}(\mu_t)\Big|_{t=0}. \tag{2.11}$$

A formal computation shows that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{E}(\mu_t)\Big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \int_{\mathbb{R}^d} \sum_g \Big(\log f_t + V \Big) f_t \mathrm{d}x \\ &= \int_{\mathbb{R}^d} \sum_g \Big(\log f_0 + 1 + V \Big) \partial_t f_0(x, g) \mathrm{d}x. \end{aligned}$$

Using the semi-discrete continuity equation, the last line can be rewritten as

$$\begin{aligned} &\int_{\mathbb{R}^d} \sum_g \nabla_x (\log f + V) \cdot \nabla_x \psi_0 \mathrm{d}\mu(x, g) \\ &+ \int_{\mathbb{R}^d} \sum_{g, g'} \nabla_g (\log f + V) \cdot \nabla_g \psi_0 K(g, g') \mathrm{d}\hat{\mu}(x, g, g'), \end{aligned}$$

which in turn can be rewritten as $\langle \log f + V, \psi_0 \rangle_\mu$. It follows that φ_μ can be taken to be

$$\nabla_{W_2} \mathcal{E}(\mu) := \varphi_\mu = \log f + V. \tag{2.12}$$

Having found the gradient of $\mathcal{E}$ through the above heuristic computations, we can now give a concrete interpretation to (2.4) by plugging in the potential $-(\log f + V)$ in the semi-discrete continuity equation. In particular, $t \in [0, \infty) \to \mu_t$ in (2.4) is interpreted as

$$\mathrm{d}\mu_t(x, g) = f_t(x, g) \mathrm{d}x \mathrm{d}g,$$

where f_t follows (1.1). Equation (1.1) can be described as a coupled system of *reaction–diffusion* equations indexed by $g \in \mathcal{G}$. The presence of the last term in (1.1) is responsible for the coupling of the dynamics. From the transport point of view, this coupling term induces mass to be exchanged between different nodes (and thus the total mass at a single $g \in \mathcal{G}$ changes in time). From the optimization point of view, a coupled system implies that information on the optimization over parameters x for a given node g is used for the optimization of parameters x for nearby nodes g' and vice versa.

We finish this section with two examples of mobility functions θ and their corresponding gradient flows.

Example 2.11 Let $W : \mathbb{R}^d \to \mathbb{R}$ be a function in the Sobolev space $W^{1,2}(\mathbb{R}^d)$ satisfying

$$\int |x|^2 e^{-W} \mathrm{d}x < \infty. \tag{2.13}$$

We define a mass independent mobility θ according to

$$\theta_{x,g,g'}(s,t) := e^{-W(x)}.$$

This mobility function satisfies $(A0)$ in Assumptions 2.2. We notice that in the corresponding optimal transport problem from definition (2.6) the transfer of mass between points (x, g) and (x, g') is cheap precisely when $W(x)$ is large. We also notice that the cost of transporting mass along the graph $\mathcal{G}$ does not depend on the actual amount of mass that is initially located at the nodes of $\mathcal{G}$, a situation that contrasts with the one presented in the next example.

Finally, for this choice of mobility the system of Eq. (1.1) becomes the system of nonlinear reaction diffusion equations:

$$\begin{aligned}\partial_t f(x,g) = \Delta_x f_t(x,g) + \mathrm{div}_x(f_t(x,g)\nabla_x V(x,g)) + \sum_{g'\in\mathcal{G}} \big[\log f(x,g) + V(x,g) \\ - (\log f(x,g') + V(x,g')\big]K(g,g')e^{-W(x)}.\end{aligned} \tag{2.14}$$

Example 2.12 Suppose that the mobility θ takes the form

$$\theta_{x,g,g'}(s,t) = \theta_{\log}(s\exp(V(x,g)), t\exp(V(x,g'))$$

where $\theta_{\log}$ is the logarithmic interpolation function:

$$\theta_{\log}(a,b) := \frac{a-b}{\log(a)-\log(b)} = \int_0^1 a^r b^{1-r}\mathrm{d}r.$$

For this choice of mobility, the dynamic cost of transporting mass from (x, g) into (x, g') depends on the value of the potential V at these points, as well as on the value of the mass that is currently located at them. In particular, it is more expensive to move mass between these points when the amount of mass at one of them is close to zero. This mobility function satisfies $(A1)$–$(A5)$ in Assumptions (2.2). In this case, Eq. (1.1) take the form

$$\begin{aligned}\partial_t f_t(x,g) =& \Delta_x f_t(x,g) + \mathrm{div}_x(f_t(x,g)\nabla_x V(x,g)) \\ &+ \sum_{g'\in\mathcal{G}} \big[f_t(x,g)\exp(V(x,g)) - f_t(x,g')\exp(V(x,g'))\big]K(g,g').\end{aligned}$$

which is a **linear** system of reaction diffusion equations.

2.4 Hamiltonian Dynamics: Formal Computation of Geodesic Equations and Accelerated Methods for Optimization

In this section, we discuss how a formal Riemannian structure can be used to introduce accelerated methods for optimization of energies on $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$. We first provide a characterization of the geodesic equations in the space $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$, and then introduce a system of *accelerated* dynamics for the minimization of the energy $\mathcal{E}$ in (2.10). These two sets of equations are related to certain Hamiltonian systems in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ which can be formally defined using a notion of acceleration of curves. Throughout this section, we continue to work at a formal level.

2.4.1 Geodesics

To motivate the characterization of geodesics in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$, let us recall that when working on a smooth Riemannian manifold $\mathcal{M}$, the local equation satisfied by a geodesic $t \mapsto q(t) \in \mathcal{M}$ can be written as

$$\begin{cases} \dot{q}(t) = p(t) \\ \dot{p}(t) = 0, \end{cases}$$

where $t \mapsto p(t)$ is understood as a vector field along the curve $t \mapsto q(t)$, and its derivative as the covariant derivative of p along the curve q (using the Levi-Civita connection) written $\nabla_{\dot{q}} p$. The second equation states that geodesics have zero *acceleration*, i.e., $\nabla_{\dot{q}} \dot{q} = 0$. This system can be understood as a Hamiltonian system on the tangent bundle $\mathcal{T}\mathcal{M}$ with Hamiltonian $\mathcal{H}(q, p) := \frac{1}{2}|p|_q^2$.

Following the above intuition, in Sect. 3.3 we will formally derive for the formal Riemannian manifold $(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G}), W_2)$ the system of equations:

$$\begin{cases} \dot{\mu}_t + \mathrm{div}_x(\nabla_x \varphi_t \mu_t) + \mathrm{div}_g(\nabla_g \varphi_t \hat{\mu}_t) = 0 \\ \partial_t \varphi_t + \frac{1}{2}|\nabla_x \varphi_t|^2 + \sum_{g'} \left(\nabla_g \varphi_t\right)^2 K(g, g') \partial_1 \theta_{x,g,g'}(f_t(x, g), f_t(x, g')) = 0, \end{cases} \tag{2.15}$$

characterizing geodesics in the space $(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G}), W_2)$; in the above $\mathrm{d}\mu(x, g) = f(x, g)\mathrm{d}x\mathrm{d}g$, and we interpret $\partial_1 \theta_{x,g,g'}(s, t)$ as the derivative in s of the mobility function. The first of the two equations, i.e., the continuity equation, simply states that the curve $t \mapsto \mu_t$ moves with velocity $(\nabla_x \varphi_t, \nabla_g \varphi_t)$. On the other hand, the left hand side of the second equation can be understood as the derivative of the velocity along the curve (i.e., the acceleration), and so by setting it to zero one matches the intuition coming from Riemannian geometry that was discussed earlier.

2.4.2 Second-Order Dynamics

In order to introduce a system of second-order dynamics for the optimization of an energy $\mathcal{E}$ like that in (2.10), we once again return to the setting of a smooth Riemannian

manifold $\mathcal{M}$ and consider the optimization of an objective function $q \in \mathcal{M} \mapsto E(q)$. The system

$$\begin{cases} \dot{q}(t) = p(t) \\ \dot{p}(t) = -\gamma\, p(t) - \nabla_{\mathcal{M}} E(q(t)), \end{cases}$$

can be interpreted as a continuous time accelerated method for the optimization of the objective E. Here we abuse the use of the term *accelerated* method slightly given the motivation coming from the Euclidean setting. Indeed, in the case $\mathcal{M} = \mathbb{R}^d$ and when the parameter γ is allowed to depend on time according to $\gamma = \gamma_t = 3/t$, the above dynamics correspond to the continuous time analogue of the celebrated Nesterov accelerated method for optimization (Su et al. 2016). For general $\mathcal{M}$, the above system may be interpreted again as a dynamical system on the tangent bundle $\mathcal{T}\mathcal{M}$, and can be understood as the flow map induced by a vector field that is the addition of a Hamiltonian vector field on $\mathcal{T}\mathcal{M}$ with Hamiltonian $\mathcal{H}(q, p) = \frac{1}{2}|p|_q^2 + E(q)$ and a dissipative term that corresponds to the gradient of an energy $(q, p) \mapsto \frac{\gamma}{2}|p|_q^2$ for a positive parameter $\gamma > 0$.

Following the above intuition, we can introduce an accelerated method for the optimization of an objective on $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ such as the relative entropy $\mathcal{E}$. For this purpose we use the formal computation of the gradient of the relative entropy (2.12) from Sect. 2.3 as well as the expression for the acceleration of curves in the formal Riemannian structure (which actually was already used when introducing the geodesic equation (2.15) and will be formally computed in Sect. 3.3). We obtain the system:

$$\begin{cases} \dot{\mu}_t + \mathrm{div}_x(\nabla_x \varphi_t \mu_t) + \mathrm{div}_g(\nabla_g \varphi_t \hat{\mu}_t) = 0 \\ \partial_t \varphi_t + \frac{1}{2}|\nabla_x \varphi_t|^2 + \sum_{g'} \left(\nabla_g \varphi_t\right)^2 K(g, g') \partial_1 \theta_{x,g,g'}(f_t(x, g), f_t(x, g')) \\ \quad = -[\gamma \varphi_t(x, g) + \log f_t(x, g) + V(x, g)]; \end{cases} \tag{2.16}$$

in the above, we interpret $\mathrm{d}\mu(x, g) = f(x, g)\mathrm{d}x\mathrm{d}g$.

Remark 2.13 Notice that when the interpolation map θ is like the one in Example 2.11 the expression for the acceleration of a curve with velocity induced by the potentials φ_t reads

$$\partial_t \varphi_t + \frac{1}{2}|\nabla_x \varphi_t|^2.$$

2.5 Main Theoretical Result

In the previous sections, we have taken a formal Riemannian approach to make sense of the gradient descent ODE (2.4) when the energy $\mathcal{E}$ is the relative entropy defined in (2.10). In this section, we provide a more solid theoretical ground motivating equations (1.1). For that purpose, we will define the gradient flow of $\mathcal{E}$ using the *minimizing movement scheme* approach that we mentioned at the end of Sect. 1.1. To achieve this, we first introduce a family of *static* transport costs that are used to define the

iterations (1.9) (thinking of $\mathcal{M} = \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$). Our main theoretical result, Theorem 2.14, states that for a suitable static cost (see (2.20) below), and for a suitable choice of mobility θ (the one in Example 2.11), the resulting minimizing movement scheme converges, as the time discretization parameter τ goes to zero, toward a solution of the equation formally derived in (2.14).

It is worth highlighting that the minimizing movement scheme that we consider here has the advantage of being defined in terms of a (static) transport cost that is closer to the Kantorovich formulation of the classical optimal transport problem (i.e., (1.6)), rather that in terms of the dynamic problem (2.6). First, the static formulation is computationally cheaper (e.g., using the entropic regularization methods from Peyré and Cuturi (2019) which can be used in our context). Additionally, for the static formulation we will be able to use techniques similar to those developed in Figalli and Gigli (2010) to show that the resulting minimizing movement scheme satisfies a type of maximum principle characteristic of Fokker–Planck equations.

To define our static transportation costs, we first introduce some notation. Given a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ we will consider the unique collection $\{\mu_g\}_{g\in\mathcal{G}}$ of positive measures over $\mathbb{R}^d$, such that

$$\mu = \sum_{g\in\mathcal{G}} \mu_g \otimes \delta_g. \tag{2.17}$$

In the remainder, we will often deal with absolutely continuous measures $\mathrm{d}\mu(x, g) = f(x, g)\mathrm{d}x\mathrm{d}g$ in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$, and by abuse of notation, in that case we will simply use the density f to denote the measure μ. For example in the above decomposition, we will use the functions $f_g : \mathbb{R}^d \to \mathbb{R}$, (i.e., $f_g(x) = f(x, g)$) to denote the measures μ_g. We now introduce our static transportation problem which we remark is of interest in its own right.

Static semi-discrete transportation problem Let $\tau > 0$ be a positive time step and let W be as in Example 2.11. For arbitrary measures μ, σ in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$, we define $ADM(\mu, \sigma)$ to be the set of pairs (γ, h) (the admissible pairs) that satisfy:

i) $\gamma = \{\gamma_g\}_{g\in\mathcal{G}}$ where each γ_g is a Borel positive measure on $\mathbb{R}^d \times \mathbb{R}^d$ and whose first marginal $\pi_{1\sharp}\gamma_g$ is equal to μ_g .

ii) $h : \mathbb{R}^d \times \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ is antisymmetric in $\mathcal{G} \times \mathcal{G}$ (i.e., for all $g, g' \in \mathcal{G}$, $x \in \mathbb{R}^d$ we have $h(x, g, g') = -h(x, g', g)$), and it belongs to

$$L^2_{W,K}(\mathbb{R}^d \times \mathcal{G} \times \mathcal{G}) := \Big\{ h \in \mathbb{R}^d \times \mathcal{G} \times \mathcal{G} \to \mathbb{R} \quad : \quad \sum_{g,g'} \int h^2_{gg'} e^{-W} K(g, g')\mathrm{d}x < \infty \Big\}. \tag{2.18}$$

iii) For every $g \in \mathcal{G}$

$$\sigma_g = \pi_{2\#}\gamma_g - \tau \sum_{g'} h_{gg'}(x)K(g, g')e^{-W(x)}. \tag{2.19}$$

The last term on the right hand side of the identity (2.19) must be interpreted as the positive measure on $\mathbb{R}^d$ whose density (with respect to the Lebesgue measure) is given by

$$\tau \sum_{g'} h_{gg'}(x) K(g, g') e^{-W(x)}.$$

In the remainder, we refer to the measures γ_g as *transport plans* and to the functions h as *mass exchange maps*.

A *static transportation cost* between μ, σ is defined by

$$\mathcal{A}^{\mathcal{G},W,\tau}(\mu,\sigma) := \inf_{(\gamma,h)\in ADM(\mu,\sigma)} C_\tau^{W,K}(\gamma, h), \tag{2.20}$$

where

$$C_\tau^{W,K}(\gamma, h) := \sum_{g,g'\in\mathcal{G}} \left(\frac{1}{2\tau} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |x - x'|^2 \mathrm{d}\gamma_g + \frac{\tau}{4} \int_{\mathbb{R}^d} h_{gg'}^2 K(g, g') e^{-W} \mathrm{d}x \right). \tag{2.21}$$

Since the set $ADM(\mu, \sigma)$ may very well be the empty set, we follow the convention that the infimum of a quantity over an empty set is equal to $+\infty$. We use $Opt(\mu, \sigma)$ to denote the set of minimizers of (2.20) when $\mathcal{A}^{\mathcal{G},W,\tau}(\mu, \sigma)$ is finite.

The static semi-discrete optimal transport problem introduced above can be interpreted as an optimal two stage mass transport process from one distribution over $\mathbb{R}^d \times \mathcal{G}$ to another. In the first stage, mass is transported along each fiber of $\mathbb{R}^d$ (i.e., a set of the form $\mathbb{R}^d \times \{g\}$). In the second stage, mass gets exchanged along every fiber of $\mathcal{G}$ (i.e., a set of the form $\{x\} \times \mathcal{G}$). The optimal transport plans and optimal exchange maps (and implicitly the optimal intermediate mass distribution after stage 1) are chosen so as to minimize the sum of two terms: one that corresponds to aggregate quadratic cost in stage one, and the other that corresponds to an average of discrete H^{-1} norms of the mass exchanged during stage two. In Sect. 4, we study the above semi-discrete (static) transport problem mathematically. In particular, we study properties of the set $ADM(\mu, \sigma)$ and characterize $Opt(\mu, \sigma)$ in a way that resembles Brenier's theorem for optimal transport in Euclidean space. Part of the motivation for the definition of this static problem comes from the theoretical desire of recovering the system (1.1) as limit of a JKO scheme relative to some meaningful cost function. While this transport problem is not the same as the dynamic one from Definition 2.6, we believe that they are actually closely related. This is a topic that we may explore in future work.

Let us now return to our aim of defining the gradient descent of the relative entropy energy $\mathcal{E}$ using the minimizing movement scheme. We use the cost function $2\tau\mathcal{A}^{\mathcal{G},W,\tau}$ introduced above to produce the series of iterates in (1.9) for $\mathcal{M} = \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ and $E = \mathcal{E}$. We will assume that the initial datum $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ satisfies $\mathcal{E}(\mu_0) < \infty$. Moreover, we will impose a further technical condition and assume that μ_0 has a probability density f_0 such that

$$\lambda e^{-V} \le f_0 \le \Lambda e^{-V}, \tag{2.22}$$

for some positive constants λ and Λ. Setting $\mu_0^\tau := \mu_0$, we will then let μ_{n+1}^τ be a minimizer of

$$\sigma \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G}) \longmapsto \mathcal{E}(\sigma) + \mathcal{A}^{\mathcal{G},W,\tau}(\mu, \sigma), \tag{2.23}$$

where we set $\mu = \mu_n^\tau$. In Sect. 5 we study properties of the minimization problem (2.23), and in particular provide conditions under which minimizers exist (see Proposition 5.6). It will then be straightforward to see that the resulting iterates must be absolutely continuous with respect to the measure $\mathrm{d}x\mathrm{d}g$, and thus can be written as $\mathrm{d}\mu_n^\tau(x, g) = f_n^\tau(x, g)\mathrm{d}x\mathrm{d}g$. A continuous-time extension of the above iterates is defined via piecewise constant interpolation in time. Namely,

$$f^\tau(t) := f_{n+1}^\tau, \quad t \in (n\tau, (n+1)\tau].$$

Comparing the minimization problems (1.11) and (2.23), we see our semi-discrete transportation cost plays the role of the kinetic energy in the Lagrangian formulation of the JKO scheme.

Our main theoretical result is the following:

Theorem 2.14 *Suppose that f_0 satisfies (2.22), W satisfies the conditions from Example (2.11) and in addition for some constants λ', Λ'*

$$\lambda' e^{-W(x)} \leq e^{-V(x,g)} \leq \Lambda' e^{-W(x)}. \tag{2.24}$$

where $V : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ is a differentiable function in x that also satisfies

$$\sum_g \int_{\mathbb{R}^d} |\nabla_x V(x, g)|^2 e^{-V(x,g)} \mathrm{d}x < \infty.$$

Then, for any sequence $\tau_k \downarrow 0$ there exists a subsequence, not relabeled, for which f^{τ_k} converges to f in $L^2(0; t_F, L^2_{loc}(\mathbb{R}^d \times \mathcal{G}))$ for any $t_F > 0$, where the map $t \in [0, \infty) \to f(t)$ belongs to $L^2_{loc}([0, \infty), W^{1,2}(\mathbb{R}^d \times \mathcal{G}))$ and is a weak solution of (2.14) (see Definition 6.1).

Moreover, for every $t > 0$

$$\lambda e^{-V(x,g)} \leq f(t, x, g) \leq \Lambda e^{-V(x,g)}, \tag{2.25}$$

for almost every (x, g) in $\mathbb{R}^d \times \mathcal{G}$, where λ, Λ are the constants in (2.22).

We prove Theorem 2.14 in Sect. 6.

Remark 2.15 The function

$$f_\infty(x, g) = ce^{-V(x,g)},$$

with c chosen so that

$$\sum_{g\in\mathcal{G}}\int cf_\infty(x,g)\quad \mathrm{d}x=1,$$

is an equilibrium point and solves Eq. (2.14). Consequently, the property described in (2.25) coincides with a well-known maximum principle for the Fokker–Planck equation.

3 Metric and Geometric Properties of W_2

3.1 Proof of Theorem 2.7

1. Let μ_0, μ_1 be two elements in $\mathcal{P}_2(\mathbb{R}^d\times\mathcal{G})$. First, we prove that the infimum in the definition of $W_2^2(\mu_0,\mu_1)$ is finite by exhibiting one solution to the continuity equation connecting μ_0 and μ_1 with finite kinetic energy. One such solution is described as follows.

Let us first assume that μ_0 and μ_1 are supported on the set $B(0,R)\times\mathcal{G}$ for some $R>0$. For each $g\in\mathcal{G}$, let $m_g:=\mu_0(\mathbb{R}^d\times\{g\})$ be the total mass assigned to the fiber $\mathbb{R}^d\times\{g\}$ by μ_0 and let μ_{0g}, μ_{1g} be the positive measures over $\mathbb{R}^d$ defined by

$$\mu_{0g}(A):=\mu_0(A\times\{g\}),\quad \mu_{1g}(A):=\mu_1(A\times\{g\})\quad \forall A\subseteq\mathbb{R}^d,\ \text{Borel}.$$

Also, let $\tilde{\mu}_1$ be the first marginal of the measure μ_1, i.e.,

$$\tilde{\mu}_1(A)=\mu_1(A\times\mathcal{G}),\quad \forall A\subseteq\mathbb{R}^d,\ \text{Borel}.$$

Since the measures μ_{0g} and $m_g\tilde{\mu}_1$ have the same amount of total mass, we can find a solution $t\in[0,1]\mapsto(\nu_{t,g},\nabla_x\phi_t(\cdot,g))$ to the continuity equation on $\mathbb{R}^d$

$$\dot{\nu}_{t,g}+\mathrm{div}_x(\nabla_x\phi_t(\cdot,g)\nu_{t,g})=0,$$

satisfying $\nu_{0,g}=\mu_{0g}$, $\nu_{1,g}=m_g\tilde{\mu}_1$, and

$$\int_0^1\int_{\mathbb{R}^d}|\nabla_x\phi_t(x,g)|^2\mathrm{d}\nu_{t,g}(x)\mathrm{d}t<\infty.$$

On the other hand, notice that for every $g\in\mathcal{G}$ the measure μ_{1g} is absolutely continuous with respect to $\tilde{\mu}_1$, and for $\tilde{\mu}_1$-a.e. x we have

$$\sum_g\frac{\mathrm{d}\mu_{1g}}{\mathrm{d}\tilde{\mu}_1}(x)=1.$$

For each such x, we can find a solution to the discrete continuity equation $t \in [0,1] \mapsto (\gamma_{t,x}, \nabla_g \psi_t(x,\cdot))$

$$\dot{\gamma}_{t,x} + \mathrm{div}_g(\nabla_g \psi_t(x,\cdot) \cdot \hat{\gamma}_{t,x}) = 0$$

satisfying $\gamma_{0,x}(g) = m_g$ and $\gamma_{1,x}(g) = \frac{\mathrm{d}\mu_{1g}}{\mathrm{d}\tilde{\mu}_1}(x)$ for all $g \in \mathcal{G}$, and satisfying

$$\int_0^1 \sum_{g,g'} |\nabla_g \psi_t(x,g,g')|^2 K(g,g') \mathrm{d}\hat{\gamma}_{t,x}(g,g') \mathrm{d}t \leq C,$$

for some constant C that only depends on R. Such solution exists due to assumptions (A0) or (A5) on θ and the fact that discrete optimal transport is well defined in that case (see Maas 2011; Erbar and Maas 2012).

We define

$$\mu_t := \begin{cases} \sum_{g\in\mathcal{G}} m_g d\nu_{2t,g}(x) \otimes \delta_g, & t \in [0,1/2] \\ \sum_{g\in\mathcal{G}} \gamma_{(2t-1),x}(g) \mathrm{d}\tilde{\mu}_1(x) \otimes \delta_g, & t \in [1/2,1] \end{cases}$$

and

$$\phi_t(x,g) := \begin{cases} \phi_{2t}(x,g), & t \in [0,1/2] \\ 0, & t \in [1/2,1] \end{cases} \qquad \psi_t(x,g) := \begin{cases} 0, & t \in [0,1/2] \\ \psi_{2t-1}(g), & t \in [1/2,1] \end{cases}$$

It is straightforward to verify that $t \in [0,1] \mapsto (\mu_t, \nabla_x \phi_t, \nabla_g \psi_t)$ solves the semi-discrete continuity equation, connects μ_0 and μ_1, and has finite kinetic energy.

If μ_0, μ_1 are not compactly supported as assumed above, then pick *any* $\tilde{\mu}_0, \tilde{\mu}_1$ compactly supported satisfying

$$\mu_0(\mathbb{R}^d \times \{g\}) = \tilde{\mu}_0(\mathbb{R}^d \times \{g\}), \quad \mu_1(\mathbb{R}^d \times \{g\}) = \tilde{\mu}_1(\mathbb{R}^d \times \{g\}), \quad \forall g \in \mathcal{G}.$$

One can then dynamically transport mass from μ_0 to $\tilde{\mu}_0$ restricting the transport to each fiber $\mathbb{R}^d \times \{g\}$ using a continuity equation with finite kinetic energy on each fiber (this is simply OT in $\mathbb{R}^d$). Then, one can transport dynamically from $\tilde{\mu}_0$ to $\tilde{\mu}_1$ (as done above) and finally transport dynamically from $\tilde{\mu}_1$ to μ_1 restricting the transport to each fiber $\mathbb{R}^d \times \{g\}$ (again doing OT just on $\mathbb{R}^d$).

2. Let us now show that $W_2(\mu_0,\mu_1) = 0$ if and only if $\mu_0 = \mu_1$. First notice that if $\mu_0 = \mu_1$ we may take $\phi_t \equiv 0$, $\psi_t \equiv 0$ and $\mu_t = \mu_0$ for all $t \in [0,1]$. Then, it is clear that $t \in [0,1] \to (\mu_t, \phi_t)$ solves the continuity equation, has zero kinetic energy, and connects μ_0 and μ_1, from where it follows that $W_2^2(\mu_0,\mu_1) = 0$.

Now let us suppose that $W_2^2(\mu_0,\mu_1) = 0$. We want to show that $\mu_0 = \mu_1$. Fix an arbitrary test function $\zeta : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ where $\zeta(\cdot, g)$ is smooth and compactly supported for all $g \in \mathcal{G}$. From the condition $W_2(\mu_0,\mu_1) = 0$, we see that for every $\varepsilon > 0$ there is a solution to the continuity equation $t \mapsto (\mu_t, \nabla_x \phi_t, \nabla_g \psi_t)$ connecting

μ_0 and μ_1 with kinetic energy less than ε, i.e.,

$$\mathcal{K} := \int_0^1 \left(\int_{\mathbb{R}^d} \sum_g |\nabla_x \phi_t(x,g)|^2 \mathrm{d}\mu_t(g,x) + \int_{\mathbb{R}^d} \sum_{g,g'} \nabla_g \psi_t(x,g,g')^2 K(g,g') \mathrm{d}\hat{\mu}_t(x,g,g') \right) \mathrm{d}t \leq \varepsilon.$$

Using (2.6) (after integration over $t \in [0,1]$) for the above test function ζ, we conclude that

$$\left| \sum_{g\in\mathcal{G}} \int_{\mathbb{R}^d} \zeta(x,g) \mathrm{d}\mu_1(x,g) - \sum_{g\in\mathcal{G}} \int_{\mathbb{R}^d} \zeta(x,g) \mathrm{d}\mu_0(x,g) \right| \leq C_\zeta \sqrt{\mathcal{K}} \leq C_\zeta \sqrt{\varepsilon}$$

where C_ζ is a constant that only depends on the test function ζ. Given that ε was arbitrary we can conclude that

$$\sum_{g\in\mathcal{G}} \int_{\mathbb{R}^d} \zeta(x,g) \mathrm{d}\mu_1(x,g) = \sum_{g\in\mathcal{G}} \int_{\mathbb{R}^d} \zeta(x,g) \mathrm{d}\mu_0(x,g).$$

Finally, since ζ was an arbitrary smooth compactly supported test function we deduce that $\mu_0 = \mu_1$.

3. Next, we show that $W_2(\mu_0,\mu_1) = W_2(\mu_1,\mu_0)$. To see this, simply notice that any solution $t \in [0,1] \mapsto (\mu_t, \nabla_x\phi_t, \nabla_g\psi_t)$ to the continuity equation starting at μ_0 and ending at μ_1, can be reverted in time $t \in [0,1] \to (\mu_{1-t}, -\nabla_x\phi_{1-t}, -\nabla_g\psi_{1-t})$ producing in this way a solution to the continuity equation that starts at μ_1 and ends at μ_0, and has the exact same kinetic energy as the original curve.

4. Lastly, we prove the triangle inequality. First, we observe that after a standard reparametrization (of time) by arc-length it follows that for every $\mu, \tilde{\mu} \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ and every $T > 0$,

$$W_2(\mu,\tilde{\mu}) = \inf_{t\in[0,T]\mapsto(\mu_t,\nabla_x\phi_t,\nabla_g\psi_t)} \int_0^T \left(\int_{\mathbb{R}^d} \sum_{g\in\mathcal{G}} |\nabla_x\phi_t(x,g)|^2 \mathrm{d}\mu_t(x,g) + \int_{\mathbb{R}^d} \sum_{g,g'} (\nabla_g\psi_t(x,g,g'))^2 K(g,g') \mathrm{d}\hat{\mu}_t(x,g,g') \right)^{1/2} \mathrm{d}t, \tag{3.1}$$

where the inf ranges over all solutions $t \in [0,T] \mapsto (\mu_t, \nabla_x\phi_t, \nabla_g\psi_t)$ to the semi-discrete continuity equation with $\mu_0 = \mu$ and $\mu_T = \tilde{\mu}$

Let now μ_0, μ_1, μ_2 be arbitrary elements in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$. From (3.1), for any $\varepsilon > 0$ we may consider $t \in [0,1] \mapsto (\mu_t, \nabla_x\phi_t, \nabla_g\psi_t)$ and $t \in [0,1] \mapsto (\widetilde{\mu}_t, \nabla_x\widetilde{\phi}_t, \nabla_g\widetilde{\psi}_t)$ solutions to the semi-discrete continuity equation satisfying $\mu_0 = \mu_0, \mu_1 = \mu_1 = \widetilde{\mu}_0$,

$\widetilde{\mu}_1 = \mu_2$ and

$$\int_0^1 \left(\frac{1}{2} \sum_g \int_{\mathbb{R}^d} |\nabla_x \phi_t(x, g)|^2 \mathrm{d}\mu_t(x, g) + \int_{\mathbb{R}^d} \sum_{g,g'} \nabla_g \psi_t(x, g, g')^2 K(g, g') \mathrm{d}\hat{\mu}_t(x, g, g') \right)^{1/2} \mathrm{d}t \leq W_2(\mu_0, \mu_1) + \varepsilon,$$

$$\int_0^1 \left(\frac{1}{2} \sum_g \int_{\mathbb{R}^d} |\nabla_x \widetilde{\phi}_t(x, g)|^2 d\widetilde{\mu}_t(x, g) + \int_{\mathbb{R}^d} \sum_{g,g'} \nabla_g \widetilde{\psi}_t(x, g, g')^2 K(g, g') \mathrm{d}\hat{\widetilde{\mu}}_t(x, g, g') \right)^{1/2} \mathrm{d}t \leq W_2(\mu_1, \mu_2) + \varepsilon.$$

We then consider

$$\gamma_t := \begin{cases} \mu_t, & t \in [0, 1] \\ \widetilde{\mu}_{t-1}, & t \in [1, 2] \end{cases}$$

and the potentials

$$\alpha_t(x, g) := \begin{cases} \phi_t(x, g), & t \in [0, 1] \\ \widetilde{\phi}_{t-1}(x, g), & t \in [1, 2] \end{cases} \qquad \beta_t(x, g) := \begin{cases} \psi_t(x, g), & t \in [0, 1] \\ \widetilde{\psi}_{t-1}(x, g), & t \in [1, 2]. \end{cases}$$

It follows that $t \in [0, 2] \mapsto (\gamma_t, \nabla_x \alpha_t, \nabla_g \beta_t)$ solves the semi-discrete continuity equation, connects μ_0 and μ_2, and satisfies

$$\int_0^2 \left(\frac{1}{2} \sum_g \int_{\mathbb{R}^d} |\nabla_x \widetilde{\phi}_t(x, g)|^2 \mathrm{d}\widetilde{\mu}_t(x, g) + \int_{\mathbb{R}^d} \sum_{g,g'} \nabla_g \widetilde{\psi}_t(x, g, g')^2 K(g, g') \mathrm{d}\hat{\widetilde{\mu}}_t(x, g, g') \right)^{1/2} \mathrm{d}t \leq W_2(\mu_0, \mu_1) + W_1(\mu_1, \mu_2) + 2\varepsilon.$$

From (3.1), it follows that $W_2(\mu_0, \mu_2) \leq W_2(\mu_0, \mu_1) + W_2(\mu_1, \mu_2) + 2\varepsilon$. Since $\varepsilon > 0$ was arbitrary the result now follows.

3.2 Tangent Plane Characterization

In this section, we provide concrete conditions under which the statement of Theorem 2.10 can be made rigorous. The bottom line is that the arguments presented in this section motivate the formal characterization for the tangent plane $\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$, i.e., infinitesimal curves on $\mathcal{P}_2(\mathcal{G} \times \mathbb{R}^d)$ passing through μ. The main result of this section can be interpreted as a minimal selection principle for the potentials (ϕ, ψ) driving a given solution to the continuity equation. Some of the results proved below will be used again later on when we get to analyze the static semi-discrete transport problem from Sect. 2.5.

Throughout this section, we work with measures of the form $\mathrm{d}\mu(x, g) = f(x, g)\mathrm{d}x\mathrm{d}g$ for a density function f satisfying basic boundedness conditions. We also use the following spaces of potentials:

$$\Phi := \left\{ \varepsilon \in L^2_c(\mathbb{R}^d \times \mathcal{G}) \text{ s.t. } \int_{\mathbb{R}^d} \varepsilon(x, g)\mathrm{d}x = 0 \quad \forall g, \quad \sum_g \varepsilon(x, g) = 0 \text{ a.e. } x \in \mathbb{R}^d \right\}, \tag{3.2}$$

where $L^2_c(\mathbb{R}^d \times \mathcal{G})$ stands for the space of $L^2(\mathbb{R}^d \times \mathcal{G})$ functions with compact support (i.e., almost everywhere equal to zero outside a set of the form $B(0, R) \times \mathcal{G}$), and also

$$\Phi^\perp := \left\{ \varphi \in L^2_{loc}(\mathbb{R}^d \times \mathcal{G}) \text{ s.t. } \int_{\mathbb{R}^d} \sum_g \varphi(x, g)\varepsilon(x, g)\mathrm{d}x = 0, \quad \forall \varepsilon \in \Phi \right\}.$$

Lemma 3.1 *Let $f : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ be a probability density such that in every compact subset of $\mathbb{R}^d \times \mathcal{G}$ is bounded and bounded away from zero. Let ϕ, ψ be two potentials belonging to $L^2_{loc}(\mathbb{R}^d \times \mathcal{G})$ for which*

$$\int_{\mathbb{R}^d} \sum_g |\nabla_x \phi(x, g)|^2 f(x, g)\mathrm{d}x + \int_{\mathbb{R}^d} \sum_{g,g'} |\nabla_g \psi(x, g, g')|^2 K(g, g')\hat{f}(x, g, g')\mathrm{d}x < \infty.$$

Consider the minimization problem:

$$\inf_{\tilde{\phi}, \tilde{\psi} \in L^2_{loc}(\mathbb{R}^d \times \mathcal{G})} \int_{\mathbb{R}^d} \sum_g |\nabla_x \tilde{\phi}(x, g)|^2 f(x, g)\mathrm{d}x + \int_{\mathbb{R}^d} \sum_{g,g'} |\nabla_g \tilde{\psi}(x, g, g')|^2 K(g, g') \hat{f}(x, g, g')\mathrm{d}x \tag{3.3}$$

subject to

$$\mathrm{div}_x(f \nabla_x \tilde{\phi}) + \mathrm{div}_g(\hat{f} \nabla_g \tilde{\psi}) = \mathrm{div}_x(f \nabla_x \phi) + \mathrm{div}_g(\hat{f} \nabla_g \psi),$$

where the equality must be interpreted in the sense of distributions.

Then, there exists a minimizing pair $\tilde{\phi}, \tilde{\psi}$ for the above problem. In addition, any minimizing pair must satisfy $\tilde{\phi} - \tilde{\psi} \in \Phi^\perp$.

***Proof* 1.** Let us start by proving the existence of minimizers. First, we consider the slightly modified problem

$$\inf_{\tilde{\phi},h} \int_{\mathbb{R}^d} \sum_g |\nabla_x \tilde{\phi}(x,g)|^2 f(x,g)\mathrm{d}x + \int_{\mathbb{R}^d} \sum_{g,g'} |h_{gg'}(x)|^2 K(g,g')\hat{f}(x,g,g')\mathrm{d}x \quad (3.4)$$

subject to

$$\mathrm{div}_x(f\nabla_x\tilde{\phi}) + \mathrm{div}_g(\hat{f}\cdot h) = \mathrm{div}_x(f\nabla_x\phi) + \mathrm{div}_g(\hat{f}\nabla_g\psi),$$

where the minimization is now over pairs $(\tilde{\phi}, h)$ for $\tilde{\phi}$ as in problem (3.3) and $h \in L^2_{loc}(\mathbb{R}^d \times \mathcal{G} \times \mathcal{G})$ an antisymmetric function on $\mathcal{G}\times\mathcal{G}$ (i.e., $h_{gg'}(x) = -h_{g'g}(x)$ for every x, g, g'). Existence of solutions to (3.4) follows immediately from the direct method of the calculus of variations. From a solution $(\tilde{\phi}, h)$ to problem (3.4) we now construct a solution to (3.3). Fix $x \in \mathbb{R}^d$. Thanks to Proposition 2.1, there exists a solution $\tilde{\psi}(x,\cdot) = \tilde{\psi}_x$ to the graph PDE

$$\mathrm{div}_g(\nabla_g\tilde{\psi}_x\hat{f}_x) = \mathrm{div}_g(h_x\hat{f}_x),$$

which satisfies $\sum_g \tilde{\psi}_x(g) = 0$. Following the proof of Proposition 2.1 and using the fact that $\sum_g \tilde{\psi}_x(g) = 0$, we can conclude that there exists a constant $C_x > 0$ for which

$$\sum_g |\tilde{\psi}(x,g)|^2 \le C_x \sum_{g,g'} |\nabla_g\tilde{\psi}(x,g,g')|^2 K(g,g')\hat{f}(x,g,g'). \quad (3.5)$$

The constant C_x can be assumed to be uniform on compact subsets of $\mathbb{R}^d$ thanks to the assumptions on θ and the fact that in each compact subset of $\mathbb{R}^d\times\mathcal{G}$ the function f is assumed to be bounded and bounded away from zero. Using (2.1), we obtain

$$\begin{aligned}\sum_{gg'} |\nabla_g\tilde{\psi}_x|^2 K(g,g')\hat{f}_x(g,g') &= -\sum_g \mathrm{div}_g(\nabla_g\tilde{\psi}_x\hat{f}_x)\tilde{\psi}_x = -\sum_g \mathrm{div}_g(h_x\hat{f}_x)\tilde{\psi}_x \\ &= \sum_{gg'} \nabla_g\tilde{\psi}_x\cdot h_x K(g,g')\hat{f}_x(g,g'),\end{aligned}$$

and thus, from Cauchy–Schwarz inequality

$$\sum_{gg'} |\nabla_g\tilde{\psi}_x(g,g')|^2 K(g,g')\hat{f}_x(g,g') \le \sum_{gg'} |h_x(g,g')|^2 K(g,g')\hat{f}_x(g,g').$$

The above implies that $(\tilde{\phi}, \nabla_g\tilde{\psi})$ is also a solution to (3.4). Given that $\tilde{\psi}$ is in L^2_{loc} thanks to (3.5), we deduce that $(\tilde{\phi}, \tilde{\psi})$ is a minimizing pair for (3.3).

2. Let $(\tilde{\phi}, \tilde{\psi})$ be an arbitrary minimizing pair. Let ε be an arbitrary element in Φ and pick a specific measurable representative for it (which we also denote by ε). For each fixed g consider the PDE (in x)

$$\varepsilon(\cdot, g) = \mathrm{div}_x(f(\cdot, g)\nabla_x \eta(\cdot, g)). \tag{3.6}$$

Existence of a solution $\eta(\cdot, g)$ in $L^2_{loc}(\mathbb{R}^d)$ follows from standard arguments in the theory of elliptic PDEs, given that ε has compact support and that in each compact subset of $\mathbb{R}^d \times \mathcal{G}$ f is bounded and bounded away from zero. Also, let x be a Lebesgue point for all the functions $\varepsilon(\cdot, g)$, and consider the graph PDE

$$-\varepsilon(x, \cdot) = \mathrm{div}_g(\hat{f}_x \cdot \nabla_g \beta(x, \cdot)). \tag{3.7}$$

This equation has a unique solution (that we denote by $\beta(x, \cdot)$) that averages to zero according to Lemma 2.1 (given that $\varepsilon(x, \cdot)$ has average zero). Moreover, the function β can be seen to be in L^2_{loc} using the inequalities from Proposition 2.1.

Now, for each $s \in \mathbb{R}$ consider the perturbed potentials:

$$\begin{aligned}\phi_s(x, g) &:= \tilde{\phi}(x, g) + s\eta(x, g),\\ \psi_s(x, g) &:= \tilde{\psi}(x, g) + s\beta(x, g),\end{aligned}$$

and notice that

$$\begin{aligned}&\mathrm{div}_x(f\nabla_x\phi_s) = \mathrm{div}_x(f\nabla_x\tilde{\phi}) + s\mathrm{div}_x(f\nabla_x\eta) = \mathrm{div}_x(f\nabla_x\tilde{\phi}) + s\varepsilon\\ &\mathrm{div}_g(\hat{f}\cdot\nabla_g\psi_s) = \mathrm{div}_g(\hat{f}\cdot\nabla_g\tilde{\psi}) + s\mathrm{div}_g(\hat{f}\cdot\nabla_g\beta) = \mathrm{div}_g(\hat{f}\cdot\nabla_g\tilde{\psi}) - s\varepsilon,\end{aligned}$$

so that in particular, for every $s \in \mathbb{R}$,

The pair (ϕ_s, ψ_s) is admissible in the minimization of (3.3). Let $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ be the function

$$\mathcal{K}(s) := \sum_g \int_{\mathbb{R}^d} |\nabla_x\phi_s|^2 f(x, g)\mathrm{d}x + \sum_{g,g'} \int_{\mathbb{R}^d} (\psi_s(x, g) - \psi_s(x, g'))^2 K(g, g')\hat{f}(x, g, g')\mathrm{d}x,$$

which is minimized at $s = 0$ by definition of $(\tilde{\phi}, \tilde{\psi})$. Computing $\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{K}(s)$ and evaluating at $s = 0$, we deduce that

$$\begin{aligned}0 &= \sum_g \int_{\mathbb{R}^d} \nabla_x\eta(x, g)\cdot\nabla_x\tilde{\phi}(x, g) f(x, g)\mathrm{d}x\\ &\quad + \int_{\mathbb{R}^d} \sum_{g,g'} (\beta(x, g) - \beta(x, g'))(\tilde{\psi}(x, g) - \tilde{\psi}(x, g'))K(g, g')\hat{f}(x, g, g')\mathrm{d}x\\ &= \sum_g \int_{\mathbb{R}^d} \varepsilon(x, g)\tilde{\phi}(x, g)\mathrm{d}x - \int_{\mathbb{R}^d} \sum_g \varepsilon(x, g)\tilde{\psi}(x, g)\mathrm{d}x\\ &= \sum_g \int_{\mathbb{R}^d} \varepsilon(x, g)(\tilde{\phi}(x, g) - \tilde{\psi}(x, g))\mathrm{d}x\end{aligned}$$

where the second equality follows from the fact that $\eta(\cdot, g)$ solves (3.6) and $\beta(x, \cdot)$ solves (3.7). Since $\varepsilon \in \Phi$ was arbitrary, it follows that $\tilde{\phi} - \tilde{\psi}$ belongs to $\Phi^{\perp}$ as we wanted to show.

□

Lemma 3.2 *For any φ in $\Phi^{\perp}$, there exists $\varphi_1 : \mathbb{R}^d \to \mathbb{R}$ in $L^2_{loc}(\mathbb{R}^d)$ and $\varphi_2 : \mathcal{G} \to \mathbb{R}$ such that*

$$\varphi(x, g) = \varphi_1(x) + \varphi_2(g), \quad \forall g \in \mathcal{G}, \quad a.e.\ x \in \mathbb{R}^d.$$

Conversely, if φ admits the above decomposition then $\varphi \in \Phi^{\perp}$.

Proof Let $\varphi \in \Phi^{\perp}$ and fix a Lebesgue point x_0 for all the functions $\varphi(\cdot, \tilde{g})$. Let

$$\varphi_2(\tilde{g}) := \varphi(x_0, \tilde{g}), \quad \tilde{g} \in \mathcal{G}.$$

Observe that from Fubini's theorem any function that is independent of x belongs to $\Phi^{\perp}$, and thus, φ_2 must be contained in $\Phi^{\perp}$. Define now the function

$$\varphi_1(\tilde{x}, \tilde{g}) := \varphi(\tilde{x}, \tilde{g}) - \varphi_2(\tilde{g}).$$

To complete our proof, we must show that φ_1 does not depend on $\tilde{g}$. For this purpose, let x be an arbitrary Lebesgue point for all the functions $\varphi(\cdot, \tilde{g})$. Fix $g, g' \in \mathcal{G}$. Let $r > 0$ and consider the test function

$$\varepsilon_r := \xi^r_{x,g} - \xi^r_{x,g'} - \xi^r_{x_0,g} + \xi^r_{x_0,g'}, \tag{3.8}$$

where $\xi^r_{x,g} : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ is given by

$$\xi^r_{x,g}(\tilde{x}, \tilde{g}) := \frac{1}{|B(x, r)|} \mathbb{1}_{B(x,r)}(\tilde{x}) \mathbb{1}_{\{\tilde{g}=g\}}.$$

Notice that by construction ε_r is contained in Φ. Also, since φ and φ_2 are contained in $\Phi^{\perp}$, φ_1 is contained in $\Phi^{\perp}$ too. Hence,

$$\begin{aligned} 0 &= \sum_g \int_{\mathbb{R}^d} \varepsilon_r(\tilde{x}, \tilde{g}) \varphi_1(\tilde{x}, \tilde{g}) \mathrm{d}\tilde{x} \\ &= \frac{1}{|B(x, r)|} \int_{B(x,r)} \varphi_1(\tilde{x}, g) \mathrm{d}\tilde{x} - \frac{1}{|B(x, r)|} \int_{B(x,r)} \varphi_1(\tilde{x}, g') \mathrm{d}\tilde{x} \\ &\quad - \frac{1}{|B(x_0, r)|} \int_{B(x_0,r)} \varphi_1(\tilde{x}, g) \mathrm{d}\tilde{x} + \frac{1}{|B(x_0, r)|} \int_{B(x_0,r)} \varphi_1(\tilde{x}, g') \mathrm{d}\tilde{x}. \end{aligned}$$

We may now take $r \to 0$ and use the fact that x_0 and x were assumed to be Lebesgue points for the functions $\varphi(\cdot, g)$ and $\varphi(\cdot, g')$ (thus also for φ_1) to conclude that

$$0 = \varphi_1(x, g) - \varphi_1(x, g') - \varphi_1(x_0, g) + \varphi_1(x_0, g').$$

By construction, $\varphi_1(x_0, g) = \varphi_1(x_0, g') = 0$. Consequently, we deduce that

$$\varphi_1(x, g) = \varphi_1(x, g').$$

Since x, g and g' were arbitrary, we conclude that φ can be written as the sum of a function of x only and a function of g only.

The converse statement is a direct consequence of Fubini's theorem.

□

Remark 3.3 Notice that from the proof of Lemma 3.2, it actually follows that if $\varphi \in L^2_{loc}(\mathbb{R}^d \times \mathcal{G})$ is such that $\sum_g \int_{\mathbb{R}^d} \varphi(x, g)\varepsilon(x, g)\mathrm{d}x$ for all ε of the form (3.8) then φ can be written as $\varphi(x, g) = \varphi_1(x) + \varphi_2(g)$ (and in particular it follows that $\varphi \in \Phi^{\perp}$). We will use this observation in Proposition 4.5.

We may now combine the previous two lemmas to deduce the following minimum selection principle providing concrete support to Theorem 2.10.

Proposition 3.4 *Under the same assumptions on f from Lemma 3.1, there exists a minimizing pair for problem (3.3) of the form (φ, φ).*

Proof Consider an arbitrary minimizing pair for problem (3.3). By Lemma 3.1 we know that this pair must satisfy $\tilde{\phi} - \tilde{\psi} \in \Phi^{\perp}$, and by Lemma 3.2 we can conclude that

$$\tilde{\phi} - \tilde{\psi} = \varphi_1 + \varphi_2,$$

for some $\varphi_1 : \mathbb{R}^d \to \mathbb{R}$ in $L^2_{loc}(\mathbb{R}^d \times \mathcal{G})$ and $\varphi_2 : \mathcal{G} \to \mathbb{R}$. Consider now the function

$$\varphi(x, g) := \tilde{\phi}(x, g) - \varphi_2(g)$$

and notice that we can also write it as

$$\varphi(x, g) = \tilde{\psi}(x, g) + \varphi_1(x).$$

It follows that

$$\nabla_x \varphi = \nabla_x \tilde{\phi}, \quad \nabla_g \varphi = \nabla_g \tilde{\psi}.$$

Due to the above relationship, it follows that (φ, φ) is admissible for the optimization problem (3.3) and that it achieves the same value as that of the minimizing pair $(\tilde{\phi}, \tilde{\psi})$. Therefore, (φ, φ) solves (3.3).

□

3.3 A Formal Computation of the Acceleration of a Curve in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$: Geodesic Equations and Accelerated Methods for Optimization

In this section, we present a heuristic argument that motivates the discussion in Sect. 2.4. The heuristics are based on the formal computation of the acceleration of a given curve in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$.

Let us recall that the covariant derivative $\nabla_{\dot{q}(t)}$ along a smooth curve $t \mapsto q(t)$ on a smooth Riemannian manifold $\mathcal{M}$ is a mapping taking vector fields into vector fields along the curve q. This mapping makes sense of the idea of differentiation of a vector field $t \mapsto p(t)$ along the curve in a way that is compatible with the Riemannian structure of $\mathcal{M}$. We will now recall a formula from Riemannian geometry that characterizes $\nabla_{\dot{q}}\dot{q}$ (the covariant derivative of the velocity of the curve, i.e., the acceleration of the curve) in terms of variations of the kinetic energy. For that purpose, we let $t \in [0, T] \mapsto q(t)$ be a fixed smooth curve in $\mathcal{M}$. We recall that a (smooth) *proper variation* of the curve q is a smooth function $\alpha : (s, t) \in (-\varepsilon, \varepsilon) \times [0, T] \to \mathcal{M}$ satisfying $\alpha(0, t) = q(t)$ for all $t \in [0, T]$ and $\alpha(s, 0) = q(0)$, $\alpha(s, T) = q(T)$ for all $s \in (-\varepsilon, \varepsilon)$. In particular, the maps $t \in [0, T] \mapsto \alpha(s, t)$ can be understood as describing nearby curves to the original curve q, and in that light, the vector field $v(t) = \frac{\partial}{\partial s}\alpha(0, t)$ known as the *variational field* of α (which is a vector field along the curve q) describes an infinitesimal deformation of the curve maintaining its endpoints anchored. A well-known result in Riemannian geometry (e.g., Proposition 2.4 in Chapter 9 in do Carmo 1992) states that:

$$\frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0}\left(\frac{1}{2}\int_0^T \left|\frac{\partial}{\partial t}\alpha(s,t)\right|^2_{q(t)} \mathrm{d}t\right) = -\int_0^T \left\langle v(t), \nabla_{\dot{q}}\dot{q}\right\rangle_{q(t)} \mathrm{d}t. \tag{3.9}$$

Since in the above one can take arbitrary variations of q, the previous expression indeed characterizes $\nabla_{\dot{q}}\dot{q}$ completely: regardless of the smooth proper variation taken, the first variation of the kinetic energy (the left hand side) must match the right hand side which is expressed in terms of the corresponding variational field and the acceleration of the curve $\nabla_{\dot{q}}\dot{q}$.

Using the above discussion as motivation, let us now consider a curve $t \in [0, T] \mapsto f_t \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ and let us provide a formal definition for its acceleration; here and in what follows we identify a measure $\mathrm{d}\mu(x, g) = f(x, g)\mathrm{d}x\mathrm{d}g$ with its density, and let $(\nabla_x\varphi_t, \nabla_g\varphi_t)$ be the velocity of the curve at time t. Let $(s, t) \in (-\varepsilon, \varepsilon) \times [0, T] \mapsto (f_{s,t}, \nabla_x\varphi_{s,t}, \nabla_g\varphi_{s,t})$ be a proper variation of $t \mapsto f_t$. Namely, we assume $(f_{0,t}, \varphi_{0,t}) = (f_t, \varphi_t)$ for all t, and $f_{s,0} = f_0$, $f_{s,T} = f_T$ for all $s \in (-\varepsilon, \varepsilon)$. We use $\psi_{s,t}$ to denote a potential associated to the curve $s \in (-\varepsilon, \varepsilon) \mapsto f_{s,t}$. The map $t \in [0, T] \mapsto \psi_t := \psi_{0,t}$ can then be interpreted as the corresponding variational field of the variation $(s, t) \mapsto f_{s,t}$. We assume all functions are smooth, and smooth in s and t so that we can take derivatives in x, s, t at will.

Relative to the proper variation introduced above, we define

$$F(s) := \frac{1}{2}\int_0^T \left(\sum_g \int_{\mathbb{R}^d} |\nabla_x \varphi_{s,t}|^2 f_{s,t}(x,g)\mathrm{d}x \right.$$
$$\left. + \int_0^T \sum_{g,g'} \int_{\mathbb{R}^d} |\nabla_g \varphi_{s,t}|^2 \hat{f}_{s,t}(x,g,g')\mathrm{d}x \right) \mathrm{d}t,$$

for $s \in (-\varepsilon, \varepsilon)$, which according to (2.9) can also be written as

$$\frac{1}{2}\int_0^T \langle \varphi_{s,t}, \varphi_{s,t} \rangle_{f_{s,t}} \mathrm{d}t.$$

We show that

$$\frac{\mathrm{d}}{\mathrm{d}s} F(s)\Big|_{s=0} = -\int_0^T \left\langle \psi_t, \partial_t \phi_t + \frac{1}{2}|\nabla_x \varphi_t|^2 \right.$$
$$\left. + \sum_{g'} |\nabla_g \varphi_t(\cdot,\cdot,g')|^2 \partial_1 \theta(f_t(\cdot,\cdot), f_t(\cdot,g')) \right\rangle_{f_t} \mathrm{d}t, \tag{3.10}$$

which when compared to (3.9) motivates the definition of the acceleration of the curve $t \in [0,T] \mapsto (f_t, \nabla_x \varphi_t, \nabla_g \varphi_t)$ at time t as the potential:

$$(x,g) \in \mathbb{R}^d \times \mathcal{G} \mapsto \partial_t \varphi_t(x,g) + \frac{1}{2}|\nabla_x \varphi(x,g)|^2$$
$$+ \sum_{g'} |\nabla_g \varphi_t(x,g,g')|^2 \partial_1 \theta(f_t(x,g), f_t(x,g')).$$

Notice that in turn, the above definition motivates the geodesic equations given in (2.15), as well as the (continuous time) accelerated scheme in (2.16) for the optimization of the relative entropy defined in (2.10) (using the expression for its gradient that we found in Sect. (2.3)) in light of the discussion in Sect. 2.4.

We now formally obtain (3.10). First,

$$\frac{\mathrm{d}}{\mathrm{d}s} F(s) = \int_0^T \sum_g \int_{\mathbb{R}^d} (\nabla_x \partial_s \varphi_{s,t} \cdot \nabla_x \varphi_{s,t}) f_{s,t}(x,g)\mathrm{d}t$$
$$+ \int_0^T \sum_{g,g'} \int_{\mathbb{R}^d} (\nabla_g \partial_s \varphi_{s,t} \cdot \nabla_g \varphi_{s,t}) K(g,g') \hat{f}_{s,t}(x,g,g')\mathrm{d}t$$
$$+ \frac{1}{2}\int_0^T \sum_g \int_{\mathbb{R}^d} |\nabla_x \varphi_{s,t}|^2 \partial_s f_{s,t}(x,g)\mathrm{d}t + \frac{1}{2}\int_0^T \sum_{g,g'} \int_{\mathbb{R}^d} |\nabla_g \varphi_{s,t}|^2$$

$$K(g,g')\partial_s \hat{f}_{s,t}(x,g,g')\mathrm{d}t = \int_0^T \sum_g \int_{\mathbb{R}^d} (\nabla_x \partial_s \varphi_{s,t} \cdot \nabla_x \varphi_{s,t}) f_{s,t}(x,g)\mathrm{d}t$$

$$+\int_0^T \sum_{g,g'} \int_{\mathbb{R}^d} (\nabla_g \partial_s \varphi_{s,t} \cdot \nabla_g \varphi_{s,t}) K(g,g') \hat{f}_{s,t}(x,g,g') \mathrm{d}t$$
$$+\frac{1}{2}\int_0^T \sum_{g} \int_{\mathbb{R}^d} |\nabla_x \varphi_{s,t}|^2 \partial_s f_{s,t}(x,g) \mathrm{d}t$$
$$+\int_0^T \sum_{g,g'} \int_{\mathbb{R}^d} |\nabla_g \varphi_{s,t}|^2 K(g,g') \partial_1 \theta(f_{s,t}(x,g), f_{s,t}(x,g')) \partial_s f_{s,t}(x,g) \mathrm{d}t. \tag{3.11}$$

On the other hand, integration by parts and the fact that $\partial_s \varphi(0,s) = 0$ and $\partial_s \varphi(s,T) = 0$ for all s (because the variation is proper) lead to

$$\int_0^T \partial_t \varphi_{s,t}(x,g) \partial_s f_{s,t}(x,g) \mathrm{d}t = -\int_0^T \varphi_{s,t}(x,g) \partial_s \partial_t f_{s,t}(x,g) \mathrm{d}t$$
$$= -\frac{\mathrm{d}}{\mathrm{d}s}\left(\int_0^T \varphi_{s,t} \partial_t f_{s,t} \mathrm{d}t\right) + \int_0^T \partial_s \varphi_{s,t} \partial_t f_{s,t} \mathrm{d}t.$$

After integration over x, g and using the continuity equation, the above implies

$$\int_0^T \sum_{g} \int_{\mathbb{R}^d} \partial_t \varphi_{s,t}(x,g) \partial_s f_{s,t}(x,g) \mathrm{d}x \mathrm{d}t$$
$$= -\frac{\mathrm{d}}{\mathrm{d}s}\left(\int_0^T \left(\sum_{g} \int_{\mathbb{R}^d} |\nabla_x \varphi_{s,t}|^2 f_{s,t} \mathrm{d}x + \sum_{g,g'} \int_{\mathbb{R}^d} |\nabla_g \varphi_{s,t}|^2 \hat{f}_{s,t} \mathrm{d}x\right) \mathrm{d}t\right)$$
$$+ \int_0^T \sum_{g} \int_{\mathbb{R}^d} \partial_s \varphi_{s,t} \partial_t f_{s,t} \mathrm{d}x \mathrm{d}t$$
$$= -2\frac{\mathrm{d}}{\mathrm{d}s} F(s)$$
$$+ \int_0^T \sum_{g} \int_{\mathbb{R}^d} (\nabla_x \partial_s \varphi_{s,t} \cdot \nabla_x \varphi_{s,t}) f_{s,t}(x,g) \mathrm{d}t$$
$$+ \int_0^T \sum_{g,g'} \int_{\mathbb{R}^d} (\nabla_g \partial_s \varphi_{s,t} \cdot \nabla_g \varphi_{s,t}) \hat{f}_{s,t}(x,g,g') \mathrm{d}t. \tag{3.12}$$

Combining (3.11) and (3.12), we deduce that $\frac{\mathrm{d}}{\mathrm{d}s} F(s)$ can be written as:

$$-\int_0^T \int_{\mathbb{R}^d} \sum_{g} \left(\partial_t \varphi_{s,t} + \frac{1}{2}|\nabla_x \varphi_{s,t}|^2 + \sum_{g'} |\nabla_g \varphi_{s,t}|^2 K(g,g') \partial_1 \theta(f_{s,t}(x,g), f_{s,t}(x,g'))\right)$$
$$\partial_s f_{s,t}(x,g) \mathrm{d}x \mathrm{d}t.$$

Finally, at $s = 0$ we have $\partial_s f_{s,t} = -\mathrm{div}_x(\nabla_x \psi_t f_t) - \mathrm{div}_g(\nabla_g \psi_t \hat{f}_t)$, and thus (3.10) follows combining the above with the semi-discrete continuity equation.

4 Properties of Minimizing Pairs of the Static Semi-discrete Optimal Transport Problem

In this section, we study the minimizers of the static semi-discrete transportation problem that we introduced in Sect. 2.5. Some of the results presented in this section will be used in the sequel while others are of interest on their own. We seek to reproduce the result of Brenier Ambrosio and Gigli (2013), Theorem 1.26 that characterizes optimal transport maps in the Euclidean setting in terms of convex functions. Our characterization is presented in Proposition 4.5. We begin by studying the existence of optimal pairs.

Lemma 4.1 (Existence of Optimal pairs) *Let $\mu, \sigma \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ and suppose that $W_2^{\mathcal{G},W,\tau}(\mu,\sigma) < \infty$. Then, the set $Opt(\mu,\sigma)$ (i.e., the set of solutions to (2.20)) is non-empty.*

Proof Let us consider a minimizing sequence of admissible pairs $\{(\gamma_n, h_n)\}_{n=1}^{\infty}$ and note that since $\mathcal{A}^{\mathcal{G},W,\tau}(\mu,\sigma) < \infty$ we have that, passing to a subsequence if necessary, we can assume that the second moments of $\{\gamma_n\}_{n=1}^{\infty}$, and the norm of $\{h_n\}_{n=1}^{\infty}$ in the weighted space $L^2_W(\mathbb{R}^d \times \mathcal{G} \times \mathcal{G})$ are equibounded (see (2.18)). Consequently, since $L^2_W(\mathbb{R}^d \times \mathcal{G} \times \mathcal{G})$ is a Hilbert space, the existence of a minimizer follows by a standard lower compactness/lower semicontinuity and weak convergence argument (see Ambrosio and Gigli (2013), Theorem 1.2). Indeed, since the constraint (2.19) is linear, we can pass it to the limit by weak convergence of γ_n and h_n in duality with smooth functions with compact support. □

Notice that if $\mu = \sigma$ then $W_2^{\mathcal{G},W,\tau} = 0 < \infty$. The following lemma will not be used in the sequel, but provides other examples of μ and σ for which one can prove that $W_2^{\mathcal{G},W,\tau}(\mu,\sigma) < \infty$.

Lemma 4.2 *Let $\mu, \sigma \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ be absolutely continuous w.r.t.* $\mathrm{d}x\mathrm{d}g$ *and assume that σ's density belongs to the space:*

$$L^2_W(\mathbb{R}^d \times \mathcal{G}) := \left\{ f : \mathbb{R}^d \times \mathcal{G} \to \mathbb{R} \quad s.t. \quad \sum_{g \in \mathcal{G}} \int |f_g|^2 e^W \mathrm{d}x < \infty \right\}. \tag{4.1}$$

Then, $W_2^{\mathcal{G},W,\tau}(\mu,\sigma) < \infty$.

Proof We begin by showing that the cost $\mathcal{A}^{\mathcal{G},W,\tau}(\mu,\sigma)$ is finite. Let f and $\tilde{f}$ be the densities for μ and σ respectively, and define

$$m_g := \int_{\mathbb{R}^d} f_g(x)\mathrm{d}x, \quad g \in \mathcal{G},$$
$$\tilde{f}(x) := \sum_g \tilde{f}_g(x), \quad x \in \mathbb{R}^d.$$

Notice that for every $g \in \mathcal{G}$ the positive measures f_g and $m_g \tilde{f}$ have the same total mass, and thus there exists a coupling γ_g between them. In particular, $\pi_{1\sharp}\gamma = f_g$, $\pi_{2\sharp}\gamma = m_g \tilde{f}$ and also

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - \tilde{x}|^2 \mathrm{d}\gamma_g < \infty.$$

Now, notice that for every $x \in \mathbb{R}^d$ we have

$$\sum_g (m_g \tilde{f}(x) - \tilde{f}_g(x)) = 0.$$

Therefore, we may use Proposition 2.1 in order to find $\eta(x, \cdot)$ satisfying

$$m_g \tilde{f}(x) - \tilde{f}_g(x) = \sum_{g'} (\eta(x, g) - \eta(x, g'))K(g, g'), \quad \forall g \in \mathcal{G}, \tag{4.2}$$

as well as

$$\sum_{g,g'} |\eta(x, g) - \eta(x, g')|^2 K(g, g') \leq C \sum_g |m_g \tilde{f}(x) - \tilde{f}_g(x)|^2 e^{2W(x)}, \tag{4.3}$$

for some constant C that only depends on the weighted graph $(\mathcal{G}, K)$. We let

$$h_{gg'}(x) := \frac{e^{W(x)}}{\tau}(\eta(x, g) - \eta(x, g')), \quad x \in \mathbb{R}^d, \quad g, g' \in \mathcal{G}$$

and notice that from (4.2) it follows that

$$\sigma_g = \pi_{2\sharp}\gamma_g - \tau \sum_{g'} h_{gg'}(x)K(g, g')e^{-W(x)}.$$

We observe that h is clearly antisymmetric in $\mathcal{G} \times \mathcal{G}$, and thanks to (4.3) and the fact that $\tilde{f} \in L^2_W(\mathbb{R}^d \times \mathcal{G})$ also satisfies

$$\sum_{gg'} \int_{\mathbb{R}^d} h^2_{gg'} e^{-W} K(g, g')\mathrm{d}x < \infty.$$

The bottom line is that $(\gamma, h) \in ADM(\mu, \sigma)$ and $C_\tau^{W,K}(\gamma, h) < \infty$. It follows that $W_2^{\mathcal{G},W,\tau}(\mu, \sigma) < \infty$. □

Remark 4.3 To provide an example where the cost is infinite suppose that $\mathcal{G}$ consists of two elements g_1, g_2 and $K(g_1, g_2) > 0$. Let μ be the measure with representation $\mu_{g_1} = \delta_{x_1}$ for some $x_1 \in \mathbb{R}^d$ and $\mu_{g_2} = 0$ (i.e., all mass is in g_1), and let σ be the

measure with $\sigma_{g_1} = 0$ and $\sigma_{g_2} = \delta_{x_2}$ for some $x_2 \in \mathbb{R}^d$. We show that $ADM(\mu, \sigma) = \emptyset$. Indeed, if there existed an admissible pair, from (2.19) we would have that

$$\delta_{x_2} = \sigma_{g_2} = \pi_{2\sharp}\gamma_{g_2} - \tau h_{g_2 g_1}(x) K(g_1, g_2) e^{-W(x)} \mathrm{d}x = -\tau h_{g_2 g_1}(x) K(g_1, g_2) e^{-W(x)} \mathrm{d}x.$$

In other words, we would conclude that δ_{x_2} admits a density w.r.t. Lebesgue measure.

The main ingredient necessary to prove the main result of this section, i.e., Proposition 4.5, is a set of variational inequalities satisfied by optimal pairs. We obtain such inequalities by computing the first variation of minimizing pairs under suitable perturbations. We do this in the next lemma. Before stating this result let us first introduce some notation that will be used in the remainder of the section. We let μ, σ be as in Lemma 4.1 and assume that σ has a density. To a given minimizing pair (γ, h), we associate the density

$$\bar{f}_g(x) := \sigma_g(x) + \tau \sum_{g'} h_{gg'} K(g, g') e^{-W}, \tag{4.4}$$

which corresponds to the density of the measure $\pi_{2\sharp}\gamma_g$. An immediate observation is that each γ_g is an optimal plan for the OT problem between μ_g and $\pi_{2\sharp}\gamma_g$ for the cost $c(x, y) = \frac{|x-y|^2}{2\tau}$. Given that $\pi_{2\sharp}\gamma_g$ has a density, we know that there exists a unique map $S_g : \mathbb{R}^d \to \mathbb{R}^d$ such that $(S_g, \mathrm{Id})_{\#}\bar{f}_g = \gamma_g$ (see Ambrosio and Gigli 2013[Theorem 6.2.4 and Remark 6.2.11], for example). We will use the maps $\{S_g\}_{g\in\mathcal{G}}$ to state the variational inequalities satisfied by minimizers of the static semi-discrete transportation problem. This set of inequalities serves as analogue to the notion of cyclical monotonicity that appears in the classical (Euclidean) optimal transport setting.

Lemma 4.4 (Variational inequalities) *Let μ and σ satisfy the hypothesis of Lemma 4.1 and suppose that in addition σ has a density w.r.t. $\mathrm{d}x\mathrm{d}g$. Let (γ, h) be an element in $Opt(\mu, \sigma)$. Then, the following properties hold:*

- *For any g in $\mathcal{G}$ and any y in $\mathbb{R}^d$, suppose we have two sequences $\{g_l\}_{l=0}^{M}$ and $\{g'_l\}_{l'=0}^{M'}$ in $\mathcal{G}$, that satisfy both:*
  - *a) The two sequences describe paths in the graph with the same initial and final endpoints, i.e, we have that $g_0 = g'_0$, $g_M = g'_{M'}$, $K(g_l, g_{l+1}) > 0$, and $K(g'_l, g'_{l+1}) > 0$.*
  - *b) The point y is a Lebesgue point for all the functions $h_{g_{l-1}g_l}$ and $h_{g'_{l-1},g'_l}$.*

  Then,

$$\sum_{l=1}^{M} h_{g_{l-1}g_l}(y) = \sum_{l'=1}^{M'} h_{g'_{l-1}g'_l}(y). \tag{4.5}$$

- *Fix g and g' satisfying $K(g, g') > 0$ and assume that y is a Lebesgue point for S_g which also belongs to the support of $\pi_{2\sharp}\gamma_g$, and that y' is a Lebesgue point for*

$S_{g'}$ *which also belongs to the support of* $\pi_{2\sharp}\gamma_{g'}$. *Then,*

$$\begin{aligned}(h_{gg'}(y') - h_{gg'}(y)) + \left[\frac{|y' - S_g(y)|^2}{2\tau} - \frac{|y - S_g(y)|^2}{2\tau}\right] \\ + \left[\frac{|y - S_{g'}(y')|^2}{2\tau} - \frac{|y' - S_{g'}(y')|^2}{2\tau}\right] \geq 0.\end{aligned} \tag{4.6}$$

Proof Let us start with a small outline describing the main ideas behind the proof.

Heuristic Proof We begin analyzing (4.6). The idea is to perturb γ_g by transporting a small amount of mass from $(S_g(y), g)$ into (y', g) instead of transporting it to (y, g). On the other hand, $\gamma_{g'}$ is perturbed by transporting a small amount of mass from $(S(y'), g')$ into (y, g') instead of transporting it to (y', g'). By modifying the plans γ_g and $\gamma_{g'}$, we create a transport cost differential

$$\left[\frac{|y' - S_g(y)|^2}{2\tau} - \frac{|y - S_g(y)|^2}{2\tau}\right] + \left[\frac{|y - S_{g'}(y')|^2}{2\tau} - \frac{|y' - S_{g'}(y')|^2}{2\tau}\right], \tag{4.7}$$

per unit of mass transported. To balance the above perturbation in the transportation and remain with an admissible pair, we must also perturb $h_{y'}(gg')$ and $h_{y'}(g'g)$ so that the extra amount of mass created by the transportation perturbation gets removed from (y', g) and put into (y', g'). We must also perturb $h_y(g'g)$ and $h_y(gg')$ so that the extra amount of mass created by the transportation perturbation gets removed from (y, g') and put into (y, g). Modifying the mass exchange function h in this way creates a mass exchange cost differential of

$$h_{gg'}(y') - h_{gg'}(y),$$

per unit of mass transported. The resulting modified pair is still admissible, and by optimality of the original pair (γ, h), it must be the case that

$$\begin{aligned}(h_{gg'}(y') - h_{gg'}(y)) + \left[\frac{|y' - S_g(y)|^2}{2\tau} - \frac{|y - S_g(y)|^2}{2\tau}\right] \\ + \left[\frac{|y - S_{g'}(y')|^2}{2\tau} - \frac{|y' - S_{g'}(y')|^2}{2\tau}\right] \geq 0,\end{aligned}$$

which is precisely (4.6).

To deduce (4.5), we consider two sequences $\{g_l\}_{l=1}^{M}$ and $\{g'_l\}_{l=1}^{M'}$ satisfying the given conditions *a*) and *b*) for some y in $\mathbb{R}^d$. We send some extra mass from the point (y, g_0) to the point (y, g_1) by increasing $h_y(g_0g_1)$. Then, we take the extra mass at (y, g_1) and send it to (y, g_2) by increasing $h_y(g_1g_2)$. We can continue in this fashion until we reach the point $(y, g_M) = (y, g'_{M'})$. At this stage, we will have a deficit of mass at the

point (y, g_1) and an excess of mass at the point $(y, g'_{M'})$ and we will pay an excess exchange cost given by:

$$\sum_{l=1}^{M} h_{g_{l-1}g_l}(y),$$

per unit of mass transported. We can balance the previous perturbation by reversing the mass exchange along the sequence $\{g'_l\}_{l=1}^{M'}$. Namely, for each pair g'_l, g'_{l+1} we reduce the mass sent from (y, g'_l) to (y, g'_{l+1}) by decreasing $h_{g'_l g'_{l+1}}(y)$. Doing this we save

$$\sum_{l=1}^{M'} h_{g'_{l-1}g'_l}(y)$$

in terms of the cost. By optimality, we must have

$$\sum_{l=1}^{M} h_{g_{l-1}g_l}(y) \geq \sum_{l=1}^{M'} h_{g'_{l-1}g'_l}(y).$$

We can then switch the roles of the sequences and obtain the opposite inequality and from this deduce (4.5).

Let us now make the previous ideas rigorous.

Rigorous proof: 1. We begin with the proof of (4.5). Let us fix two positive real numbers $r, \varepsilon > 0$. We perturb our minimizer (γ, h) by considering a new mass exchange function:

$$h^{r,\varepsilon}_{g_{l-1}g_l}(\hat{y}) := \begin{cases} h_{g_{l-1}g_l}(\hat{y}) & \text{if } \hat{y} \in B_r^c(y) \\ h_{g_{l-1}g_l}(\hat{y}) + \frac{\varepsilon}{\tau K(g_{l-1}, g_l) e^{-W(\hat{y})}} & \text{if } \hat{y} \in B_r(y), \end{cases}$$

$$h^{r,\varepsilon}_{g'_{l-1}g'_l}(\hat{y}) := \begin{cases} h_{g'_{l-1}g'_l}(\hat{y}), & \text{if } \hat{y} \in B_r^c(y) \\ h_{g'_{l-1}g'_l}(\hat{y}) - \frac{\varepsilon}{\tau K(g'_{l-1}, g'_l) e^{-W(\hat{y})}} & \text{if } \hat{y} \in B_r(y), \end{cases}$$

$h^{r,\varepsilon}_{g_l g_{l-1}} = -h_{g_l g_{l-1}}$ and $h^{r,\varepsilon}_{g'_l g'_{l-1}} = -h_{g'_l g'_{l-1}}$ to maintain the asymmetry, and finally $h^{r,\varepsilon}_{gg'} = h_{gg'}$ whenever (g, g') is not one of the consecutive pairs in the sequences. In the above, we use $B_r(y)$ to denote the Euclidean ball of radius r centered at y.

It is straightforward to see that the pair $(\gamma, h^{r,\varepsilon})$ is admissible, and thus by the optimality of (γ, h) we have $C_\tau(\gamma, h) \leq C_\tau(\gamma, h^{r,\varepsilon})$, which simplifies to

$$0 \leq \frac{\tau}{2}\sum_{l=1}^{M}\int_{B_r(y)}\left(\left(h_{g_{l-1}g_l}+\frac{\varepsilon}{\tau K(g_{l-1},g_l)e^{-W(\hat{y})}}\right)^2-(h_{g_{l-1}g_l})^2\right)$$
$$K(g_{l-1},g_l)e^{-W(\hat{y})}d\hat{y}$$
$$+\frac{\tau}{2}\sum_{l=1}^{M'}\int_{B_r(y)}\left(\left(h_{g'_{l-1}g'_l}-\frac{\varepsilon}{\tau K(g'_{l-1},g'_l)e^{-W(\hat{y})}}\right)^2-(h_{g'_{l-1}g'_l})^2\right)$$
$$K(g'_{l-1},g'_l)e^{-W(\hat{y})}d\hat{y}.$$

Dividing by ε and letting $\varepsilon \to 0$ yields

$$0 \leq \int_{B_r(y)}\left(\sum_{l=1}^{M}h_{g_{l-1}g_l}(\hat{y})-\sum_{l=1}^{M'}h_{g'_{l-1}g'_l}(\hat{y})\right)d\hat{y}.$$

Dividing by the volume of $B_r(y)$, letting $r \to 0$, and recalling that y was assumed to be a Lebesgue point for all the functions $h_{g_{l-1}g_l}$ and $h_{g'_{l-1},g'_l}$ we conclude that

$$\sum_{l=1}^{M}h_{g_{l-1}g_l}(y) \geq \sum_{l=1}^{M'}h_{g'_{l-1}g'_l}(y).$$

Switching the roles of the sequences we obtain the reverse inequality. (4.5) follows.

2. Let us now consider (y_1, g_1) and (y_2, g_2) such that $K(g_1, g_2) > 0$, y_1 is a Lebesgue point of S_{g_1} and belongs to the support of $\pi_{2\sharp}\gamma_{g_1}$, y_2 is a Lebesgue point of S_{g_2} and belongs to the support of $\pi_{2\sharp}\gamma_{g_2}$, and $y_1 \neq y_2$. Fix $\varepsilon > 0$, and let r be a small enough positive number so that $B_r(y_1) \cap B_r(y_2) = \emptyset$. We now construct measures $\gamma_{g_1}^{r,\varepsilon}$, $\gamma_{g_2}^{r,\varepsilon}$ and a function $h_{g_1g_2}^{r,\varepsilon}$ which we use to formalize the perturbation argument provided in the heuristic proof. To define these measures and function, we first need to introduce some objects.

Let us start by defining

$$m_1 := \gamma_{g_1}(\mathbb{R}^d \times B_r(y_1)), \quad m_2 := \gamma_{g_2}(\mathbb{R}^d \times B_r(y_2)).$$

Notice that both numbers are nonzero given that y_1 belongs to the support of $\pi_{2\sharp}\gamma_{g_1}$ and y_2 belongs to the support of $\pi_{2\sharp}\gamma_{g_2}$. To ease the notation, we use $\bar{\mu}_{g_1}$ and $\bar{\mu}_{g_2}$ to denote the positive measures

$$\bar{\mu}_{g_1} := \pi_{2\sharp}\gamma_{g_1} = \bar{f}_{g_1}\mathrm{d}x, \quad \bar{\mu}_{g_2} := \pi_{2\sharp}\gamma_{g_2} = \bar{f}_{g_2}\mathrm{d}x,$$

and consider also the positive measures $\bar{\mu}_{g_1}|_{B_r(y_1)}$ and $\bar{\mu}_{g_2}|_{B_r(y_2)}$ defined by

$$\bar{\mu}_{g_1}|_{B_r(y_1)}(A) := \bar{\mu}_{g_1}(A \cap B_r(y_1)), \quad \bar{\mu}_{g_2}|_{B_r(y_2)}(A) := \bar{\mu}_{g_2}(A \cap B_r(y_2)),$$

for all Borel subsets A of $\mathbb{R}^d$.

Let us consider the maps $\mathcal{T}_{y_1}^{y_2}(y) := (y - y_1 + y_2)$ and $\mathcal{T}_{y_2}^{y_1}(y) := (y - y_2 + y_1)$. Also, let $T_1 : B_r(y_1) \to B_r(y_1)$ be an optimal transport map (for the quadratic cost) between the measures $\mathcal{T}_{y_2}^{y_1}\sharp(\frac{m_1}{m_2}\bar{\mu}_{g_2}|_{B_r(y_2)})$ and the measure $\bar{\mu}_{g_1}|_{B_r(y_1)}$ (measures that can be checked to have the same total mass), and let $T_2 : B_r(y_2) \to B_r(y_2)$ be an optimal transport map between the measures $\mathcal{T}_{y_1}^{y_2}\sharp(\bar{\mu}_{g_1}|_{B_r(y_1)})$ and the measure $\frac{m_1}{m_2}\bar{\mu}_{g_2}|_{B_r(y_2)}$.

We can now define the measures $\gamma_{g_1}^{r,\varepsilon}$ and $\gamma_{g_1}^{r,\varepsilon}$ by

$$\begin{aligned}\gamma_{g_1}^{r,\varepsilon}(A \times C) := \gamma_{g_1}(A \times C) - \varepsilon\gamma_{g_1}(A \times (C \cap B_r(y_1))) \\ + \varepsilon(S_{g_1}, T_2 \circ \mathcal{T}_{y_1}^{y_2})_\sharp\bar{\mu}_{g_1}|_{B_r(y_1)}(A \times C),\end{aligned}$$

and

$$\begin{aligned}\gamma_{g_2}^{r,\varepsilon}(A \times C) := \gamma_{g_2}(A \times C) - \varepsilon\frac{m_1}{m_2}\gamma_{g_2}(A \times (C \cap B_r(y_2))) \\ + \varepsilon(S_{g_2}, T_1 \circ \mathcal{T}_{y_2}^{y_1})_\sharp(\frac{m_1}{m_2}\bar{\mu}_{g_2}|_{B_r(y_2)})(A \times C),\end{aligned}$$

for all A, C Borel subsets of $\mathbb{R}^d$. For g that is neither g_1 nor g_2 we set $\gamma_g^{r,\varepsilon} = \gamma_g$. Notice that $\pi_{1\sharp}\gamma_{g_1}^{r,\varepsilon} = \mu_{g_1}$ and $\pi_{2\sharp}\gamma_{g_2}^{r,\varepsilon} = \mu_{g_2}$.

Finally, we define

$$h_{g_1g_2}^{r,\varepsilon}(y) := h_{g_1g_2}(y) + \frac{\varepsilon}{\tau K(g_1, g_2)e^{-W(y)}}\left(\frac{m_1}{m_2}\overline{f}_{g_2}(y)\mathbb{1}_{B_r(y_2)}(y) - \overline{f}_{g_1}(y)\mathbb{1}_{B_r(y_1)}(y)\right)$$

and set $h_{g_2g_1}^{r,\varepsilon} = -h_{g_1g_2}^{r,\varepsilon}$, and $h_{gg'}^{r,\varepsilon} = h_{gg'}$ for pairs g, g' different from g_1g_2. It is straightforward to check that $h^{r,\varepsilon} \in L_2^{W,K}(\mathbb{R}^d \times \mathcal{G} \times \mathcal{G})$ and that for every $g \in \mathcal{G}$

$$\sigma_g = \pi_{2\sharp}\gamma_g^{r,\varepsilon} - \tau\sum_{g'} h_{gg'}K(g, g')e^{-W}.$$

That is, $(\gamma^{r,\varepsilon}, h^{r,\varepsilon}) \in ADM(\mu, \sigma)$ and thus by optimality of (γ, h) we deduce that $C_\tau(\gamma, h) \le C_\tau(\gamma^{r,\varepsilon}, h^{r,\varepsilon})$. This inequality simplifies to

$$\begin{aligned}\varepsilon\int_{B_r(y_1)}\left[\frac{|Id - S_{g_1}|^2}{2\tau} - \frac{|T_2\circ\mathcal{T}_{y_1}^{y_2} - S_{g_1}|^2}{2\tau}\right]\bar{f}_{g_1}\mathrm{d}y \\ \le \varepsilon\frac{m_1}{m_2}\int_{B_r(y_2)}\left[\frac{|T_1\circ\mathcal{T}_{y_2}^{y_1} - S_{g_2}|^2}{2\tau} - \frac{|Id - S_{g_2}|^2}{2\tau}\right]\bar{f}_{g_2}\mathrm{d}y \\ + \frac{\tau}{2}\int_{B_r(y_1)}\left[(h_{g_1g_2} - \frac{\varepsilon}{\tau K(g_1, g_2)e^{-W}}\bar{f}_{g_1})^2 - h_{g_1g_2}^2\right]K(g_1, g_2)e^{-W}\mathrm{d}y \\ + \frac{\tau}{2}\int_{B_r(y_2)}\left[(h_{g_1g_2} + \frac{\varepsilon}{\tau K(g_1, g_2)e^{-W}}\frac{m_1}{m_2}\bar{f}_{g_2})^2 - h_{g_1g_2}^2\right]K(g_1, g_2)e^{-W}\mathrm{d}y.\end{aligned}$$

If we divide by ε and let $\varepsilon \to 0$, we obtain

$$\int_{B_r(y_1)} \left[\frac{|Id - S_{g_1}|^2}{2\tau} - \frac{|T_2 \circ \mathcal{T}_{y_1}^{y_2} - S_{g_1}|^2}{2\tau} \right] \bar{f}_{g_1} \mathrm{d}y \\ \leq \frac{m_1}{m_2} \int_{B_r(y_2)} \left[\frac{|T_1 \circ \mathcal{T}_{y_2}^{y_1} - S_{g_2}|^2}{2\tau} - \frac{|Id - S_{g_2}|^2}{2\tau} \right] \bar{f}_{g_2} \mathrm{d}y \\ - \int_{B_r(y_1)} h_{g_1 g_2} \bar{f}_{g_1} \mathrm{d}y + \frac{m_1}{m_2} \int_{B_r(y_2)} h_{g_1 g_2} \bar{f}_{g_2} \mathrm{d}y.$$

Consequently, dividing by m_1 and expanding we obtain

$$\frac{1}{m_1} \int_{B_r(y_1)} \left[\frac{|Id - S_{g_1}|^2}{2\tau} - \frac{|\mathcal{T}_{y_1}^{y_2} - S_{g_1}|^2}{2\tau} + R_2(y) \right] \bar{f}_{g_1} \mathrm{d}y \\ \leq \frac{1}{m_2} \int_{B_r(y_2)} \left[\frac{|\mathcal{T}_{y_2}^{y_1} - S_{g_2}|^2}{2\tau} - \frac{|Id - S_{g_2}|^2}{2\tau} + R_1(y) \right] \bar{f}_{g_2} \mathrm{d}y \\ - \frac{1}{m_1} \int_{B_r(y_1)} h_{g_1 g_2} \bar{f}_{g_1} \mathrm{d}y + \frac{1}{m_2} \int_{B_r(y_2)} h_{g_1 g_2} \bar{f}_{g_2} \mathrm{d}y, \tag{4.8}$$

where

$$\frac{1}{m_1} \int_{B_r(y_1)} |R_2(y)| \bar{f}_{g_1} \mathrm{d}y = \frac{1}{m_1} \int_{B_r(y_1)} \left| \frac{|\mathcal{T}_{y_1}^{y_2} - S_{g_1}|^2}{2\tau} - \frac{|T_2 \circ \mathcal{T}_{y_1}^{y_2} - S_{g_1}|^2}{2\tau} \right| \bar{f}_{g_1} \mathrm{d}y \\ = \frac{1}{2\tau m_1} \int_{B_r(y_1)} |\langle \mathcal{T}_{y_1}^{y_2} - T_2 \circ \mathcal{T}_{y_1}^{y_2}, \mathcal{T}_{y_1}^{y_2} + T_2 \circ \mathcal{T}_{y_1}^{y_2} - 2S_{g_1} \rangle| \bar{f}_{g_1} \mathrm{d}y \\ \leq \frac{r}{\tau} \frac{1}{m_1} \int_{B_r(y_1)} |\mathcal{T}_{y_1}^{y_2} + T_2 \circ \mathcal{T}_{y_1}^{y_2} - 2S_{g_1}| \bar{f}_{g_1} \mathrm{d}y,$$

and by a similar computation

$$\frac{1}{m_2} \int_{B_r(y_2)} |R_1(y)| \bar{f}_{g_2} \mathrm{d}y \leq \frac{r}{\tau} \frac{1}{m_2} \int_{B_r(y_2)} |\mathcal{T}_{y_2}^{y_1} + T_1 \circ \mathcal{T}_{y_2}^{y_1} - 2S_{g_2}| \bar{f}_{g_2} \mathrm{d}y.$$

We now use the above estimates and let $r \downarrow 0$ in (4.8) to deduce (4.6) (with $(y, g) = (y_1, g_1)$ and $(y', g') = (y_2, g_2)$). □

Before proceeding with our characterization of optimal pairs, let us first recall some useful definitions from the classical optimal transport theory. First, given a symmetric $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we say that a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is *c-concave*, if it can be written as

$$\varphi(y) = \inf_{x \in \mathbb{R}^d} c(x, y) - \psi(x), \quad \forall y \in \mathbb{R}^d,$$

for some $\psi \quad : \quad \mathbb{R}^d \to \mathbb{R}$. The *c-transform* of a given φ is the function φ^c defined by

$$\varphi^c(x) := \inf_{y \in \mathbb{R}^d} c(x, y) - \varphi(y), \tag{4.9}$$

and its *c-superdifferential* is the set

$$\partial^c_+ \varphi := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \quad : \quad \varphi^c(x) + \varphi(y) = c(x, y) \right\}. \tag{4.10}$$

To characterize minimizers of Problem 2.4, in the proposition below we will use the quadratic cost

$$c(x, y) := \frac{1}{2\tau} |x - y|^2.$$

We will also use the spaces Φ and $\Phi^\perp$ defined in (3.2).

Proposition 4.5 (Characterization of Optimal pairs) *Let μ, σ be absolutely continuous with respect to $\mathrm{d}x\mathrm{d}g$ and assume that $W_2^{\mathcal{G},W,\tau}(\mu, \sigma) < \infty$. Also, let (γ, h) be in $ADM(\mu, \sigma)$ and assume that μ_g's density and $\bar{f}_g$ as defined in (4.4) are strictly positive for every g in $\mathcal{G}$. Then, the following are equivalent*

i. $C_\tau(\gamma, h)$ is minimal among all pairs in $ADM(\mu, \sigma)$.
ii. There exist functions $\phi, \psi \quad : \quad \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ satisfying the following properties:

a) For every g in $\mathcal{G}$, the plan γ_g is supported on $\partial^c_+ \phi_g$, for some c-concave function $\phi_g(\cdot) = \phi(\cdot, g)$.
b) For Lebesgue almost every point $y \in \mathbb{R}^d$, the function $\psi_y(\cdot) = \psi(y, \cdot)$ satisfies

$$\psi_y(g') - \psi_y(g) = h_{gg'}(y), \quad \forall g, g' \text{ with } K(g, g') > 0. \tag{4.11}$$

c) The difference $\phi - \psi$ belongs to $\Phi^\perp$ as defined in (3.2).

iii. We can find a single potential $\varphi \quad : \quad \mathbb{R}^d \times \mathcal{G} \to \mathbb{R}$ satisfying properties a), b), and c) from item ii.

Proof **1.** Optimality of (γ, h) implies that γ_g is an optimal coupling between μ_g and $\pi_{2\sharp}\gamma_g$ for every g, and thus the proof that $i. \implies ii.a)$ follows directly from the classical (Euclidean) optimal transport theory (see Ambrosio and Gigli 2013, Theorem 1.13). To prove that $i. \implies ii.b)$, let us fix y_0 in $\mathbb{R}^d$ and g_0 in $\mathcal{G}$ and define

$$\psi(y_0, g) := \sum_{l=1}^{M} h_{g_{l-1} g_l}(y_0),$$

for some sequence $\{g_l\}_{l=0}^M$ starting at g_0, with $K(g_l, g_{l+1}) > 0$, and for which $g_M = g$. Such sequence exists given that $(\mathcal{G}, K)$ was assumed to be connected. On the other

hand, observe that by (4.5) the potential ψ is well defined (i.e., does not depend on the actual sequence connecting g_0 and g). In particular, we also have

$$\psi(y_0, g') = \sum_{l=1}^{M} h_{g_{l-1}g_l}(y_0) + h_{gg'}(y_0).$$

$ii.b)$ now follows.

We proceed to show that $i. \implies ii.c)$. According to Remark 3.3, it suffices to show that the difference $\psi - \phi$ is orthogonal to any ε of the form (3.8)

$$\varepsilon = \xi^r_{y',g} - \xi^r_{y',g'} - \xi^r_{y,g} + \xi^r_{y,g'},$$

for arbitrary y, y', g, g' and $r > 0$. To show this, we proceed as follows.

Fix g, g' with $K(g, g') > 0$. We first claim that the function

$$u_{gg'}(y) := \psi_{g'}(y) - \psi_g(y) + \phi_g(y) - \phi_{g'}(y).$$

is a.e. constant, where ψ is as in item $ii.b)$. To see this, notice that from Brenier's theorem for the classical optimal transport problem with the (rescaled) quadratic cost the following holds: the functions $\phi_g, \phi_{g'}$ can be written as

$$\phi_g(y) = -\beta_g(y) + \frac{|y|^2}{2\tau}, \quad \phi_{g'}(y) = -\beta_{g'}(y) + \frac{|y|^2}{2\tau},$$

for convex functions β_g and $\beta_{g'}$, and the maps S_g and $S_{g'}$ are a.e. equal to $\tau\nabla_y\beta_g$ and $\tau\nabla_y\beta_{g'}$ respectively. In particular, we can write

$$u_{gg'}(y) = \psi_{g'}(y) - \psi_g(y) - \beta_g(y) + \beta_{g'}(y), \quad y \in \mathbb{R}^d.$$

Now, for a given pair $y, y' \in \mathbb{R}^d$, we have $u_{gg'}(y) \geq u_{gg'}(y')$ or $u_{gg'}(y') \geq u_{gg'}(y)$. Suppose for the moment that the first inequality holds. In that case,

$$\beta_g(y) - \beta_{g'}(y) - \beta_g(y') + \beta_{g'}(y') \leq \psi_{g'}(y) - \psi_g(y) + \psi_g(y') - \psi_{g'}(y'). \quad (4.12)$$

After simplification, item $ii.a)$ and (4.6) imply

$$\psi_{g'}(y) - \psi_g(y) + \psi_g(y') - \psi_{g'}(y') \leq -\langle y' - y, \nabla_y\beta_g(y)\rangle - \langle y - y', \nabla_y\beta_{g'}(y')\rangle, \quad (4.13)$$

for a.e. y, y'. Combining (4.12) and (4.13), and recalling the definition of $u_{gg'}$ we obtain

$$\begin{aligned} |u_{gg'}(y) - u_{gg'}(y')| \leq\ & \beta_g(y') - (\beta_g(y) + \langle y' - y, \nabla_y\beta_g(y)\rangle) \\ & + \beta_{g'}(y) - (\beta_{g'}(y') + \langle y - y', \nabla_y\beta_{g'}(y')\rangle). \end{aligned} \quad (4.14)$$

Notice that if instead $u_{gg'}(y') \geq u_{gg'}(y)$ we would have obtained the same inequality as the one above changing the roles of g and g' on the right hand side, so we do not lose generality in assuming the former inequality. Given that along every straight line ℓ the functions β_g, $\beta_{g'}$ are convex, their distributional second derivatives (along ℓ) are characterized in terms of Radon positive measures, implying that along almost every line ℓ in $\mathbb{R}^d$ the right hand side in (4.14) is $O(|y-y'|)$, and in particular $u_{gg'}$ is a locally Lipschitz function along ℓ. Furthermore, along almost every line in ℓ and for almost every y, y' on that line, the right hand side of (4.14) is $o(|y-y'|)$ (given that Radon measures can only have at most a countable number of point masses). This implies that the locally Lipschitz function $u_{gg'}$ (restricted to ℓ) has derivative a.e. equal to zero, thus implying that the function is constant along almost every line ℓ. From this it follows that $u_{gg'}$ is almost everywhere constant in $\mathbb{R}^d$. The bottom line is that for almost every y, $y' \in \mathbb{R}^d$ we have

$$\big(\psi_{g'}(y') - \psi_g(y') - \psi_{g'}(y) + \psi_g(y)\big) - \big(\phi_{g'}(y') - \phi_g(y') - \phi_{g'}(y) + \phi_g(y)\big) = 0.$$

From the above it now follows that

$$\int_{\mathbb{R}^d} \sum_{\tilde{g}} (\psi(y,\tilde{g}) - \phi(y,\tilde{g}))\varepsilon(y,\tilde{g}) \quad \mathrm{d}y = 0,$$

for ε as in (3.8). This concludes the proof.

2. We now show that *ii. implies iii.* By Lemma (3.2) we can find $\varphi_1 \quad : \quad \mathbb{R}^d \to \mathbb{R}$ in $L^2_{loc}(\mathbb{R}^d)$ and $\varphi_2 \quad : \quad \mathcal{G} \to \mathbb{R}$ such that

$$\phi_g(y) - \psi_y(g) = \varphi_1(y) + \varphi_2(g).$$

Let us define

$$\varphi(y,g) := \phi_g(y) - \varphi_2(g) = \psi_y(g) + \varphi_1(y).$$

Clearly, we have that

$$\varphi(y,g') - \varphi(y,g) = \psi_y(g') - \psi_y(g).$$

Thus $ii.b)$, follows. On the other hand, since ϕ_g is c-concave, $\phi_g(\cdot) - \varphi_2(g)$ is c-concave too. Also, it is straightforward to verify that the superdifferential of $\phi_g(y)$ and $\phi_g(y) - \varphi_2(g)$ agree. In particular, $ii.a)$ holds for the potential φ.

3. To prove that $iii. \implies i.$, let $(\tilde{\gamma}, \tilde{h})$ be any element of $ADM(\mu, \sigma)$. Then, using item $ii.a)$, (2.19), (4.9), and (4.10), we have that

$$
\begin{aligned}
C_\tau(\gamma, h) &= \sum_{g\in\mathcal{G}}\Big[\int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)\mathrm{d}\gamma_g + \frac{\tau}{4}\sum_{g'\in\mathcal{G}}\Big(\int h^2_{gg'}(y)K(g',g)e^{-W}\mathrm{d}y\Big)\Big] \\
&= \sum_{g\in\mathcal{G}}\Big[\int_{\mathbb{R}^d\times\mathbb{R}^d} (\varphi^c_g(x)+\varphi_g(y))\mathrm{d}\gamma_g + \frac{\tau}{4}\sum_{g'\in\mathcal{G}}\Big(\int h^2_{gg'}(y)K(g',g)e^{-W}\mathrm{d}y\Big)\Big] \\
&= \sum_{g\in\mathcal{G}}\Big[\int_{\mathbb{R}^d}\varphi^c_g\mathrm{d}\mu_g + \int_{\mathbb{R}^d}\varphi_g\mathrm{d}\sigma_g + \tau\sum_{g'\in\mathcal{G}}\int\Big(\varphi_g(y)(h_{gg'}(y))K(g',g)e^{-W}\Big)\mathrm{d}y \\
&\quad + \frac{\tau}{4}\sum_{g'\in\mathcal{G}}\Big(\int h^2_{gg'}(y)K(g',g)e^{-W}\mathrm{d}y\Big)\Big] \\
&= \sum_{g\in\mathcal{G}}\Big[\int_{\mathbb{R}^d\times\mathbb{R}^d} (\varphi^c_g(x)+\varphi_g(y))\mathrm{d}\tilde{\gamma}_g + \tau\sum_{g'\in\mathcal{G}}\Big(\int\varphi_g(y)\big(h_{gg'}(y)-\tilde{h}_{gg'}(y)\big) \\
&\quad K(g',g)e^{-W}\mathrm{d}y\Big) \\
&\quad + \frac{\tau}{4}\sum_{g'\in\mathcal{G}}\Big(\int h^2_{gg'}(y)K(g',g)e^{-W}\mathrm{d}y\Big)\Big] \\
&\le \sum_{g\in\mathcal{G}}\Big[\int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)\mathrm{d}\tilde{\gamma}_g + \sum_{g'\in\mathcal{G}}\frac{\tau}{2}\Big(\int\big(\varphi_g(y)-\varphi_{g'}(y)\big)\big([h_{gg'}(y)-\tilde{h}_{gg'}(y)\big] \\
&\quad K(g',g)e^{-W}\mathrm{d}y\Big) \\
&\quad + \frac{\tau}{4}\sum_{g'\in\mathcal{G}}\Big(\int h^2_{gg'}(y)K(g',g)e^{-W}\mathrm{d}y\Big)\Big],
\end{aligned}
$$

where in the last line we have used the antisymmetry of h and $\tilde{h}$. Now, from item $ii.b)$ and the above inequality we obtain

$$
\begin{aligned}
C_\tau(\gamma, h) &\le \sum_g\int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)\mathrm{d}\tilde{\gamma}_g + \frac{\tau}{4}\sum_{g,g'}\Big(\int\tilde{h}^2_{gg'}(y)K(g',g)e^{-W}\mathrm{d}y\Big) \\
&\quad + \frac{\tau}{4}\sum_{g,g'}\Big(\int(h^2_{gg'}(y)-\tilde{h}^2_{gg'}(y))K(g',g)e^{-W}\mathrm{d}y\Big) \\
&\quad + \sum_{g,g'}\frac{\tau}{2}\Big(\int\big(h_{gg'}(y)\big)\big(\tilde{h}_{gg'}(y)-h_{gg'}(y)\big)K(g',g)e^{-W}\mathrm{d}y\Big) \\
&\le C_\tau(\tilde{\gamma}, \tilde{h}).
\end{aligned}
$$

□

5 Properties of JKO Minimizers and Maximum Principle

In this section, we prove a series of preliminary results characterizing solutions to the optimization problem (2.23). In Proposition 5.3, we show that the iterates of the minimizing movement scheme satisfy a maximum principle that is characteristic of the Fokker–Plank equation. In Proposition 5.6, we show that the corresponding potential φ generating the associated optimal transport map and optimal exchange function from Proposition 4.5 agrees with (2.12), i.e., with the formula for the gradient of $\mathcal{E}$ suggested by the formal computation from Sect. 2.2.

We begin by showing that minimizers of (2.23) exist.

Lemma 5.1 (Existence of minimizers to (2.23)) *Let μ be a measure in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ with the property that $\mathcal{E}(\mu) < \infty$. Then, there exists a minimizer $\mu_\tau \in \mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ of*

$$\sigma \to \mathcal{E}(\sigma) + \mathcal{A}^{W,\mathcal{G},\tau}(\mu, \sigma). \tag{5.1}$$

Moreover, such a minimizer is absolutely continuous with respect to the measure $\mathrm{d}x\mathrm{d}g$.

Proof Since the entropy of μ is finite, by considering the competitor $\sigma = \mu$ we deduce that the infimum in (5.1) if finite as well. Now, consider a minimizing sequence of measures $\{\sigma^n\}_{n=1}^\infty$, with corresponding optimal pairs $\{(\gamma^n, h^n)\}_{n=1}^\infty$ in $ADM(\mu, \sigma^n)$. Then, by construction, the second moments of $\{\gamma^n\}_{n=1}^\infty$ and the norm of $\{h^n\}_{n=1}^\infty$ in the weighted space $L^2_W(\mathbb{R}^d \times \mathcal{G} \times \mathcal{G})$ are equibounded. Thus, following the argument of Lemma 4.1 we can guarantee the existence of a pair (γ, h) such that up to subsequence not relabeled, γ^n converges weakly to γ, h^n converges weakly (in L^2_W) to h and

$$\liminf_{n\to\infty} C_\tau(\gamma^n, h^n) \geq C_\tau(\gamma, h).$$

From

$$\sigma_g^n = \pi_{2\#}\gamma_g^n - \tau \sum_{g'} h^n_{gg'}(x) K(g, g') e^{-W},$$

and the weak convergence of the sequences $\{(\gamma^n, h^n)\}_{n=1}^\infty$, we deduce that

$$\begin{aligned}\mu_\tau &:= \lim_{n\to\infty} \pi_{2\#}\gamma_g^n - \tau \sum_{g'} h^n_{gg'}(x) K(g, g') e^{-W} \\ &= \pi_{2\#}\gamma_g - \tau \sum_{K(g',g)>0} h_{gg'}(x) K(g, g') e^{-W}.\end{aligned}$$

Consequently, the pair (γ, h) belongs to $\mathrm{AMD}(\mu, \mu_\tau)$. Finally, the inequality

$$\liminf_{n\to\infty} \mathcal{E}(\sigma^n) \geq \mathcal{E}(\mu_\tau),$$

is a consequence of the weak convergence of σ^n toward μ_τ and the weak lower semi continuity of the relative entropy. The desired result follows. □

In the next lemma, we prove a set of variational inequalities satisfied by minimizers of (2.23). These inequalities are the main ingredient necessary to attain the main results of this section, i.e., Propositions 5.3 and 5.6. We obtain these inequalities by computing the first variation of minimizing pairs under suitable perturbations.

Proposition 5.2 (Variational inequalities of JKO minimizers) *Let μ and μ_τ be as in Lemma 5.1, and let f_τ be μ_τ's density. Let $\{\gamma_g\}_g$ and h be the optimal transport plans and optimal exchange functions for the static semi-discrete optimal transport between μ and $\sigma = \mu_\tau$. The following inequalities hold:*

- *Let $y \in \mathbb{R}^d$ be a Lebesgue point for the function $h_{g_1 g_2}$ where $K(g_1, g_2) > 0$, and suppose that (y, g_2) is an element in the support of f_τ. Then,*

$$\log f_\tau(y, g_1) + V(y, g_1) - [\log f_\tau(y, g_2) + V(y, g_2)] \geq h_{g_1 g_2}(y). \quad (5.2)$$

- *Let y_1 be a Lebesgue point for S_g and suppose that $(y_1, g), (y_2, g)$ belong to the support of f_τ. Then,*

$$\log f_\tau(y_2, g) + V(y_2, g) - [\log f_\tau(y_1, g) + V(y_1, g)] + \frac{|S_g(y_1) - y_2|^2}{2\tau} \geq \frac{|S_g(y_1) - y_1|^2}{2\tau}. \quad (5.3)$$

- *Let (x, y) be an element in the support of γ_g for some g in $\mathcal{G}$, and suppose that x and y belong to the support of $f_{\tau,g}$. Then,*

$$\log f_\tau(x, g) + V(x, g) - [\log f_\tau(y, g) + V(y, g)] \geq \frac{|x - y|^2}{2\tau}. \quad (5.4)$$

Proof Let us start with a small outline describing the main ideas behind the proof.

Heuristic Proof We begin by proving (5.2). For this purpose, we consider the following perturbation of the optimal pair (γ, h). The idea is to stop exchanging a small amount of mass between (y, g_1) and (y, g_2). By doing this we save

$$h_{g_1 g_2}(y) + \log f_\tau(y, g_2) + 1 + V(y, g_2),$$

in terms of the mass exchange cost and the entropy, and we pay an extra

$$\log f_\tau(y, g_1) + 1 + V(y, g_1),$$

in terms of the entropy of the excess mass we now have in (y, g_1). Thus, (5.2) follows by optimality.

We proceed to the proof of (5.3). We perturb γ_g as follows. Instead of transporting a small amount of the mass from $(S_g(y_1), g)$ into (y_1, g), we transport it to (y_2, g). By doing this, we create a transport cost differential

$$\frac{|y_2 - S_g(y_1)|^2}{2\tau} - \frac{|y_1 - S_g(y_1)|^2}{2\tau}.$$

The resulting excess mass in (y_2, g) and deficit of mass in (y_1, g) create an entropy differential of

$$\log f_\tau(y_2, g) + V(y_2, g) - [\log f_\tau(y_1, g) + V(y_1, g)].$$

Thus, (5.3) follows by optimality.

Finally, to prove (5.4) we take a pair (x, y) in the support of γ_g where both x, y are assumed to belong to the support of $f_{\tau,g}$. Now, by setting $y = y_1$ and $x = S(y_1) = y_2$ in inequality (5.3) we have

$$\log f_\tau(x, g) + V(x, g) - [\log f_\tau(y, g) + V(y, g)] \geq \frac{|y - x|^2}{2\tau}.$$

Rigorous proof We only prove (5.2). The proof of (5.3) follows exactly as in Proposition 3.7 from Figalli and Gigli (2010) and the proof of (5.4) follows the same lines as Lemma 4.4.

Let $y \in \mathbb{R}^d$ be a Lebesgue point for the function $h_{g_1g_2}$ and suppose that (y, g_2) is an element in the support of f_τ. Let r and ε be positive numbers. We perturb the minimizing pair (γ, h) by considering the new mass exchange function $h^{r,\varepsilon}_{g_1g_2} : \mathbb{R}^d \to \mathbb{R}$ defined by

$$h^{r,\varepsilon}_{g_1g_2}(\hat{y}) = \begin{cases} h_{g_1g_2}(\hat{y}), & \text{if } \hat{y} \in B_r^c(y) \\ h_{g_1g_2}(\hat{y}) - \frac{\varepsilon f_{\tau,g_2}(\hat{y})}{\tau K(g_1,g_2) e^{-W(\hat{y})}} & \text{if } \hat{y} \in B_r(y), \end{cases}$$

$h^{r,\varepsilon}_{g_2g_1} := -h^{r,\varepsilon}_{g_2g_1}$ and $h^{r,\varepsilon}_{gg'} = h_{gg'}$ whenever (g, g') is not (g_1, g_2) or (g_2, g_1). Observe that this produces a competitor $\mu^{r,\varepsilon}_\tau$ whose densities are given by

$$f^{r,\varepsilon}_{\tau,g_1}(\hat{y}) = \begin{cases} f_{\tau,g_1}(\hat{y}), & \text{if } \hat{y} \in B_r(y)^c \\ f_{\tau,g_1}(\hat{y}) + \varepsilon f_{\tau,g_2}(\hat{y}) & \text{if } \hat{y} \in B_r(y) \end{cases},$$

$$f^{r,\varepsilon}_{\tau,g_2}(\hat{y}) = \begin{cases} f_{\tau,g_2}(\hat{y}), & \text{if } \hat{y} \in B_r(y)^c \\ (1-\varepsilon) f_{\tau,g_2}(\hat{y}) & \text{if } \hat{y} \in B_r(y), \end{cases}$$

and $f_{\tau,g} = f^{r,\varepsilon}_\tau$ whenever g is not g_1 or g_2. From the minimality of μ_τ we get that

$$\sum_g \int_{\mathbb{R}^d} \vartheta(f_\tau, \hat{y}, g) \mathrm{d}\hat{y} + \mathcal{C}_\tau(\gamma, h) \leq \sum_g \int \vartheta(f^{r,\varepsilon}_\tau, \hat{y}, g) d\hat{y} + \mathcal{C}_\tau(\gamma, h^{r,\varepsilon}),$$

which simplifies to

$$\int_{B_r(y)} \left[\vartheta\left(f_{\tau,g_1}(\hat{y}), \hat{y}, g_1\right) + \vartheta\left(f_{\tau,g_2}(\hat{y}), \hat{y}, g_2\right) + \frac{\tau}{2} h^2_{g_1 g_2}(\hat{y}) K(g_1, g_2) e^{-W(\hat{y})} \right] d\hat{y}$$
$$\leq \int_{B_r(y)} \left[\vartheta\left(f_{\tau,g_1}(\hat{y}) + \varepsilon f_{\tau,g_2}(\hat{y}), \hat{y}, g_1\right) + \vartheta\left((1-\varepsilon) f_{\tau,g_2}(\hat{y}), \hat{y}, g_2\right) \right.$$
$$\left. + \frac{\tau}{2} \left(h_{g_1 g_2}(\hat{y}) - \frac{\varepsilon f_{\tau,g_2}(\hat{y})}{\tau K(g_1, g_2) e^{-W(\hat{y})}} \right)^2 K(g_1, g_2) e^{-W(\hat{y})} \right] d\hat{y}.$$

Reordering terms, we obtain

$$\int_{B_r(y)} \left[\vartheta\left(f_{\tau,g_1}(\hat{y}), \hat{y}, g_1\right) - \vartheta\left(f_{\tau,g_1}(\hat{y}) + \varepsilon f_{\tau,g_2}(\hat{y}), \hat{y}, g_1\right) \right] d\hat{y}$$
$$\leq \int_{B_r(y)} \left[\vartheta\left((1-\varepsilon) f_{g_2}(\hat{y}), \hat{y}, g_2\right) - \vartheta\left(f_{\tau,g_2}(\hat{y}), \hat{y}, g_2\right) \right.$$
$$\left. + \frac{\tau}{2} \left[\left(h_{g_1 g_2}(\hat{y}) - \frac{\varepsilon f_{\tau,g_2}(\hat{y})}{\tau K(g_1, g_2) e^{-W(\hat{y})}} \right)^2 - h^2_{g_1 g_2}(\hat{y}) \right] K(g_1, g_2) e^{-W(\hat{y})} \right] d\hat{y}.$$

Dividing by ε and letting $\varepsilon \to 0$ yields

$$\int_{B_r(y)} \left[-\log f_{\tau,g_1}(\hat{y}) - 1 - V(\hat{y}, g_1) \right] f_{\tau,g_2}(y) \mathrm{d}y$$
$$\leq \int_{B_r(y)} \left[-\log f_{\tau,g_2}(\hat{y}) - 1 - V(\hat{y}, g_2) - h_{g_1 g_2}(\hat{y}) \right] f_{\tau,g_2}(\hat{y}) d\hat{y}.$$

Dividing by $\int_{B_r(y)} f_{\tau,g_2}(\hat{y}) d\hat{y}$, and letting $r \to 0$ we obtain the desired inequality. □

In the next proposition, we prove that minimizers of (2.23) satisfy a maximum principle that is characteristic of Fokker–Planck equations.

Proposition 5.3 (Consistent barriers) *Suppose that μ and μ_τ are as in Lemma 5.1. Suppose in addition that μ's density satisfies:*

$$\lambda e^{-V(x,g)} \leq f(x, g) \leq \Lambda e^{-V(x,g)},$$

for every (x, g). Then, f_τ satisfies

$$\lambda e^{-V(x,g)} \leq f_\tau(x, g) \leq \Lambda e^{-V(x,g)}, \tag{5.5}$$

as well.

Proof We only prove the lower bound in (5.5) since the argument for the upper bound is completely analogous. Let us define the set

$$A := \{(x, g) \ : \ \lambda e^{-V(x,g)} > f_\tau(x, g)\},$$

and consider the auxiliary positive measure

$$\mathrm{d}\mu_\lambda = \lambda e^{-V(x,g)}\mathrm{d}x\mathrm{d}g.$$

Suppose for the sake of contradiction that

$$\mu_\lambda(A) > 0.$$

Then,

$$\mu(A) \geq \mu_\lambda(A) > \mu_\tau(A),$$

and thus the set A has to lose mass during the transportation. Consequently, at least one of the following facts should hold:

i. There exist $g \in \mathcal{G}$ and y a Lebesgue point of S_g such that $(S_g(y), g) \in A$ and $(y, g) \notin A$.
ii. There exist a pair of nodes g, g' with $K(g, g') > 0$ and x a density point of $h_{gg'}$ for which (x, g) and (x, g') belong to the support of f_τ, $h_{gg'}(x) > 0$, $(x, g) \in A$ and $(x, g') \notin A$.

Let us show that in both cases we reach a contradiction.

Case *i:* In this case, we apply (5.3) with $y_1 = y$ and $y_2 = S_g(y)$ to obtain that

$$\log f_\tau(y, g) + V(y, g) + \frac{1}{2\tau}|S_g(y) - y|^2 \leq \log f_\tau(S_g(y), g) + V(S_g(y), g).$$

Now, observe that the assumption that $(y, g) \notin A$ implies that the left-hand side of the above inequality is bigger than $\log\lambda$, whereas the assumption that $(S_g(y), g) \in A$ implies the right-hand side is strictly smaller than $\log\lambda$. Thus, we reach a contradiction.

Case *ii:* In this case we apply (5.2) with $g_2 = g'$, $g_1 = g$ and $y = x$, to obtain that

$$0 < h_{gg'}(x) \leq \log f_\tau(x, g) + V(x, g) - \log f_\tau(x, g') - V(x, g').$$

Moreover, our assumption that $(x, g') \notin A$ and $(x, g) \in A$ implies that the right hand side is negative. Thus, we reach a contradiction. □

As a by-product of the above proposition, we obtain a uniform control on the distance traveled by the transported mass.

Lemma 5.4 (Transportation bound) *Let μ, μ_τ, λ, and Λ be as in Proposition 5.3. Then, there exists $C > 0$ such that for all $g \in \mathcal{G}$*

$$|y - x| \leq C\sqrt{\tau} \quad \forall (x, y) \in supp(\gamma_g),$$

where we recall $\gamma = \{\gamma_g\}_{g\in\mathcal{G}}$ is the set of optimal plans between μ and μ_τ. The constant C can be taken to be $C = \sqrt{2}(\log(\Lambda) - \log(\lambda))$.

Proof The estimate follows by combining (5.4) with Proposition 5.3. □

In the next lemma, we show that the target density f_τ and the transported density

$$\bar{f}_\tau(x, g) = f_\tau(x, g) + \tau \sum_{g'} h_{gg'}(x) K(g, g') e^{-W(x)},$$

are comparable. Recall that $\bar{f}_{\tau,g}$ is nothing but the density of the positive measure $\pi_{2\sharp}\gamma_g$.

Lemma 5.5 (Positivity of the transported mass) *Let μ, μ_τ, λ, and Λ be as in Proposition 5.3, and let λ', Λ' be as in (2.24). Finally, let $\bar{f}_\tau$ be defined as above. Then, there exists a positive constant $\tau_0 := \tau_0(\lambda, \Lambda, \lambda', \Lambda') < 1/2$ such that for any τ in $(0, \tau_0)$ we have that $\bar{f}_\tau > 0$, i.e., the support of $\pi_{2\#}\gamma_g$ is all of $\mathbb{R}^d$ for all $g \in \mathcal{G}$. Moreover, we have that*

$$\frac{C}{1-\tau} < \frac{\bar{f}_{\tau,g}}{f_{\tau,g}} < C(1+\tau), \tag{5.6}$$

for any τ in $(0, \tau_0)$ for some constant C that only depends on λ, Λ, λ', Λ'.

Proof To prove (5.6), we note that thanks to (5.2) and (5.5), we have that the mass exchange function h is uniformly bounded in terms of λ and Λ. Additionally, (5.5) and the assumption (2.24) imply that the quotient of e^{-W} and f_τ is uniformly bounded as well. Hence, the desired result follows. □

In the next proposition, we show that the potential φ that generates the optimal transport map and exchange function between μ and μ_τ for μ satisfying the conditions from Proposition 5.3 (see item iii. in Proposition 4.5) agrees with the negative of (2.12) which is the gradient of the relative entropy suggested by the formal Riemannian structure from Sect. 2.2.

Proposition 5.6 (The gradient of the relative entropy and JKO minimizers) *Let μ, μ_τ, λ, and Λ be as in Proposition 5.3, let λ', Λ' be as in (2.24), and let $\tau_0 > 0$ be as in Lemma 5.5. Then, for every τ in $(0, \tau_0)$ we have:*

i. For each g in $\mathcal{G}$ the optimal transport plan $\gamma_{\tau,g}$ is given by

$$\gamma_{\tau,g} = (S_g, Id)_\#\big(f_{\tau,g} + \tau \sum h_{\tau,gg'} K(g, g') e^{-W}\big), \tag{5.7}$$

where the corresponding optimal transport map S_g satisfies

$$\frac{S_g(y) - y}{\tau} f_\tau(y, g) = \nabla_x f_\tau(y, g) + f_\tau(y, g) \nabla_x V(y, g), \tag{5.8}$$

for almost every y in $\mathbb{R}^d$.

ii For each pair g, g' with $K(g, g') > 0$ and for almost every x in $\mathbb{R}^d$, the optimal exchange function $h_{\tau,gg'}$ satisfies

$$h_{\tau,gg'}(x) = \big[\log f_\tau(x, g) + V(x, g) - \log f_\tau(x, g') - V(x, g')\big]. \tag{5.9}$$

Proof We begin by noting that thanks to Lemma 5.5 and Proposition 5.3, we have that the support of $f_{\tau,g}$ and $\pi_{2\#}\gamma_g$ is $\mathbb{R}^d$ for any g in $\mathcal{G}$, i.e., $f_\tau > 0$ and $\bar{f}_\tau > 0$. We will use this fact together with the variational inequalities from Proposition 5.2.

1. Let us begin by showing i.

Observe that due to (5.3) for any (x, y) in the support of γ_g we have

$$\log f_{\tau,g}(z) + V(z, g) - \log f_{\tau,g}(y) - V(y, g) + \frac{|x - z|^2}{2\tau} \geq \frac{|x - y|^2}{2\tau},$$

for almost every z in $\mathbb{R}^d$. Expanding the squares and rearranging terms, we obtain that

$$\log f_{\tau,g}(z) + V(z, g) + \frac{|z|^2}{2} \geq \log f_{\tau,g}(y) + V(y, g) + \frac{|y|^2}{2} + \langle \frac{x}{\tau}, z - y\rangle$$
$$\text{foralmost every} z \text{ in} \mathbb{R}^d.$$

Such an inequality implies that, up to redefining $f_{\tau,g}$ in a set up measure zero, the function $\Phi_g(z) = \log f_{\tau,g}(z) + V(z, g) + \frac{|z|^2}{2}$ is convex and for almost every y in $\mathbb{R}$ and every pair (x, y) in the support of the optimal transport plan $\gamma_{\tau,g}$ we have that $\frac{x}{\tau}$ is contained in the subdifferential of Φ_g at y. Following the notation from Ambrosio et al. (2005), Section 3.1 , we shall denote such a subdifferential by $\partial^-\Phi(y)$. Finally, since convex function are almost everywhere differentiable, we have that for almost every y the set $\partial^-\Phi_g(y)$ is a singleton and

$$\nabla_{z=y}\Phi_g = \frac{x}{\tau}.$$

Moreover, using the almost everywhere differentiability of Φ_g we get that $z \to \log f_{\tau,g}(z) + V(z, g)$ is almost everywhere differentiable and

$$\nabla_{z=y}\left(\log f_{\tau,g}(z) + V(z, g) + \frac{|z|^2}{2}\right) = \frac{x}{\tau}$$

which implies that

$$\tau\nabla_y(\log f_{\tau,g} + V_g) = x - y.$$

Notice that combining the above equation with Lemma 5.4, we obtain that $\log f_{\tau,g} + V_g$ has a uniformly bounded gradient. Consequently, i follows.

2. Let us now show ii. Using (5.2), we obtain

$$\log f_\tau(x, g') + V(x, g') - [\log f_\tau(x, g) + V(x, g)] \geq h_{g'g}(x),$$

for almost every x in $\mathbb{R}^d$. Interchanging g and g', we obtain the opposite inequality and thus the desired identity. Here, once more we have used the fact that Proposition 5.3 and Lemma 5.4 imply that $f_\tau > 0$ and $\bar{f}_\tau > 0$. □

As a direct consequence of the above proposition, we obtain the following result:

Corollary 5.7 (Sobolev regularity) *Let μ, μ_τ, λ, and Λ be as in Proposition 5.3, let λ', Λ' be as in (2.24), and let $\tau_0 > 0$ be as in Lemma 5.5. Then, for every τ in $(0, \tau_0)$, $f_{\tau,g}$ is contained in the weighted Sobolev space $W^{1,2}(\mathbb{R}^d, e^W)$ for every g in $\mathcal{G}$. Moreover,*

$$\sum_{g\in\mathcal{G}} \int_{\mathbb{R}^d} |f_\tau(x,g)|^2 e^W \mathrm{d}x \le C_1 \sum_{g\in\mathcal{G}} \int_{\mathbb{R}^d} e^{-W(x)} \mathrm{d}x, \tag{5.10}$$

$$\tau \sum_{g\in\mathcal{G}} \int_{\mathbb{R}^d} |\nabla_x f_\tau(x,g)|^2 e^W \mathrm{d}x \le C_2\big[\mathcal{E}(\mu) - \mathcal{E}(\mu_\tau) + \tau\big], \tag{5.11}$$

for some constant C_1 that only depends on λ, Λ, λ', Λ', and a constant C_2 that only depends on λ, Λ, λ', Λ' and the quantity

$$[\nabla_x V]_{e^{-V}} := \sum_g \int_{\mathbb{R}^d} |\nabla_x V(y,g)|^2 e^{-V(y,g)} \mathrm{d}y.$$

Proof The fact that $f_{\tau,g}$ belongs to $L^2(\mathbb{R}^d, e^W)$ follows from (5.3), (2.24), and the fact that e^{-W} was assumed to be integrable.

Now, note that by optimality

$$\mathcal{E}(\mu_\tau) + C_\tau(\mu, \mu_\tau) \le \mathcal{E}(\mu).$$

Consequently, using (5.8) and the definition of the transportation cost, we deduce that

$$\frac{\tau}{2} \sum_{g\in\mathcal{G}} \int |\nabla_x \log f_\tau(y,g) + \nabla_x V(y,g)|^2 \bar{f}_\tau(y,g) \mathrm{d}y \le \mathcal{E}(\mu) - \mathcal{E}(\mu_\tau).$$

Hence, using (5.5), (5.6) and (2.24) we obtain

$$\tau \sum_{g\in\mathcal{G}} \int |\nabla_x f_\tau(y,g)|^2 e^{-W} \mathrm{d}y \le C\Big(\mathcal{E}(\mu) - \mathcal{E}(\mu_\tau) + \tau\Big),$$

for some constant C that only depends on λ, Λ, λ', Λ' and the quantity $[\nabla_x V]_{e^{-V}}$. □

6 Convergence of the JKO Scheme: Proof of Theorem 2.14

Let us start by defining precisely the notion of weak solution to Eq. (2.14).

Definition 6.1 We say that a weakly continuous curve of measures $\{\mu_t\}_{t\geq 0}$ in $\mathcal{P}_2(\mathbb{R}^d \times \mathcal{G})$ with associated probability density functions $\{f(t,\cdot,\cdot)\}_{t\geq 0}$ is a weak solution with initial condition f_0 (2.14) if

$$f(0,x,g) = f_0(x,g), \quad \forall (x,g) \in \mathbb{R}^d \times \mathcal{G}$$

and

$$\begin{aligned}
&\sum_g \left(\int_{\mathbb{R}^d} \zeta_g f_g(s,x)\mathrm{d}x - \int_{\mathbb{R}^d} \zeta_g f_g(r,x)\mathrm{d}x \right) \\
&= \int_r^s \left(\sum_g \int_{\mathbb{R}^d} \left[\Delta_x \zeta_g - \langle \nabla_x V_g, \nabla_x \zeta_g \rangle \right] f_g(t,x)\mathrm{d}x \right. \\
&\quad + \frac{1}{2} \sum_{g,g'} \int_{\mathbb{R}^d} [\zeta_{g'} - \zeta_g][\log f_{g'}(t,x) + V_{g'} \\
&\quad \left. - \log f_g(t,x) - V_g] K(g,g') e^{-W(x)} \mathrm{d}x \right) \mathrm{d}t,
\end{aligned}$$

for every r, s, in $[0,\infty)$, and every test function ζ in $C_c^\infty(\mathbb{R}^d \times \mathcal{G})$.

With all the preliminary results from Sect. 5, we can now proceed to the proof of Theorem 2.14.

Proof of Theorem 2.14 **1. JKO scheme produces an approximate solution.** Let f_0 be an initial datum with finite energy $\mathcal{E}(f_0) < \infty$ satisfying (2.22). Let τ_0 be as in Lemma 5.5, Proposition 5.6, and Corollary 5.7. Let $\tau \in (0, \tau_0)$, and for every $n \in \mathbb{N}$ let $(\gamma_n^\tau, h_n^\tau)$ be the minimizing pair of transporting f_n^τ into f_{n+1}^τ, where the f_n^τ are the densities iteratively constructed as in (2.23). Let $S_{n,g}^\tau$ be the optimal transport map associated to $\gamma_{n,g}^\tau$ as in (5.7), and let $\bar{f}_n^\tau$ be the density of the measure $\pi_{2\sharp}\gamma_{n,g}^\tau$, i.e., the transported density. We recall that $\bar{f}_{n,g}^\tau$ can be written as

$$\bar{f}_{n,g}^\tau = f_{n+1,g}^\tau + \tau \sum_{g'} h_{n,gg'}^\tau K(g,g') e^{-W}.$$

Notice that by iterating Proposition 5.3, we have

$$\lambda e^{-V_g} \leq f_{n,g}^\tau \leq \Lambda e^{-V_g} \quad \forall n \in \mathbb{N},$$

and by Lemma (5.5)

$$\frac{C}{1-\tau} < \frac{\bar{f}_{n,g}^\tau}{f_{n+1,g}^\tau} < C(1+\tau).$$

Finally, recall that the discrete time sequence f_n^τ can be extended to continuous time by setting

$$f^\tau(t) := f_{n+1}^\tau \quad \text{for} \quad t \in \big(n\tau, (n+1)\tau\big],$$

We will now show that the curve $t \mapsto f^\tau(t)$ can be interpreted as an approximate solution to Eq. (2.14).

Let $\zeta \in C_c^\infty(\mathbb{R}^d \times \mathcal{G})$ be an arbitrary test function. Then,

$$\int_{\mathbb{R}^d} \zeta_g \quad f_{n+1,g}^\tau(y)\mathrm{d}y - \int_{\mathbb{R}^d} \zeta_g f_{n,g}^\tau(x)\mathrm{d}x = \int \zeta_g(y)\mathrm{d}\gamma_{n,g}^\tau(x,y) - \int \zeta_g(x)\mathrm{d}\gamma_{n,g}^\tau(x,y) + \tau \sum_{g' \in \mathcal{G}} \int_{\mathbb{R}^d} \zeta_g h_{n,gg'}^\tau K(g',g) e^{-W}\mathrm{d}y. \tag{6.1}$$

Using the fundamental theorem of calculus and (5.8), we deduce

$$\begin{aligned}
&\int_{\mathbb{R}^d\times\mathbb{R}^d} \zeta_g(y)\mathrm{d}\gamma_{n,g}^\tau(x,y) - \int_{\mathbb{R}^d\times\mathbb{R}^d} \zeta_g(x)\mathrm{d}\gamma_{n,g}^\tau(x,y) \\
&\quad = \int_{\mathbb{R}^d\times\mathbb{R}^d} \big(\zeta_g(y) - \zeta_g(x)\big)\mathrm{d}\gamma_{n,g}^\tau(x,y) \\
&\quad = \int_{\mathbb{R}^d\times\mathbb{R}^d} \big(\zeta_g(y) - \zeta_g(S_{n,g}^\tau(y))\big) \quad \bar{f}_{n,g}^\tau(y)\mathrm{d}y \\
&\quad = \int_{\mathbb{R}^d\times\mathbb{R}^d} (\zeta_g(y) - \zeta_g(S_{n,g}^\tau(y)) f_{n+1,g}^\tau(y)\mathrm{d}y + R_1(\tau,n,g) \\
&\quad = -\int_{\mathbb{R}^d} \langle \nabla_x \zeta_g, S_{n,g}^\tau - Id\rangle f_{n+1,g}^\tau(y)\mathrm{d}y + R_2(\tau,n,g) + R_1(\tau,n,g) \\
&\quad = -\tau \int_{\mathbb{R}^d} \langle \nabla_x \zeta_g, \nabla_x f_{n+1,g}^\tau + f_{n+1,g}^\tau \nabla_x V_g\rangle \mathrm{d}y + R(\tau,n,g),
\end{aligned}$$

where the error term is given by

$$\begin{aligned}
R(\tau,n,g) &= R_1(\tau,n,g) + R_2(\tau,n,g) \\
&= \tau \int_{\mathbb{R}^d} (\zeta_g - \zeta_g \circ S_{n,g}^\tau) \sum_{g'} h_{n,gg'}^\tau K(g,g') e^{-W}\mathrm{d}y \\
&\quad + \int_{\mathbb{R}^d} \int_0^1 \Big(\langle \nabla_x \zeta_g \circ ((1-s)S_{n,g}^\tau + sId), Id - S_{n,g}^\tau\rangle - \langle \nabla_x \zeta_g, Id - S_{n,g}^\tau\rangle \Big) \\
&\quad f_{n+1,g}^\tau(y) ds \mathrm{d}y.
\end{aligned}$$

Plugging back in (6.1) and using (5.9), we deduce that

$$\begin{aligned}\sum_g \int_{\mathbb{R}^d} \zeta_g \quad f^\tau_{n+1,g}(y)\mathrm{d}y - \sum_g \int_{\mathbb{R}^d} \zeta_g \quad f^\tau_{n,g}(x)\\ \mathrm{d}x = -\tau \sum_g \int_{\mathbb{R}^d} \langle \nabla_x \zeta_g, \nabla_x f^\tau_{n+1,g} + f^\tau_{n+1,g} \nabla_x V_g\rangle \mathrm{d}y\\ + \frac{\tau}{2} \sum_{g,g'} \int_{\mathbb{R}^d} (\zeta_g - \zeta_{g'}) \big[\log f^\tau_{n+1}(x,g) + V(x,g)\\ - \log f^\tau_{n+1}(x,g') - V(x,g')\big] K(g,g') e^{-W} \mathrm{d}y\\ + \sum_g R(\tau,n,g).\end{aligned} \tag{6.2}$$

Let us now estimate the error terms. First, using (5.9) and the bounds on $f^\tau_{n+1,g}$ we can bound the transfer functions $h^\tau_{n,gg'}$ by a constant that only depends on λ and Λ, and then use Lemma 5.4 to obtain

$$|R(\tau,n,g)| \le C_1 \|\nabla_x \zeta_g\|_{L^\infty(\mathbb{R}^d)} \left(\tau^{\frac{3}{2}} + \int_{\mathbb{R}^d} |Id - S^\tau_{n,g}|^2 f^\tau_{n+1,g}(y)\mathrm{d}y\right), \tag{6.3}$$

for some constant $C_1 := C_1(\lambda, \Lambda)$. Now, from the fact that $f^\tau_{n+1,g}$ and $\bar{f}^\tau_{n,g}$ are comparable, and from the definition of $f^\tau_{n+1,g}$ and the transport cost $W_2^{\mathcal{G},W,\tau}$ it follows that

$$\sum_g \int |Id - S^\tau_{g,n}|^2 f^\tau_{n+1,g}\mathrm{d}y \le C_2 \sum_g \int |Id - S^\tau_{g,n}|^2 \bar{f}^\tau_{n,g}\mathrm{d}y \le C_2 \tau \big(\mathcal{E}(f^\tau_n) - \mathcal{E}(f^\tau_{n+1})\big).$$

where $C_2 := C_2(\lambda, \Lambda, \lambda', \Lambda')$. Thus, combining the above inequalities with (6.3) we deduce that

$$\begin{aligned}\sum_{n=M}^{N-1} \sum_g \big|R(\tau,n,g)\big| &\le C_3 \max_g \|\nabla_x \zeta_g\|_{L^\infty(\mathbb{R}^d)} \left(\tau^{3/2}(N-M) + \tau\left[\mathcal{E}(f^\tau_M) - \mathcal{E}(f^\tau_N)\right]\right)\\ &\le C_3 \max_g \|\nabla_x \zeta_g\|_{L^\infty(\mathbb{R}^d)} \left(\tau^{3/2}(N-M) + \tau \mathcal{E}(f_0)\right),\end{aligned} \tag{6.4}$$

for all $M \le N-1$, where $C_3 := C_3(\lambda, \Lambda, \lambda', \Lambda')$.

Let us now fix $0 \le r < s$. We add up (6.2) from $M = \lceil r \backslash \tau \rceil$ to $N - 1 = \lceil s \backslash \tau \rceil - 1$ (assuming that τ is small enough so that $M \le N - 1$) to get that

$$\begin{aligned}
&\sum_g \int_{\mathbb{R}^d} \zeta_g f_g^\tau(s,x) \quad \mathrm{d}x - \sum_g \int_{\mathbb{R}^d} \zeta_g f_g^\tau(r,x) \quad \mathrm{d}x \\
&= \int_{\tau\lceil r\backslash\tau\rceil}^{\tau\lceil s\backslash\tau\rceil} \Bigg(- \sum_g \int_{\mathbb{R}^d} \langle \nabla_x \zeta_g, \nabla_x f_g^\tau(t,x) + f_g^\tau(t,x) \nabla_x V_g \rangle \mathrm{d}x \\
&\quad + \frac{1}{2} \int_{\mathbb{R}^d} \sum_{g,g'} (\zeta_{g'} - \zeta_g) \big[\log f_{g'}^\tau(t,x) + V_{g'} - \log f_g^\tau(t,x) - V_g \big] K(g,g') e^{-W} \mathrm{d}x \Bigg) \mathrm{d}t \\
&\quad + \sum_{n=M}^{N-1} \sum_g R(\tau,n,g) \\
&= \int_{\tau\lceil r\backslash\tau\rceil}^{\tau\lceil s\backslash\tau\rceil} \Bigg(\sum_g \int_{\mathbb{R}^d} \big[\Delta_x \zeta_g - \langle \nabla_x \zeta_g, \nabla_x V_g \rangle \big] f_g^\tau(t,x) \quad \mathrm{d}x \\
&\quad + \frac{1}{2} \int_{\mathbb{R}^d} \sum_{g,g'} (\zeta_{g'} - \zeta_g) \big[\log f_{g'}^\tau(t,x) + V_{g'} - \log f_g^\tau(t,x) - V_g \big] K(g,g') e^{-W} \mathrm{d}x \Bigg) \mathrm{d}t \\
&\quad + \sum_{n=M}^{N-1} \sum_g R(\tau,n,g).
\end{aligned} \tag{6.5}$$

From (6.4), it is clear that as $\tau \to 0$ the error term in the above expression vanishes. Therefore, if we can show that as $\tau \to 0$ (along a sequence) the curve $t \mapsto f^\tau(t)$ converges to a limiting curve $t \mapsto f(t)$ which is weakly continuous, and that this convergence is strong enough so that in particular we can pass to the limit in all the terms in the above expression, then we will have shown that the curve $t \mapsto f(t)$ is indeed a weak solution to (2.14).

2. Compactness. Let us consider a sequence $\{\tau_k\}_k$ of positive numbers converging to zero. Without the loss of generality, we can assume that $\tau_k \le \tau_0$ for all k. Our goal is to show that we can pass to the limit in (6.5). For this purpose, we use the Aubin–Lions theorem (see Theorem 5 in Simon 1986). We introduce some notation first.

Let us fix $t_F > 0$. For $h > 0$, we define the *translates*

$$T_h f^{\tau_k}(t) := f^{\tau_k}(t+h).$$

Also, for $R > 0$ we let $U_R := B_R \times \mathcal{G}$, where B_R is the open ball in $\mathbb{R}^d$ with radius R centered at the origin. Let p be a positive number such that $p > d + 1$. Consider the Sobolev spaces $W^{1,2}(B_R)$ and $W^{2,p}(B_R)$, and denote by $W^{-2,p}(U_R)$ the dual of $W^{2,p}(B_R)$. Notice that

$$W^{1,2}(B_R) \hookrightarrow L^2(B_R) \hookrightarrow W^{-2,p}(B_R),$$

where the first embedding is compact and the second one is continuous; notice also that $W^{2,p}(B_R)$ embeds continuously into $C^1(B_R)$.

We show the following:

a) For every $g \in \mathcal{G}$, $\{f_g^{\tau_k}\}_k$ is bounded in $L^2(0, t_F; W^{1,2}(B_R))$.
b) For every $g \in \mathcal{G}$, $\|T_h f_g^{\tau_k} - f_g^{\tau_k}\|_{L^2(0,t_F-h;W^{-2,p}(B_R))} \to 0$ as $h \to 0$, uniformly for all k.

Theorem 5 in Simon (1986) will then imply that for every $g \in \mathcal{G}$, $\{f_g^{\tau_k}\}_k$ is precompact in $L^2(0, t_F; L^2(B_R))$.

2a. Observe that by iterating the bounds from Corollary 5.7 along $f_n^{\tau_k}$ we deduce that

$$\int_{B_R} |f_g^{\tau_k}(t,x)|^2 \quad \mathrm{d}x \le C_4, \quad \forall t \ge 0, \forall k \in \mathbb{N} \tag{6.6}$$

as well as

$$\int_0^{t_F} \left(\int_{B_R} |\nabla_x f_g^{\tau_k}(t,x)|^2 \quad \mathrm{d}x \right) \mathrm{d}t \le C_4(\mathcal{E}(f_0) + t_F), \quad \forall k \in \mathbb{N}, \tag{6.7}$$

where the constant C_4 depends only on $\lambda, \Lambda, \lambda', \Lambda', R, W, |\mathcal{G}|$. From the above inequalities it follows that for every $g \in \mathcal{G}$, the sequence $\{f_g^{\tau_k}\}_{k\in\mathbb{N}}$ is bounded in $L^2(0, t_F; W^{1,2}(B_R))$ (and also in $L^2(0, t_F; L^2(B_R))$). Moreover, for every $t \ge 0$ the sequence $\{f_g^{\tau_k}(t)\}_{k\in\mathbb{N}}$ is bounded in $L^2(B_R)$.

2b. Let h be smaller than t_F. For $t \in [0, t_F - h)$ set $N_k = \lceil \frac{t+h}{\tau_k} \rceil - 1$ and $M_k = \lceil \frac{t}{\tau_k} \rceil$. Notice that if $N_k < M_k$ then $T_h f_g^{\tau_k}(t) = f_g^{\tau_k}(t)$, and so we may assume that $M_k \le N_k$. For any given $\zeta_g \in W^{2,p}(B_R)$, we have

$$\begin{aligned}
&\int_{B_R} \zeta_g(x)(T_h f_g^{\tau_k}(t,x) - f_g^{\tau_k}(t,x)) \quad \mathrm{d}x \\
&= \sum_{n=M_k}^{N_k} \int_{B_R} \zeta_g \quad f_{n+1,g}^{\tau_k}(x)\mathrm{d}x - \int_{B_R} \zeta_g \quad f_{n,g}^{\tau_k}(x)\mathrm{d}x \\
&= \sum_{n=M_k}^{N_k} \int_{B_R\times B_R} (\zeta_g(y) - \zeta_g(x)) \quad \mathrm{d}\gamma_{n,g}^{\tau_k}(x,y) - \tau_k \sum_{g'} \int_{B_R} \zeta_g h_{n,gg'}^{\tau_k} e^{-W} \mathrm{d}x \\
&= \sum_{n=M_k}^{N_k} \int_{B_R\times B_R} \int_0^1 \langle \nabla\zeta_g(x + s(y-x)), y - x\rangle \quad ds \quad \mathrm{d}\gamma_{n,g}^{\tau_k} \\
&\quad - \tau_k \sum_{g'} \int_{B_R} \zeta_g h_{n,gg'}^{\tau_k} e^{-W} \quad \mathrm{d}x \\
&\le C_6 \sum_{n=M_k}^{N_k} \|\zeta_g\|_{C^1(B_R)} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} |y-x|^2 \quad \mathrm{d}\gamma_{n,g}^{\tau_k} \right)^{\frac{1}{2}} + C_5\tau_k \|\zeta_g\|_{W^{2,p}(B_R)} \\
&\le C_7 \|\zeta\|_{W^{2,p}(B_R)} \sum_{n=M_k}^{N_k} \left[\left(\int_{\mathbb{R}^d\times\mathbb{R}^d} |y-x|^2 \quad \mathrm{d}\gamma_{n,g}^{\tau_k} \right)^{\frac{1}{2}} + \tau_k \right].
\end{aligned}$$

In the above, the constant C_5 depends only on λ, Λ and W, C_6 depends only on R, and $C_7 := C_5 + C_6$. We have used the fact that $W^{2,p}(B_R)$ embeds continuously into $C^1(U_R)$, and we have also used the bounds on the exchange function $h_n^{\tau_k}$ from (5.9) together with the lower and upper bounds for the density $f_{n,g}^{\tau_k}$. Consequently,

$$
\begin{aligned}
&||T_h f_g^{\tau_k}(t) - f_g^{\tau_k}(t)||_{W^{-2,p}(B_R)} \\
&\quad = \sup_{||\zeta_g||_{W^{2,p}(B_R)}=1} \int_{B_R} \zeta_g \big(T_h f_g^{\tau_k}(t,y) - f_g^{\tau_k}(t,y)\big) \ \mathrm{d}y \\
&\quad \le C_7 \Bigg(\tau_k (N_k - M_k) + \big(\tau_k (N_k - M_k)\big)^{\frac{1}{2}} \Bigg(\sum_{n=M_k}^{N_k} \Bigg[\Bigg(\int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{|y-x|^2}{\tau_k} \ \mathrm{d}\gamma_{n,g}^{\tau_k} \Bigg) \Bigg)^{\frac{1}{2}} \\
&\quad \le C_7 \Bigg(h + \sqrt{h} \Bigg[\sum_{n=M_k}^{N_k} \mathcal{E}(f_n^{\tau_k}) - \mathcal{E}(f_{n+1}^{\tau_k}) \Bigg]^{\frac{1}{2}} \Bigg) \\
&\quad \le C_7 \Bigg(h + \sqrt{h} \Bigg[\mathcal{E}(f_{M_k}^{\tau_k}) - \mathcal{E}(f_{N_k+1}^{\tau_k}) \Bigg]^{1/2} \Bigg) \\
&\quad \le C_8 \Bigg(h + \sqrt{h} \Bigg[\mathcal{E}(f_0) \Bigg]^{1/2} \Bigg)
\end{aligned}
\tag{6.8}
$$

Here, we used Jensen's inequality, and the definition of $f_{n,g}^{\tau_k}$. This shows

$$
||T_h f_g^{\tau_k} - f_g^{\tau_k}||_{L^2(0,t_F-h;W^{-2,p}(B_R))} \to 0, \quad \text{as } h \to 0,
$$

uniformly in k.

From **2a)** and **2b)**, it now follows that for every $g \in \mathcal{G}$, the sequence $\{f_g^{\tau_k}\}_{k\in\mathbb{N}}$ is precompact in $L^2(0,t_F;L^2(B_R))$ (Theorem 5 in Simon (1986)). In particular, there exist a subsequence of $\{\tau_k\}_k$ (which we do not relabel) and an element $f_g \in L^2(0,t_F;L^2(B_R))$ such that $f_g^{\tau_k} \to f_g$ as $k \to \infty$ in $L^2(0,t_F;L^2(B_R))$. On the other hand, from (6.6) and (6.7) it follows that for almost every $t \in [0,t_F]$ the sequence $\{f_g^{\tau^k}(t)\}_k$ is bounded in $W^{1,2}(B_R)$ and thus precompact in $L^2(B_R)$ and in $W^{-2,p}(B_R)$. We can then use this fact and (6.8) to conclude from Arzela–Ascoli theorem that $\{f_g^{\tau_k}\}_k$ converges in $C(0,t_F;W^{-2,p}(B_R))$ (in fact in $C^{1/2-\varepsilon}$ for any ε) to f_g. Moreover, a standard diagonal argument sending $R \to \infty$ along a sequence allows us to assume without the loss of generality, that for every $g \in \mathcal{G}$, $f_g^{\tau_k} \to f_g$ in $L^2(0,t_F;L^2_{loc}(\mathbb{R}^d))$, as well as $f_g^{\tau_k} \to f_g$ in $C(0,t_F;W^{-2,p}_{loc}(\mathbb{R}^d))$, as $k \to \infty$.

3. Properties of $t \in [0,t_F) \mapsto f(t)$. We claim that for every $t \in [0,t_F)$ we have

$$
\lambda e^{-V_g} \le f_g(t) \le \Lambda e^{-V_g}.
$$

Indeed, notice that from (6.6) it follows that for every $t \in [0,t_F)$, the sequence $\{f_g^{\tau_k}(t)\}_{k\in\mathbb{N}}$ is bounded in $L^2(B_R)$ (for every R) and thus it must have a weakly converging subsequence in $L^2(B_R)$. Due to the fact that $f_g^{\tau_k} \to f_g$ in $C(0,t_F;W^{-2,p}_{loc}(\mathbb{R}^d))$, said subsequence must converge weakly to $f_g(t)$ in $L^2(B_R)$. Since each of the $f_g^{\tau_k}(t)$

satisfies the desired lower and upper bounds in B_R, it follows that $f_g(t)$ satisfies the same bounds in B_R. Since R was arbitrary we conclude that $f_g(t)$ satisfies the desired bounds in the whole $\mathbb{R}^d$.

Now we claim that for every $t \in [0, t_F)$

$$\sum_g \int_{\mathbb{R}^d} f_g(t,x)\mathrm{d}x = 1.$$

Indeed, this is a direct consequence of the lower and upper bounds obtained above and the fact that for every $t \in [0, t_F)$ $f_g^{\tau_k}(t)$ converges in $W^{-2,p}_{loc}(\mathbb{R}^d)$ toward $f_g(t)$. In particular, we conclude that the curve $t \in [0, t_F) \mapsto f(t,\cdot,\cdot)$ is indeed a curve of probability measures on $\mathbb{R}^d \times \mathcal{G}$. Moreover, the fact that $f_g \in C(0, t_F; W^{-2,p}_{loc}(\mathbb{R}^d))$ and the upper and lower bounds on the densities $f_g(t)$ imply that the curve $t \in [0, t_F) \mapsto f(t)$ (seen as a curve of probability measures) is weakly continuous (here interpreted as weak convergence of probability measures).

It remains to show that the curve is a weak solution to (2.14).

4. Weak solution of (2.14). Let $\zeta \in C_c^\infty(\mathbb{R}^d \times \mathcal{G})$, and let $0 \le r < s < t_F$.

From the convergence $f_g^{\tau_k} \to f_g$ in $C(0, t_F; W^{-2,p}_{loc}(\mathbb{R}^d))$, it follows

$$\int_{\mathbb{R}^d} \zeta_g f_g^{\tau_k}(s,x)\ \ \mathrm{d}x - \int \zeta_g f_g^{\tau_k}(r,x)\ \ \mathrm{d}x \to \int \zeta_g f_g(s,x)\ \ \mathrm{d}x - \int \zeta_g f_g(r,x)\ \ \mathrm{d}x. \tag{6.9}$$

Now, using the fact that $f_g^{\tau_k}(t) \to f_g(t)$ in $L^2_{loc}(\mathbb{R}^d)$ for almost every $t \in [0, t_F)$, and using the upper and lower bounds for $f_g^{\tau_k}(t)$ and $f_g(t)$ we conclude that

$$\begin{aligned}\int_{\mathbb{R}^d} \sum_{g' \in \mathcal{G}} (\zeta_g - \zeta_{g'})(\log f_g^{\tau_k}(t,x) + V_g - [\log f_{g'}^{\tau_k}(t,x) + V_{g'}]e^{-W}\ \ \mathrm{d}x \\ \to \int_{\mathbb{R}^d} \sum_{g' \in \mathcal{G}} (\zeta_g - \zeta_{g'})(\log f_g(t,x) + V_g - [\log f_{g'}(t,x) + V_{g'}])e^{-W}\ \ \mathrm{d}x,\end{aligned} \tag{6.10}$$

for almost every $t \in [0, t_F)$, and

$$\int_{\mathbb{R}^d} \big[\Delta_x \zeta_g - \langle \nabla_x \zeta_g, \nabla_x V_g\rangle\big] f_g^{\tau_k}(t,x)\ \ \mathrm{d}x \to \int_{\mathbb{R}^d} \big[\Delta_x \zeta_g - \langle \nabla_x \zeta_g, \nabla_x V_g\rangle\big] f_g(t,x)\ \ \mathrm{d}x, \tag{6.11}$$

for almost every $t \in [0, t_F)$.

Now, from the upper and lower bounds on $f_g^{\tau_k}$, it follows that for every $t \in [0, t_F)$

$$\int | \sum_{g' \in \mathcal{G}} (\zeta_g - \zeta_{g'})(\log f_g^{\tau_k}(t,x) + V_g - [\log f_{g'}^{\tau_k}(t,x) + V_{g'}])|e^{-W}\ \ \mathrm{d}x \le C_{10}||\zeta||_{L^\infty(\mathbb{R}^d)},$$

for a constant C_{10} that only depends on λ, Λ, $|\mathcal{G}|$, W, and also

$$\int_{\mathbb{R}^d} |\big[\Delta_x \zeta_g - \langle \nabla_x \zeta_g, \nabla_x V_g \rangle\big]| f_g^{\tau_k}(t,x) \quad \mathrm{d}x \leq ||\Delta_x \zeta_g||_{L^\infty(\mathbb{R}^d)} + C_{11} ||\nabla_x \zeta_g||_{L^\infty(\mathbb{R}^d \times \mathcal{G})} ([\nabla_x V]_{e^{-V}})^{1/2},$$

for a constant C_{11} that depends only on λ, Λ, λ', Λ'. We recall that $[\nabla_x V]_{e^{-V}}$ is the quantity defined in Corollary 5.7.

Using the above two inequalities and (6.10), (6.11), we can invoke the dominated convergence theorem twice, and then combine with (6.9) in order to conclude that we can pass to the limit in (6.5). From this it follows that $t \mapsto f(t)$ is a weak solution to (2.14).

□

7 Summary and Discussion on Applications

In this paper, we introduce two types of optimal transport problems in the semi-discrete setting and then study gradient flows of relative entropy functionals with respect to these semi-discrete transport costs. The first problem uses a dynamic formulation a la Benamou–Brenier, and a formal Riemannian structure can be associated to it. The Riemannian formalism is used to motivate systems of equations representing a gradient descent scheme for the minimization of a relative entropy functional; the Riemannian formalism can also be used to motivate accelerated methods for optimization. With the second optimal transport problem (the static one), we seek to more rigorously introduce the notion of gradient flow of the relative entropy functional by considering a minimizing movement scheme of the relative entropy with respect to this cost. Theorem 2.14 establishes an equivalence between the gradient flow equation formally derived through the Riemannian formalism of the first transport cost and the rigorous definition of gradient flow using the minimizing movement scheme with respect to the second transport cost.

There are several theoretical research directions that emanate from our work. First, we believe that it is worth establishing a closer relationship between the two semi-discrete optimal transport problems introduced in the paper (the static and dynamic formulations). Secondly, it is worth emphasizing that our main result on the convergence of the minimizing movement scheme from Sect. 2.5 toward the gradient flow heuristically motivated using the Riemannian formalism was only proved for mobilities that are independent of the mass exchanged among nodes in the graph. We believe that it is worth obtaining a more general result that justifies the connection between these two gradient flows even further.

In the remainder of the paper, we discuss some thoughts on the main application motivating this work.

7.1 From Semi-discrete Optimal Transport to Neural Architecture Search

In machine learning, a neural network is a graph g (the architecture) whose nodes are arranged into layers with edges connecting nodes at different layers. A collection of free parameters (or weights) x is associated with the nodes and edges in the graph. The network architecture g, together with the numerical values of its associated parameters x, determines a series of transformations that, when composed, define a mapping of input vectors (input data) into output vectors (labels). Training a given neural network g essentially means tuning the corresponding parameters x so as to achieve a small mismatch between predicted and observed outputs associated with given training inputs. In other words, the training of a neural network g is the optimization of an objective function (a loss function) over the free parameters x.

In *neural architecture search*, the goal is to find an architecture g that, once trained, gives the best performance possible when predicting data outputs. From a simplistic perspective, this problem can be stated as solving:

$$\min_{(x,g)\in\mathbb{R}^d\times\mathcal{G}} V(x, g). \tag{7.1}$$

where V is thought of as a loss function that typically depends on observed data as well as on additional regularization terms. The variable x (the parameters of a network) can be interpreted as a $\mathbb{R}^d$-valued vector (for d large enough but fixed for simplicity), whereas g can be interpreted as an element in a finite family of architectures $\mathcal{G}$ (which in principle may be quite large). In short, in neural architecture search the optimization is over both the architecture space $\mathcal{G}$ and over the parameters. The tensorized representation of the problem is certainly an oversimplification because, in reality, the parameters x associated to an architecture g do not have an obvious correspondence with the parameters of a different architecture g' (and in fact their dimensions do not even have to match). We will not elaborate much further on this simplification and here we just limit ourselves to saying that while unreasonable when $\mathcal{G}$ is interpreted as the whole space of architectures, the tensorized representation of problem 7.1 is useful when one restricts to a local graph of architectures where one has access to morphisms or correspondences between the parameters of different architectures (just like restricting the optimization of a function defined on a curved manifold to a local chart).

There is an enormous literature on neural architecture search methodologies and some of its applications (see Yu and Zhu 2020 for a brief overview on the subject), but essentially most methods found in the literature fall into two main groups. The first group builds on ideas from reinforcement learning as in Zoph and Le (2016) which uses optimization tools like those described in Williams (1992). The second major group is based on evolutionary algorithms (Stanley and Miikkulainen 2002; Real et al. 2018), where one specifies rules for merging and mutation of different architectures in search of "stronger" architectures. A third type of methodology is the morphism-based hill-climbing strategy from Elsken et al. (2017). There, the authors propose an iterative scheme that alternates between training for a *fixed time* a group of architectures that are determined by a morphism family and then moving in the space of architectures according to the relative performance improvement in such training time.

In all the methodologies listed above, the main objective is to avoid the full training of multiple neural networks (something that would be computationally forbidding), either by building surrogate objective functions that are easier to evaluate, by training networks for a fixed amount of time, or by learning to predict which architectures are more likely to give better results. Many techniques in the literature are based on the above strategies. To name a few: Pham et al. (2018), Liu et al. (2018), Zoph et al. (2017), Liu et al. (2018), Bergstra et al. (2011) and Yu and Zhu (2020).

In this sprawling landscape of methods and techniques for neural architecture search, mathematicians can bring to the table principled ideas and structures for the development of new algorithms or the improvement of existing ones. Take, for example, the hill-climbing algorithm from Elsken et al. (2017) where it is key to tune the amount of time that neural networks have to be trained for. It is intuitively clear that setting a fixed time for training is not ideal as in that way one forces all models to be treated the same regardless of their sizes or architectures. In our paper (Trillos et al. 2021), we elaborate on this issue and propose a method where the training time of architectures is dynamically chosen as dictated by an evolving particle system that is inspired by the gradient flow perspective developed in this paper. All along, our intention was to give meaning to the notion of gradient descent for the optimization of an objective in the space $\mathbb{R}^d \times \mathcal{G}$, i.e., how to propose a gradient-based method for semi-discrete optimization (with neural architecture search as main application in mind). As discussed in Sect. 1.1, in the Euclidean setting there is a well-known connection between gradient flows in the space of measures and dynamics in the base space. In the semi-discrete setting, this connection is sought through particle methods. Particle methods are one way to project to the space $\mathbb{R}^d \times \mathcal{G}$ the dynamics that were lifted to the space of probability measures $\mathcal{P}(\mathbb{R}^d \times \mathcal{G})$ to make sense of a gradient-based scheme. In Trillos et al. 2021, all the nuances that have to be resolved to make this conceptual idea feasible for neural architecture search are discussed.

We hope that the theoretical, methodological and implementation questions briefly described here are able to motivate further research in the mathematics and computer science communities.

Acknowledgements N. García Trillos was supported by NSF-DMS 2005797. The work of J. Morales was supported by NSF grants DMS16-13911, RNMS11-07444 (KI-Net) and ONR grant N00014-1812465. Support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation.

References

Ambrosio, L., Gigli, N.: A User's Guide to Optimal Transport, pp. 1–155. Springer, Berlin (2013)

Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel (2005)

Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numer. Math. **84**(3), 375–393 (2000)

Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 24, pp. 2546–2554. Curran Associates Inc., Red Hook (2011)

Chow, S.-N., Huang, W., Li, Y., Zhou, H.: Fokker-Planck equations for a free energy functional or Markov process on a graph. Arch. Ration. Mech. Anal. **203**(3), 969–1008 (2012)
Chung, F.: Spectral Graph Theory. American Mathematical Society, Providence (1996)
do Carmo, M.P.: *Riemannian Geometry. Mathematics: Theory & Applications*. Birkhäuser Boston, Inc., Boston (1992) (Translated from the second Portuguese edition by Francis Flaherty)
Elsken, T., Metzen, J.-H., Hutter, F.: Simple and efficient architecture search for convolutional neural networks (2017). arXiv:1711.04528
Erbar, M., Fathi, M., Laschos, V., Schlichting, A.: Gradient flow structure for Mckean–Vlasov equations on discrete spaces (2016)
Erbar, M., Maas, J.: Ricci curvature of finite Markov chains via convexity of the entropy. Arch. Ration. Mech. Anal. **206**(3), 997–1038 (2012)
Esposito, A., Patacchini, F.S., Schlichting, A., Slepcev, D.: Nonlocal-interaction equation on graphs: gradient flow structure and continuum limit (2019). arXiv:abs/1912.09834
Figalli, A., Gigli, N.: A new transportation distance between non-negative measures, with applications to gradients flows with dirichlet boundary conditions. J. Math. Pures Appl. **94**(2), 107–130 (2010)
Garbuno-Inigo, A., Hoffmann, F., Li, W., Stuart, A.M.: Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler (2019). arXiv:1903.08866
García Trillos, N.: Gromov-Hausdorff limit of Wasserstein spaces on point clouds. Calc. Var. **59**, 73 (2020). https://doi.org/10.1007/s00526-020-1729-3
Gigli, N., Maas, J.: Gromov–Hausdorff convergence of discrete transportation metrics. SIAM J. Math. Anal. **45**(2), 879–899 (2013)
Gladbach, P., Kopfer, E., Maas, J.: Scaling limits of discrete optimal transport. SIAM J. Math. Anal. **52**(3), 2759–2802 (2020)
Gladbach, P., Kopfer, E., Maas, J., Portinale, L.: Homogenisation of one-dimensional discrete optimal transport. J. Math. Pures Appl. **139**, 204–234 (2020)
Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker–Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.-J., Fei-Fei, L., Yuille, A., Huang, J., Murphy, K.: Progressive neural architecture search. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision—ECCV 2018, pp. 19–35. Springer, Cham (2018)
Maas, J.: Gradient flows of the entropy for finite Markov chains. J. Funct. Anal. **261**(8), 2250–2292 (2011)
Mielke, A.: A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems. Nonlinearity **24**(4), 1329–1346 (2011)
Mielke, A.: Geodesic convexity of the relative entropy in reversible Markov chains. Calc. Var. Partial Differ. Equ. **48**(1), 1–31 (2013)
Peyré, G., Cuturi, M.: Computational Optimal Transport: With Applications to Data Science, Foundations and Trends in Machine Learning, vol. 11, pp. 355–607 (2019)
Pham, H., Guan, M., Zoph, B., Le, Q., Dean, J.: Efficient neural architecture search via parameters sharing. In: Dy, J., Krause, A. (eds) *Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research: PMLR*, pp. 4095–4104. Stockholmsmässan, Stockholm, 10–15 Jul (2018)
Real, E., Aggarwal, A., Huang, Y., Le, Q.V.: Regularized evolution for image classifier architecture search. In: *AAAI* (2018)
Simon, J.: Compact sets in the space Lp(o, t; b). Annali di Matematica Pura ed Applicata **146**, 65–96 (1986)
Stanley, K.O., Miikkulainen, R.: Evolving neural networks through augmenting topologies. Evol. Comput. **10**(2), 99–127 (2002)
Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. **17**(153), 1–43 (2016)
Trillos, N.G., Morales, F., Morales J.: Traditional and accelerated gradient descent for neural architecture search. In: Nielsen F., Barbaresco F. (eds.) Geometric Science of Information. GSI 2021. Lecture Notes in Computer Science, vol. 12829. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80209-7_55
Villani, C.: Optimal Transport. Springer, Berlin (2009)
Williams, R.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**(8), 229–256 (1992)
Yu, T., Zhu, H.: Hyper-parameter optimization: a review of algorithms and applications (2020). arXiv:2003.05689

Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning (2016). arXiv:1611.01578
Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition (2017). arXiv:1707.07012